*На правах рукописи*

СТРИЖОВ ВАДИМ ВИКТОРОВИЧ

# ПОРОЖДЕНИЕ И ВЫБОР МОДЕЛЕЙ В ЗАДАЧАХ РЕГРЕССИИ И КЛАССИФИКАЦИИ

Москва — 2014

# Оглавление

# Введение

Работа посвящена проблемам выбора моделей в задачах регрессионного анализа и классификации. Предлагается подход, согласно которому выбор производится из индуктивно-порождаемого множества моделей. Анализируется распределение параметров моделей. На основании этого анализа выбирается модель оптимальной сложности.

*Ключевые слова*: машинное обучение, интеллектуальный анализ данных, регрессионный анализ, классификация, выбор моделей, порождение моделей, байесовский подход.

**Актуальность темы.** Модель, описывающая исследуемое явление, может быть получена двумя путями: во-первых, методами математического моделирования, во-вторых, методами анализа данных и информационного моделирования. Первый тип моделей интерпретируем экспертами в контексте моделируемого явления [314]. Второй тип моделей не всегда интерпретируем, но более точно приближает данные [46]. Совмещение достоинств обоих подходов, результатом которого является получение интерпретируемых и достаточно точных моделей, является актуальной задачей теоретической информатики.

Центральным объектом исследования является проблема построения адекватных моделей регрессии и классификации при решении задач прогнозирования. Проблема заключается в отыскании моделей оптимальной сложности, которые описывают измеряемые данные с заданной точностью. Дополнительным ограничением является интерпретируемость моделей экспертами той предметной области, для решения задач которой создается модель.

Цель исследования заключается в создании и обосновании методов выбора моделей из индуктивно порождаемого множества, а также в исследовании свойств алгоритмов выбора моделей. Задача выбора моделей из счетного порождаемого множества поставлена впервые. При постановке задачи использовался обширный материал о способах выбора моделей и выбора признаков из конечного множества, наработанный ранее в области машинного обучения. Эта задача является одной из центральных проблем машинного обучения и интеллектуального анализа данных.

Основной задачей исследования является разработка методов последовательного порождения моделей и оценки ковариационных матриц параметров моделей с целью управления процедурой выбора моделей. Основной сложностью такой задачи является необходимость выбора из значительного числа регрессионных моделей, либо необходимость оценки параметров структурно сложной, так называемой «универсальной» модели.

Взаимосвязь задачи порождения и задачи выбора регрессионных моделей была освещена в начале 1980-х годов А. Г. Ивахненко. Согласно предложенному им методу группового учета аргументов [307, 309, 308], модель оптимальной структуры может быть найдена путем последовательного порождения линейных моделей, в которых компоненты являются мономами полинома Колмогорова-Габора от набора независимых переменных. Критерий оптимальности структуры модели задается с помощью скользящего контроля.

В отличие от этого метода, метод символьной регрессии [285, 156, 167, 193] рассматривает порождение произвольных нелинейных суперпозиций базовых функций. В последние годы тема анализа сложности моделей, получаемых с помощью этого метода, стала распространенным предметом исследований [106, 273].

Первоначально принципы индуктивного порождения моделей были предложены в методе группового учета аргументов. Структура суперпозиций задавалась при этом внешними критериями качества модели. Впоследствии эти критерии были обоснованы в рамках гипотезы порождения данных с помощью связанного байесовского вывода. При последовательном порождении моделей необходимо оценивать информативность элементов суперпозиции. В рамках метода байесовской регрессии [45, 29, 49]

для этого предложено использовать функцию плотности распределения параметров модели. Эта функция является параметрической и ее параметры были названы гиперпараметрами [269, 46, 48, 47]. Было предложено использовать гиперпараметры моделей для оценки информативности элементов суперпозиции, что сделало анализ гиперпараметров одним из способов выбора моделей.

Для модификации суперпозиций нелинейных моделей был предложен метод оптимального прореживания [176, 129] Согласно этому методу, элемент суперпозиции можно отсечь как неинформативный, если значение выпуклости функции ошибки от параметров модели не превосходит относительный заданный порог.

Задача выбора модели является одной из самых актуальных в регрессионном анализе. В современной зарубежной литературе для ее решения используется принцип минимальной длины описания. Он предлагает использовать для описания данных наиболее простую и одновременно наиболее точную модель [115, 120, 116, 117, 165].

Задача сравнения моделей детально разработана [190, 189, 191, 60, 188]. Как альтернатива информационным критериям [56, 57, 14, 15, 84, 274] был предложен метод двухуровневого байесовского вывода. На первом уровне вывода настраиваются параметры моделей. На втором уровне настраиваются их гиперпараметры. Согласно этому методу, вероятность выбора более сложной модели ниже вероятности выбора простой модели при сравнимом значении функции ошибки на регрессионных остатках. Принципы байесовского подхода для выбора линейных моделей регрессии и классификации предложены авторами [62, 26, 30, 31].

В то же время, в упомянутых публикациях и подходах остается открытым ряд важных проблем. Поэтому представляется целесообразным создать и развить теорию порождения и выбора регрессионных моделей. Она заключается в следующем. Множество моделей заданного класса индуктивно порождается набором параметрических базовых функций, заданных экспертами. Каждая модель является допустимой суперпозицией таких функций.

Интерпретируемость моделей обеспечена тем, что каждая из порождаемых моделей является суперпозицией базовых функций, заданных экспертами. Класс моделей задается правилами порождения суперпозиций. Точность моделей обеспечивается тем, что рассматривается достаточно большой набор моделей-претендентов, из которого выбирается оптимальная модель. Критерий оптимальности включает в себя понятия сложности и точности модели. При построении критерия учитывается гипотеза порождения данных — предположение о распределении регрессионных остатков.

Одновременно с оценкой параметров вычисляются и гиперпараметры (параметры распределения параметров) модели. На основе гиперпараметров оценивается информативность элементов суперпозиции и оптимизируется её структура. Оптимальные модели выбираются согласно критерию, заданному гипотезой порождения данных. Множество моделей индук-

тивно порождается из набора базовых функций, заданных экспертами. Каждая модель является допустимой суперпозицией базовых функций. Одновременно с оценкой параметров моделей выполняется также и оценка гиперпараметров функции распределения параметров моделей. На основе этих параметров оценивается информативность элементов суперпозиции и принимается решение об оптимизации ее структуры. Оптимальные модели выбирается согласно критерию, заданному гипотезой порождения данных.

**Благодарности.** Автор признателен чл.-корр. РАН Константину Владимировичу Рудакову за поддержку и внимание к работе, д.ф.-м.н. Константину Вячеславовичу Воронцову за обсуждение содержания работы и критические замечания, а также аспирантам Вычислительного центра РАН и студентам кафедры «Интеллектуальные системы» Факультета управления и прикладной математики Московского физико-технического института Михаилу Кузнецову, Анастасии Мотренко, Роману Сологубу, Алексею Зайцеву, Александру Адуенко, Анне Варфоломеевой, Арсентию Кузьмину, Марии Стениной, Георгию Рудому, Александре Токмаковой и Александру Катруце за сотрудничество и участие в многочисленных вычислительных экспериментах, проводимых при исследовании свойств предлагаемых методов.

# 1. Постановка задачи выбора моделей

Важным свойством регрессионных моделей является возможность интерпретации её структуры и её параметров в контексте решаемой прикладной задачи. Различают термины «математическая модель» и «регрессионная модель». Математическая модель [314, 313] предполагает участие специалиста-аналитика в конструировании функции, которая описывает некоторую известную закономерность [40, 93, 357]. Математическая модель является интерпретируемой — объясняемой в рамках исследуемой закономерности. При построении математической модели сначала создается параметрическое семейство функций, затем с помощью измеряемых данных выполняется идентификацией модели, состоящая в нахождении её параметров [184]. Основное отличие математического моделирования от регрессионного анализа состоит в том, что в первом случае функциональная известна связь зависимой переменной и свободных переменных. Специфика математического моделирования состоит в том, что измеряемые данные используются для верификации, но не для построения модели: модель строится исходя из экспертных предположений о характере и законах моделируемого явления. При этом затруднительно получить модель сложного явления, в котором взаимосвязано большое число различных факторов.

Регрессионные модели образуют широкий класс функций, которые описывают некоторую закономерность [80]. При этом для построения модели в основном используются измеряемые данные, а не знание свойств исследуемой закономерности. Такая модель часто неинтерпретируема с точки зрения специалистов данной прикладной задачи, но более точна. Это объясняется либо большим числом моделей-претендентов, которые используются для построения оптимальной модели, либо большей сложностью модели [104, 119, 235, 214].

И на регрессионную, и на математическую модель, накладывается требование непрерывности отображения. Требование непрерывности обусловлено классом решаемых задач: чаще всего это описание физических, химических и других явлений, где требование непрерывности выставляется естественным образом [340, 276, 301, 220, 218, 159]. Примеры регрессионных моделей: линейные функции, алгебраические полиномы, ряды Чебышёва, нейронные сети без обратной связи, функции радиального базиса. Модель также может быть представлена в виде суперпозиции функций свободных переменных из некоторого набора. На функцию регрессии также могут накладываться ограничения монотонности, гладкости, измеримости и некоторые другие [354, 302, 301, 340].

Термин «регрессия» введен Фрэнсисом Гальтоном в конце XIX века [55]. Гальтон обнаружил, что дети родителей с высоким или низким ростом как правило не наследуют выдающийся рост и назвал эту закономерность «регрессия к посредственности» [102]. Сначала этот термин использовался исключительно в биологическом смысле. После работ Карла Пирсона его стали использовать и в статистике [230]. [23].

Регрессионное моделирование и математическое связаны подходом, который называется *суррогатным моделированием* [113, 155]. Согласно этому подходу, сложная в создании или идентификации математическая модель приближается функцией регрессии. Дана функция $u$ дискретного или непрерывного аргумента. Требуется найти функцию $f$ из некоторого параметрического семейства, например, среди алгебраических полиномов заданной степени.

Параметры функции $f$ должны доставлять минимум некоторому функционалу, например, $\rho(f,u) = \left(\frac{1}{b-a}\int_a^b |f(x)-u(x)|^2 dx\right)^{\frac{1}{2}}$.

При прогнозе с использованием регрессионных моделей используется подход, называемый интер- или экстраполяцией. *Интерполяция* функций — частный случай задачи приближения, когда требуется, чтобы в определенных точках, называемых узлами интерполяции, значения функции $u$ и приближающей её функции $f$ совпадали. В более общем случае накладываются ограничения на значения некоторых производных $f$. То есть, дана функция $u$ дискретного аргумента. Требуется отыскать такую функцию $f$, график которой проходит через все точки $u$. При этом понятие расстояния обычно не используется, однако часто вводится понятие гладкости искомой функции.

В работе описаны аналитические и стохастические алгоритмы оптимизации структурных параметров прогностических регрессионных моделей. Исследуется оптимизация параметров линейных, обобщенно-линейных и нелинейных моделей. Приняты статистические гипотезы о распределении зависимой переменной и параметров модели. На основании этих предположений принята оптимизируемая функция ошибки. Аналитические алгоритмы основаны на получении оценок производных функции ошибок относительно параметров модели. Статистические алгоритмы основаны на сэмплироваинии параметров модели и на процедуре скользящего контроля элементов регрессионной выборки. Алгоритмы протестированы на наборе синтетических и реальных задач. Представлены результаты сравнения алгоритмов. Выполнен анализ ошибок.

При моделировании измеряемых данных одной из важных проблем является оценка точности модели, аппроксимирующей эти данные. Для оценки точности аппроксимации вводится функция ошибки, оптимизируемая в данной работе. Предполагая, что данные измеряются с некоторой погрешностью, будем рассматривать моделирование данных как задачу восстановления регрессии [80, 169, 243, 49, 251, 268].

Эта задача состоит в нахождении функции регрессии и оценке параметров регрессионной модели [203, 57, 187]. Параметры модели назначаются таким образом, что модель наилучшим образом приближает данные, минимизируя функцию ошибки. Измеряемые данные представляют собой пары значений зависимой и независимой переменной.

Для оценки качества решения этой задачи вводится *функция ошибки*, исходя из значения которой делается вывод о том, насколько хорошо модель приближает данные, а также насколько адекватна гипотеза порождения данных [46, 131, 178]. Функция ошибки играет определяющую роль в выборе параметров регрессионной модели и зависит также от структурных параметров, которые определяются гипотезой порождения данных или априорными знаниями о виде модели. В данной работе функция ошибки назначается путем байесовского вывода [46, 87, 219, 227].

Структурными параметрами являются регуляризирующие параметры, включаемые в функционал качества для штрафа на вектор параметров модели [27, 61, 109]. Оценка структурных параметров [63, 185] является центральной задачей в данной работе. Для оценки используется метод максимизации правдоподобия модели [224, 21, 16].

Одним из методов максимизации правдоподобия модели является метод аппроксимации Лапласа [208, 153, 191], в основе которого лежат гипотезы нормального распределения за-

висимой переменной и вектора параметров модели. В зависимости от вида ковариационных матриц нормальных распределений зависимой переменной и вектора параметров, вводится ряд упрощений для максимизации правдоподобия модели. Рассматриваются ковариационные матрицы диагонального типа, скалярного типа и общего вида.

Альтернативным методом оценки правдоподобия модели является метод Монте-Карло [17, 41]. Согласно этому методу, производится процедура сэмплирования параметров модели при фиксированных структурных параметрах, и строится аппроксимирующая интеграл сумма значений правдоподобия по сэмплированным параметрам. Оптимальными структурными параметрами считаются те, которые доставляют максимум аппроксимирующей функции.

Для проверки качества предлагаемых алгоритмов используется метод скользящего контроля оценки структурных параметров [131, 20]. Этот метод не использует вероятностных предположений о структуре модели. Метод основан на многократном разбиении выборки на обучающую и контрольную части и подсчете функции ошибки на контрольной части выборки. Наилучшим структурным параметрам соответствуют те, при которых модель дает минимальную среднюю ошибку на различных разбиениях [28, 95, 115, 116].

Помимо моделей общего вида в работе в качестве частного случая рассматриваются линейные модели [203, 242]. Их отдельное рассмотрение позволяет выписать некоторые формулы, как, например, вычисление гессиана функции ошибки [286], в явном виде, и избежать избыточной оптимизации.

В работе рассматриваются также алгоритмы оптимизации структурных параметров регрессионной модели. Для оценки правдоподобия модели алгоритмы используют метод аппроксимации Лапласа функции ошибки, метод Монте-Карло, метод скользящего контроля. Исследованы свойства предлагаемых методов: сходимость, вычислительная сложность.

При постановке и решении задач регрессионного анализа [218, 209, 124, 206, 70, 12, 255, 134, 136, 245, 103, 64, 281, 96, 107, 237, 140, 159, 326] встают следующие фундаментальные вопросы.

Как выбрать структуру модели?

Какова гипотеза порождения данных, каково распределение случайной переменной, какому семейству оно должно принадлежать?

Какова связь гипотезы порождения данных и распределения параметров модели?

Какой функцией ошибки требуется оценивать качество аппроксимации?

Как оценить параметры модели, каков должен быть алгоритм оптимизации параметров?

Эти вопросы рассматриваются ниже в постановочной главе, и далее в данной работе.

## 1.1. Функция регрессии и регрессионная модель

*Регрессионный анализ* — метод анализа измеряемых данных и исследования связи между независимыми и зависимыми переменными [80, 302, 329, 321]. Измеряемые данные представ-

ляют собой пары значений зависимой переменной $y$ и независимой переменной $\mathbf{x}$. Считается [333], что эта зависимость является статистической и имеет вид

$$\mathsf{E}(y|\mathbf{x}) = f(\hat{\mathbf{w}}, \mathbf{x}). \tag{1}$$

*Регрессия* — математическое ожидание случайной величины $y$, зависящей от другой величины или от нескольких величин $\mathbf{x}$. Зависимость $f$ называется *функцией регрессии* от независимой переменной $\mathbf{x}$ при некоторых фиксированных параметрах $\hat{\mathbf{w}}$. Переменная $\mathbf{x}$ также называется регрессором. Точность, с которой функция $f$ передает изменение в среднем при изменении $\mathbf{x}$, измеряется дисперсией $y$, вычисляемой для каждого $\mathbf{x}$: $\mathsf{D}(y|\mathbf{x}) = \sigma_y^2(\mathbf{x})$. Если $\mathsf{D}(y|\mathbf{x}) = 0$ при всех значениях $\mathbf{x}$, то с вероятностью равной единице эти величины связаны функциональной зависимостью. Если $\mathsf{D}(y|\mathbf{x}) \neq 0$ ни при каком значении $\mathbf{x}$ и $f(\hat{\mathbf{w}}, \mathbf{x})$ не зависит от $\mathbf{x}$, то регрессия $y$ по $\mathbf{x}$ отсутствует.

Основной задачей регрессионного анализа является нахождение функции регрессии $f$ и оценка параметров $\mathbf{w}$. При решении этой задачи используются измеряемые данные — выборка реализаций свободных переменных и зависимой переменной.

**Определение 1.** *Регрессионная выборка* $\mathfrak{D} = \left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^m$ *— множество* $m$ *пар, состоящих из вектора* $\mathbf{x}_i = [x_{ij}]_{j=1}^n$ *значений* $n$ *свободных переменных и соотвествующего этому вектору значения зависимой переменной* $y_i$.

Далее предполагается, что переменные принадлежат множеству действительных чисел, либо его подмножеству: $\mathbf{x} \in \mathbb{X} \subseteq \mathbb{R}^n$ и $y \in \mathbb{Y} \subseteq \mathbb{R}^1$. Индекс $i$ элемента выборки и индекс $j$ свободной переменной рассматриваются как элементы конечных множеств $i \in \mathcal{I} = \{1, \ldots, m\}$ и $j \in \mathcal{J} = \{1, \ldots, n\}$.

В дальнейшем будет использоваться также обозначение $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$, где $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$ — вектор значений зависимой переменной и $\mathbf{X}$ — *матрица плана*

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\intercal \\ \vdots \\ \mathbf{x}_m^\intercal \end{bmatrix}.$$

Матрицу $\mathbf{X}$ можно представить в виде

$$\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n],$$

где $\boldsymbol{\chi}_j$ — $j$-й столбец матрицы, $\boldsymbol{\chi}_j = [\chi_1, \ldots, \chi_n]^\intercal$. Регрессионную выборку также удобно представить как пару

$$\mathfrak{D} = (\mathbf{X}, \mathbf{y}).$$

Предполагается, что элементы выборки связаны соотношением

$$y_i = f(\mathbf{w}, \mathbf{x}_i) + \varepsilon(\mathbf{x}_i), \tag{2}$$

которое аддитивно включает случайную величину $\varepsilon = \varepsilon(\mathbf{x})$. Предположение о том, что то зависимая переменная есть сумма значений модели и некоторой случайной величины, сохраняется и ниже. Мультипликативное включение случайной величины в соотношении (2) может быть представлено в аддитивном виде путем логарифмирования обеих частей выражения при условии, что независимые переменные принимают положительные значения.

**Определение 2.** *Ошибкой или регрессионным остатком* $\varepsilon_i$ *называется разность между значением функции регрессии* $f(\hat{\mathbf{w}}, \mathbf{x}_i)$ *и значением зависимой переменной, соответствующей некоторой свободной переменной* $\mathbf{x}_i$*:*

$$\varepsilon_i = f(\hat{\mathbf{w}}, \mathbf{x}_i) - y_i,$$

*или*

$$\boldsymbol{\varepsilon} = \mathbf{f} - \mathbf{y},$$

*где вектор-функция* $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X}) = [f(\mathbf{w}, \mathbf{x}_1), \ldots, f(\mathbf{w}, \mathbf{x}_m)]^\intercal \in \mathbb{Y}$.

Выборка может быть как функцией дискретного аргумента, так и отношением. Например, данные для построения регрессии могут быть такими: $\mathfrak{D} = \{(0,0), (0,1), (0,2), (1,1), (1,2), (1,3)\}$. В такой выборке одному значению переменной $\mathbf{x}$ соответствует несколько значений переменной $y$, как показано на рис. 1.

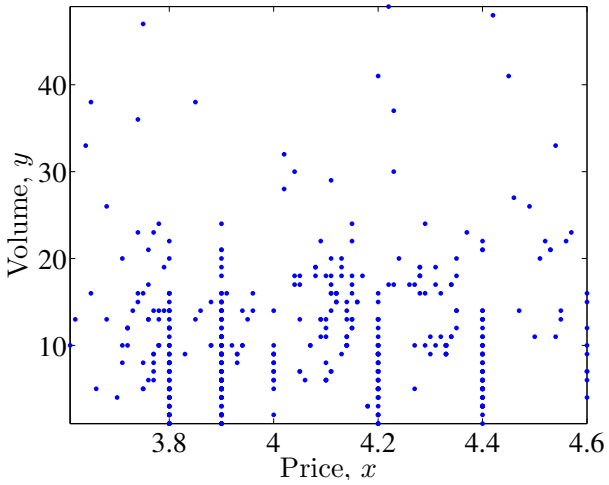



Рис. 1. Пример выборки: зависимость объема продажи товара от цены.

Для нахождения функции регрессии $f$ используются регрессионные модели.

**Определение 3.** *Регрессионная модель — параметрические семейство функций, отображение*

$$f : \mathbb{W} \times \mathbb{X} \to \mathbb{Y}$$

*декартова произведения области допустимых значений* $\mathbb{W}$ *параметров модели и области допустимых значений* $\mathbb{X}$ *свободных переменных в область значения* $\mathbb{Y}$ *зависимой переменной. Иначе, регрессионная модель есть поэлементное отображение*

$$f : (\mathbf{w}, \mathbf{x}) \mapsto y,$$

*в котором вектор параметров* $\mathbf{w} \in \mathbb{W}$*, свободная переменная* $\mathbf{x} \in \mathbb{X}$ *и зависимая переменная* $y \in \mathbb{Y}$.

Синонимами термина «регрессионная модель» являются термины «теория», «гипотеза». Эти термины используются в статистике, в частности в разделе «проверка статистических

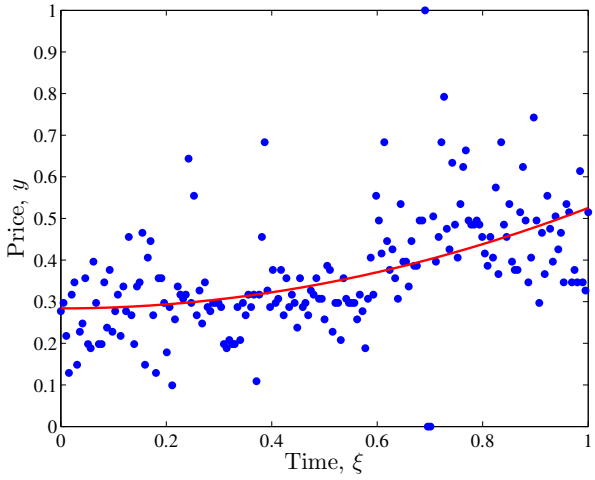



Рис. 2. Пример функции регрессии: зависимость цены товара от времени.

гипотез» [178, 220, 210]. Регрессионная модель есть гипотеза, которая должна быть подвергнута статистической проверке, после чего она принимается или отвергается при выполнении процедуры выбора.

В дальнейшем предполагается, что выполнены следующие базовые условия:

1) функция $f$ является непрерывной и гладкой по аргументу $\mathbf{w}$,

2) множества $\mathbb{W}$, $\mathbb{X}$ и $\mathbb{Y}$ являются подмножествами степеней декартовых произведений множества действительных чисел $\mathbb{R} \times \cdots \times \mathbb{R}$.

**Определение 4.** *Функция регрессии $f^*$ — функция полученная путем сужения области определения регрессионной модели $f$*

$$f|_{\mathbb{W} \ni \mathbf{w}=\hat{\mathbf{w}}} : \mathbb{X} \to \mathbb{Y}$$

*на заданное значение вектора параметров $\hat{\mathbf{w}}$.*

Далее для краткости регрессионная модель будет называться моделью, а функция регрессии будет называться регрессией. В машинном обучении модель называют настроенной или обученной, если зафиксированы её параметры.

Различают следующие виды регрессионных моделей:

1) *линейные модели* [169, 242] — модели, которые могут быть представлены в виде скалярного произведения вектора свободных переменных и вектора параметров модели

$$f = \langle \mathbf{w}, \mathbf{x} \rangle, \tag{3}$$

в частности, линейными являются полиномиальные и криволинейные модели;

2) *обобщенно-линейные модели* [203, 78, 127, 181, 177, 125] — модели вида

$$f = \mu^{-1} \langle \mathbf{w}, \mathbf{x} \rangle, \tag{4}$$

где функция $\mu$, называющаяся функцией связи, принадлежит множеству функций, заданному гипотезой о том, что распределение зависимой переменной принадлежит экспоненциальному семейству, см. раздел 1.2.;

3) *нелинейные модели* [243] — модели вида

$$f = f(\mathbf{w}, \mathbf{x}), \tag{5}$$

которые не могут быть представлены как линейные (3) или обобщенно-линейные (4).

Различают *одномерную* и *многомерную* регрессию с одной и с несколькими свободными переменными. Будем считать, что свободная переменная — вектор $\mathbf{x} \in \mathbb{R}^n$. В частных случаях, когда свободная переменная является скаляром, она будет обозначаться $\xi$. На рис. 2 показана функция регрессии $f(\hat{\mathbf{w}}, \xi) = w_1 + w_2\xi^2 + \varepsilon(\xi)$. Ее оптимальные параметры $\hat{\mathbf{w}} = [0.2839, 0.2412]$. Соответствующая регрессионная модель имеет вид $f = \mathbf{x}^\intercal\mathbf{w}$, где $x_1 = \xi^0, x_2 = \xi^2$.

## 1.2. Гипотеза порождения данных

Так как переменная $y$ рассматривается в регрессионном анализе как случайная величина, то при восстановлении функции регрессии $f^*$ и параметров $\mathbf{w}$ регрессионной модели $f(\mathbf{w}, \mathbf{x})$ используются вероятностные гипотезы.

**Определение 5.** *Гипотезой порождения данных называется предположение о виде распределения случайной величины $y$ и значеních параметров этого распределения (если распределение параметрическое).*

Эта гипотеза играет центральную роль в выборе критерия оценки качества модели и, как следствие, в методе оценки параметров модели. Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом регрессионных остатков [22, 178, 220, 210]. При этом считается, что независимая переменная $\mathbf{x}$ не является случайной величиной, не содержит ошибок и не нуждается в дополнительных статистических гипотезах.

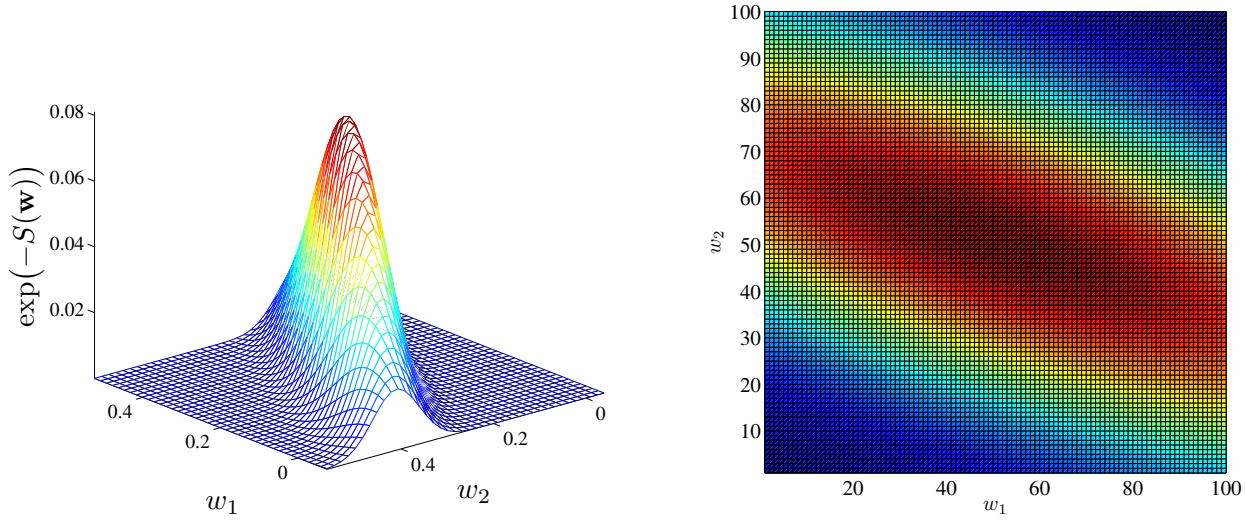



Рис. 3. Вид функции $\exp\big(-S(\mathbf{w})\big)$ в окрестности оптимального вектора параметров.

Так как дисперсия и математическое ожидание зависимой переменной $y$ зависит от реализаций $\mathbf{x}_1, \ldots, \mathbf{x}_m$ свободной переменной $\mathbf{x}$, будем считать реализации зависимой переменной $y$ многомерной случайной величиной $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$. В регрессионных задачах рассматриваются две основные многомерные случайные величины: зависимая переменная $\mathbf{y}$ и вектор параметров $\mathbf{w}$.

*Многомерная случайная величина* — упорядоченный набор (вектор) $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$ фиксированного числа $m$ одномерных случайных величин. *Многомерное наблюдение* $\bar{\mathbf{y}}$ — реализация многомерной случайной величины $\mathbf{y}$. *Многомерная выборка* $\bar{Y} = [\bar{\mathbf{y}}_1, \ldots, \bar{\mathbf{y}}_n]$ — неупорядоченный набор фиксированного числа $n$ многомерных наблюдений. Основными числовыми характеристиками многомерной случайной величины являются *вектор средних* и *ковариационная матрица*.

*Вектор средних* — вектор математических ожиданий многомерной случайной величины, $\mathsf{E}(\mathbf{y}) = [\mathsf{E}(y_1), \ldots, \mathsf{E}(y_m)]^\intercal$. Оценкой вектора средних по многомерной выборке $\mathbf{Y}$, если верна гипотеза нормального распределения $\mathbf{y}$, является среднее значение её реализаций,

$$\mathsf{E}(\mathbf{y}) = \mathbf{f}(\mathbf{w}, \mathbf{Y}) = \frac{1}{n} \sum_{j \in \{1,\ldots,n\}} \bar{y}_j,$$

иначе — среднее значение по строкам матрицы $\mathbf{Y}$. При этом каждый элемент вектора параметров $\mathbf{w}$ равен $\frac{1}{n}$, где $n$ — число реализаций.

В качестве примера получения многомерной выборки по наблюдениям зависимой переменной можно привести восстановление регрессии непараметрическими методами в задачах прогнозирования временных рядов. При этом каждая строка матрицы $\mathbf{Y}$ содержит значения временного ряда в окрестности, заданной независимой переменной — временем.

Пусть элементы многомерной случайной величины $\mathbf{y}$ имеют конечные дисперсии. Ковариационной матрицей величины $\mathbf{y}$ называется квадратная матрица

$$\mathbf{\Sigma} = [\sigma_{ij}^2], i, j \in \mathcal{I},$$

элементы которой $\sigma_{ij}^2 = \operatorname{cov}(y_i, y_j) = \mathsf{E}\big((y_i - \mathsf{E}y_i)(y_j - \mathsf{E}y_j)\big)$ — ковариации случайных величин $y_i$ и $y_j$. На главной диагонали матрицы находятся дисперсии $\sigma_{ii}$ случайных величин $y_i$. Оценкой $\hat{\mathbf{\Sigma}}$ ковариационной матрицы по многомерной выборке $\mathbf{Y}$ является

$$\hat{\mathbf{\Sigma}} = (m-1)^{-1} \bar{\mathbf{Y}}^\intercal \bar{\mathbf{Y}},$$

где $\bar{\mathbf{Y}}$ обозначает центрированность столбцов матрицы $\mathbf{Y}$. Ковариационная матрица симметрична и неотрицательно определена,

$$\mathbf{\Sigma} = \mathbf{\Sigma}^\intercal, \quad \mathbf{y}^\intercal \mathbf{\Sigma} \mathbf{y} \geqslant 0, \quad \mathbf{y} \in \mathbb{R}^m.$$

В дальнейшем рассматриваются три варианта описания многомерных случайных величин $\mathbf{y}$ и $\mathbf{w}$ посредством ковариационных матриц. При этом используется матрица $\mathbf{B}$, обратная к ковариационной матрице величины $\mathbf{y}$,

$$\mathbf{B}^{-1} = \mathbf{\Sigma}.$$

Ниже во всех примерах предполагается нормальное распределение $\mathcal{N}(\cdot,\cdot)$ соответствующих многомерных случайных величин.

Для зависимой переменной $\mathbf{y}$ выполняется один из трех вариантов гипотезы порождения данных:

1) элементы зависимой переменной, случайной величины $\mathbf{y}$, имеют одинаковую дисперсию $\sigma^2(\mathbf{y})$ и независимы, $\operatorname{cov}(y_i, y_j) = 0, i, j \in \mathcal{I}, i \neq j$,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma^2(\mathbf{y})\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{f}, \beta^{-1}\mathbf{I}); \tag{6}$$

2) элементы переменной $\mathbf{y}$ имеют различную дисперсию и независимы, то есть,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \operatorname{diag}^{-1}(\beta_1, \ldots, \beta_m)\mathbf{I}); \tag{7}$$

3) элементы переменной $\mathbf{y}$ описываются ковариационной матрицей общего вида,

$$\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B}^{-1}). \tag{8}$$

В выражениях (3), (4) и (5) переменные $\mathbf{y}$ и $\mathbf{w}$ связаны функциональной зависимостью. Поэтому параметры модели $f$ также являются случайными величинами. Распределение этих параметров зависит от гипотезы порождения данных. При заданной линейной модели

$$\mathsf{E}(\mathbf{y}|\mathbf{X}) = \mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{X}\mathbf{w},$$

либо при заданном линеаризованном виде

$$\mathsf{E}(\mathbf{y} - \mathbf{f}|\mathbf{X}) = \mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{J}\Delta\mathbf{w},$$

в котором матрица $\mathbf{J}$ есть матрица Якоби функции $f$, линейное отображение задаваемое матрицей $\mathbf{X}$, переводит распределение многомерной случайной величины $\mathbf{y}$ в распределение многомерной случайной величины $\mathbf{w}$. В случае нормального распределения случайной величины $\mathbf{y}$ распределение величины $\mathbf{w}$ также является нормальным. Обозначим $\mathbf{A}^{-1}$ ковариационную матрицу параметров $\mathbf{w}$.

Как и для зависимой переменной $\mathbf{y}$, для параметров $\mathbf{w}$ выполняется один из трех вариантов:

1) параметры имеют одинаковую дисперсию $\sigma^2(\mathbf{w})$ и независимы, $\operatorname{cov}(w_j, w_k) = 0, j, k \in \mathcal{J}, j \neq k$,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{ML}}, \sigma^2(\mathbf{w})\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I}); \tag{9}$$

2) параметры модели имеют различную дисперсию и независимы,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{ML}}, \operatorname{diag}^{-1}(\alpha_1, \ldots, \alpha_n)\mathbf{I}); \tag{10}$$

3) параметры модели описываются ковариационной матрицей общего вида

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{ML}}, \mathbf{A}^{-1}). \tag{11}$$

Таблица 1. Варианты гипотезы порождения зависимой переменной и параметров модели.

| | Зависимая переменная $\mathbf{y}$ | Параметры модели $\mathbf{w}$ | Обозначения |
|---|---|---|---|
| 1) | $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \sigma^2(\mathbf{y})\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{f}, \beta^{-1}\mathbf{I})$ | $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \sigma^2(\mathbf{w})\mathbf{I}) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ | $\mathbf{A} = \alpha\mathbf{I}$ |
| 2) | $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \operatorname{diag}^{-1}(\beta_1, \ldots, \beta_m)\mathbf{I})$ | $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \operatorname{diag}^{-1}(\alpha_1, \ldots, \alpha_n)\mathbf{I})$ | $\mathbf{A} = \operatorname{diag}(\alpha_i)\mathbf{I}$ |
| 3) | $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B}^{-1})$ | $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_0, \mathbf{A}^{-1})$ | $\mathbf{A} \in \mathbb{M}^n$ |

**Определение 6.** *Параметры $\alpha_j, j \in \mathcal{J}$ распределения параметров $\mathbf{w}$ модели $f(\mathbf{w}, \mathbf{x})$ называются гиперпараметрами. Ковариационная матрица $\mathbf{A} = [\alpha_{kj}]$ называется матрицей гиперпараметров.*

Для гипотез (9), (10) и (11) матрица гиперпараметров имеет, соответственно, вид

1) $\mathbf{A} = \alpha\mathbf{I}$,

2) $\mathbf{A} = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)\mathbf{I}$,

3) $\mathbf{A}$.

Аналогично, назовем гиперпараметрами и обозначим $\beta_i, i \in \mathcal{I}$ параметры распределения зависимой переменной $\mathbf{y}$.

**Варианты гипотезы порождения данных.** Считаем вектор зависимых переменных $\mathbf{y}$ и вектор параметров $\mathbf{w}$ многомерными нормально распределенными случайными величинами с ковариационными матрицами $\mathbf{A}^{-1}$ и $\mathbf{B}^{-1}$ соответственно. Чтобы получить оценки гиперпараметров $\mathbf{A}, \mathbf{B}, \mathbf{w}$, введем ограничения на вид распределений $p(\mathfrak{D}|\mathbf{w}, \mathbf{B})$ и $p(\mathbf{w}|\mathbf{A})$. Для зависимой переменной $\mathbf{y}$ и для параметров $\mathbf{w}$ выполняется один из трех вариантов гипотезы порождения данных, продемонстрированных в таблице 1. Рассматриваются матрицы $\mathbf{A}$ и $\mathbf{B}$ скалярного, диагонального и полного вида, независимо друг от друга. Предложенные ниже методы позволяют решить задачу для случая матрицы $\mathbf{B}$ скалярного вида, т.е. $\mathbf{B} = \beta\mathbf{I}$. При этом рассматриваются различные виды матрицы $\mathbf{A}$.

### 1.2.1. Дополнительные требования к данным

При решении задач, к исходной выборке могут быть выдвинуты требования, связанные с природой измерения величин. При этом рекомендуется привести значения переменных к единой шкале с целью исключением шкалы и единицы измерения из дальнейшего рассмотрения [261].

**Стандартизация данных.** Выборка $\mathfrak{D}$ стандартизируется таким образом, чтобы выполнялись условия нормированности и центрированности *признаков* — столбцов $\boldsymbol{\chi}$ матрицы плана $\mathbf{X}$, а также условие центрированности вектора $\mathbf{y}$:

$$\sum_{i \in \mathcal{I}} x_{ij} = 0, \sum_{i \in \mathcal{I}} x_{ij}^2 = 1, \sum_{i \in \mathcal{I}} y_i = 0, \quad j \in \mathcal{J}, \tag{12}$$

Таблица 2. Канонические функции связи.

| Распределение | Функция связи | Вид функции |
|---|---|---|
| Нормальное | Тождественная | $\boldsymbol{\mu} = \mathbf{X}\mathbf{w}$ |
| Экспоненциальное, гамма | Мультипликативная обратная | $\boldsymbol{\mu} = (\mathbf{X}\mathbf{w})^{-1}$ |
| Обратное нормальное | Обратная квадратичная | $\boldsymbol{\mu} = (\mathbf{X}\mathbf{w})^{-\frac{1}{2}}$ |
| Пуассоновское | Логарифмическая | $\boldsymbol{\mu} = \exp(\mathbf{X}\mathbf{w})$ |
| Биномиальное, мультиномиальное | Логит-функция | $\boldsymbol{\mu} = \big(1 + \exp(-\mathbf{X}\mathbf{w})\big)^{-1}$ |

или, в векторных обозначениях,

$$\|\boldsymbol{\chi}_j\|_1 = 0, \quad \|\boldsymbol{\chi}_j\|^2 = 1, \quad \|\mathbf{y}\|_1 = 0.$$

Предполагается, что векторы $\boldsymbol{\chi}_j, \boldsymbol{\chi}_k$ линейно независимы для всех значений $j, k \in \mathcal{J}$, $j \neq k$. Линейно зависимые векторы исключаются из дальнейшего рассмотрения.

**Разбиение скользящего контроля.** Дополнительно может быть задано разбиение множества индексов $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$ элементов выборки $\mathfrak{D}$ на обучающее и контрольное подмножества. Для каждой выборки, рассматриваемой при решении задач регрессионного анализа, наборы индексов $\mathcal{L}$, $\mathcal{C}$ определены до начала эксперимента.

### 1.2.2. Экспоненциальное семейство

Далее предполагается, что распределение зависимой переменной принадлежит экспоненциальному семейству. *Экспоненциальное семейство распределений* [39] многомерной случайной величины $\mathbf{y}$ с вектором параметров $\boldsymbol{\eta}$ задается набором распределений вида

$$p(\mathbf{y}|\boldsymbol{\eta}) = h(\mathbf{y})g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{y})\right). \tag{13}$$

Вектор $\boldsymbol{\eta}$ называется вектором естественных параметров распределения. Сомножитель $\mathbf{u}(\mathbf{y})$ — некоторая вектор-функция многомерной случайной величины $\mathbf{y}$.

Гипотеза принадлежности распределения зависимой переменной экспоненциальному семейству используется при построении обобщенных линейных моделей (4), то есть, моделей вида

$$\mathsf{E}(\mathbf{y}) = \boldsymbol{\mu}(\mathbf{X}\mathbf{w}), \tag{14}$$

для которых $\boldsymbol{\mu}$ — функция связи, а $\mathbf{X}\mathbf{w}$ — линейная комбинация признаков выборки. Канонические функции связи, соответствующие частным случаям экспоненциального семейства, представлены в таблице 2.

В [326, 177] показано, что из гипотезы о принадлежности распределения $p(\mathbf{y})$ зависимой переменной $\mathbf{y}$ каноническому экспоненциальному семейству, при использовании соответствующей этому распределению функции связи $\boldsymbol{\mu}$, следует, что параметры $\mathbf{w} \sim \mathcal{N}(\hat{\mathbf{w}}, \mathbf{A}^{-1})$ обобщенно-линейной модели распределены нормально.

При решении прикладных задач наибольший интерес представляют два распределения этого семейства: нормальное и биномиальное [69, 181]. Первое используется в задачах восстановления регрессии, когда зависимая переменная принимает значения из множества $\mathbb{R}$, второе используется в задачах классификации, когда зависимая переменная принимает значения из множества $\{0, 1\}$.

### 1.2.3. Нормальное распределение зависимой переменной

Рассмотрим нормальное распределение зависимой переменной в качестве гипотезы порождения данных при восстановлении линейной или существенно-нелинейной регрессии:

$$\mathsf{E}(y|\mathbf{x}) = f(\mathbf{w}, \mathbf{x}).$$

Для нахождения *наиболее правдоподобных параметров* модели используем метод наибольшего правдоподобия [175, 174, 123]. Пусть многомерная случайная величина $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B})$ имеет нормальное распределение

$$p(\mathbf{y}) = \frac{1}{(2\pi)^{\frac{2}{m}} \det^{\frac{1}{2}}(\mathbf{B}^{-1})} \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f})\right). \tag{15}$$

В выражении (15) часть под экспонентой $(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f})$ является квадратом расстояния Махаланобиса[195]. Матрицу  можно представить в виде разложения Холецкого [112], $\mathbf{B} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$, где $\mathbf{L}$ — диагональная нижнетреугольная матрица, либо в виде произведения $\mathbf{B} = \mathbf{U}^{\mathsf{T}}\mathbf{U}$, см. [339]. Оба разложения единственны. В общем случае не предполагается, что её элементы независимы и не коррелируют. Ковариационная матрица $\mathbf{B}$ является симметричной неотрицательно определенной матрицей. Диагональные элементы $\beta_i$ этой матрицы являются обратны значениям дисперсии элементов случайной величины $\mathbf{y}$:

$$\sigma_i^2 = \frac{1}{\beta_i}, \quad i \in \mathcal{I}.$$

Так как правая часть выражения (15) зависит от вида регрессионной модели $f$, вектора параметров $\mathbf{w}$, независимой переменной $\mathbf{x}$ и от ковариационной матрицы $\mathbf{B}$, перепишем его в виде

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \mathbf{B}, f) \stackrel{\text{def}}{=} p(\mathfrak{D}|\mathbf{w}, \beta, f) = \frac{\exp(-E_{\mathfrak{D}})}{Z_{\mathfrak{D}}(\beta)}, \tag{16}$$

где $Z_{\mathfrak{D}}$ — нормирующий коэффициент для плотности нормального распределения

$$Z_{\mathfrak{D}} = (2\pi)^{\frac{m}{2}} \det^{\frac{1}{2}}(\mathbf{B}^{-1}). \tag{17}$$

Функция ошибки, соответствующая матожиданию регрессионной модели при данной гипотезе, определена как

$$E_{\mathfrak{D}} = \frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}\mathbf{B}(\mathbf{y}-\mathbf{f}). \tag{18}$$

Рассмотрим частный случай гипотезы порождения данных: элементы вектора $\mathbf{y}$ не коррелируют и имеют одинаковую дисперсию, то есть обратная ковариационная матрица $\mathbf{B} = \beta\mathbf{I}_m$. В этом случае вид функции правдоподобия (16) упрощается: коэффициент $Z_{\mathfrak{D}}$ имеет вид

$$Z_{\mathfrak{D}}(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{m}{2}},$$

так как

$$Z_{\mathfrak{D}}(\beta) = \det{}^{\frac{1}{2}}(\sigma^2\mathbf{I})(2\pi)^{\frac{m}{2}} = \sigma^{\frac{2m}{2}}\det{}^{\frac{1}{2}}(\mathbf{I})(2\pi)^{\frac{m}{2}} = \sigma^m(2\pi)^{\frac{1}{2}} = \left(\frac{2\pi}{\beta}\right)^{\frac{m}{2}},$$

а функция ошибки равна

$$E_{\mathfrak{D}} = \frac{1}{2}\beta\sum_{i\in\mathbf{I}}\left(y_i - f(\mathbf{w},\mathbf{x}_i)\right)^2,$$

так как при

$$(\sigma^2\mathbf{I})^{-1} = \frac{\mathbf{I}}{\sigma^2} = \beta\mathbf{I}$$

справедливо выражение

$$E_{\mathfrak{D}} = \frac{1}{2}(\mathbf{y}-\mathbf{f})^{\intercal}(\beta\mathbf{I})(\mathbf{y}-\mathbf{f}) = \frac{1}{2}\beta\|\mathbf{y}-\mathbf{f}\|^2.$$

Рассмотрим вектор параметров $\mathbf{w}$ модели $f$ — многомерную случайную величину. Согласно принятой гипотезе распределения (15) зависимой переменной и теореме о функциях связи распределений [326, 177], распределение параметров $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{ML}}, \mathbf{A}^{-1})$ является нормальным с матожиданием $\mathbf{w}_{\text{ML}}$, ковариационной матрицей $\mathbf{A}^{-1}$ и имеет вид

$$p(\mathbf{w}|\mathbf{A}, f) = \frac{\exp(-E_{\mathbf{w}})}{Z_{\mathbf{w}}(\mathbf{A})}. \tag{19}$$

Выражение (19) справедливо для линейных моделей, поскольку многомерные случайные величины $\mathbf{y}$ и $\mathbf{w}$ связаны линейным отображением $\mathbf{X}$. Для существенно нелинейных моделей предполагается, что это выражение будет справедливо в окрестности $\Delta\mathbf{w}$ некоторой точки $\mathbf{w}_0$ при линеаризации

$$\mathbf{y} - \mathbf{f}(\hat{\mathbf{w}}, \mathbf{X}) = \mathbf{J}\Delta\mathbf{w}.$$

Матрица Якоби $\mathbf{J}$ — это матрица частных производных модели $f$ по элементам $w_j$ вектора параметров $\mathbf{w}$:

$$\mathbf{J} = \left[\frac{\partial f(\mathbf{w},\mathbf{x}_i)}{\partial w_j}\right], \quad \text{где} \;\; \mathbf{w} = [w_1,\ldots,w_j,\ldots,w_n]^{\intercal} \;\; \text{и} \;\; i\in\mathcal{I}, j\in\mathcal{J}.$$

Это предположение используется, например, в аппроксимации Лапласа [191, 277], согласно которой функция распределения параметров существенно нелинейной модели может быть приближена функций нормального распределения.

Нормирующий коэффициент $Z_{\mathbf{w}}(\mathbf{A})$ равен

$$Z_{\mathbf{w}}(\mathbf{A}) = (2\pi)^{\frac{n}{2}}\det{}^{\frac{1}{2}}(\mathbf{A}^{-1}), \tag{20}$$

где $n$ — число параметров модели $f$. Функция-штраф за большое значение параметров модели для принятого распределения определена как

$$E_{\mathbf{w}} = \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\intercal}\mathbf{A}(\mathbf{w}-\mathbf{w}_0). \tag{21}$$

Рассмотрим частный случай: дисперсии элементов $w_j$ вектора параметров $\mathbf{w}$ равны, обратная ковариационная матрица имеет вид $\mathbf{A} = \alpha\mathbf{I}_n$. В этом случае выражения (20) и (21) будут иметь вид

$$Z_{\mathbf{w}}(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{\frac{n}{2}} \quad \text{и} \quad E_{\mathbf{w}} = \frac{1}{2}\alpha\|\hat{\mathbf{w}}-\mathbf{w}\|^2.$$

Для нахождения *наиболее вероятных параметров* модели $f(\mathbf{w}, \mathbf{x})$ используем Байесовский вывод [44, 152, 142]. При заданной модели $f$ и заданных значениях $\mathbf{A}$ и $\mathbf{B}$ выражение (16) принимает вид

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, f) = \frac{p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, f)p(\mathbf{w}|\mathbf{A}, f)}{p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, f)}. \tag{22}$$

Элементы этого выражения и соответствующие им параметры:

$p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, f)$ — апостериорное распределение параметров,

$\mathbf{w}_{\text{MP}} = \arg\max p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, f)$ — наиболее вероятные параметры,

$\mathbf{w}_{\text{ML}} = \arg\max p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, f)$ — наиболее правдоподобные параметры,

$p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, f)$ — функция правдоподобия данных,

$p(\mathbf{w}|\mathbf{A}, f)$ — априорное распределение параметров,

$p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, f)$ — функция правдоподобия модели $f$.

Записывая функцию ошибки $S = E_{\mathbf{w}} + E_{\mathfrak{D}}$ в виде

$$S(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{ML}})^{\intercal}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{ML}}) + \frac{1}{2}(\mathbf{y} - \mathbf{f})^{\intercal}\mathbf{B}(\mathbf{y} - \mathbf{f}), \tag{23}$$

получаем вместо (22) выражение

$$p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, f) = \frac{\exp\bigl(-S(\mathbf{w})\bigr)}{Z_S},$$

где $Z_S$ — нормирующий коэффициент.

Заметим, что матожидание вектора параметров в некоторых случаях может быть гипотетически принято равным нулю, $\hat{\mathbf{w}} = 0$. Такое предположение явно или неявно принимается при решении задач выбора признаков линейных регрессионных моделей, см. Лассо [267], Stagewise [130], LARS [82], а также при прореживании некоторых нелинейных моделей [129, 176].

При рассмотрении частных случаев ковариационных матриц $\mathbf{A} = \alpha\mathbf{I}_n$ и $\mathbf{B} = \beta\mathbf{I}_m$, параметров модели $\mathbf{w}$ и гомоскедастичной зависимой переменной $\mathbf{y}$ выражение (22) принимает вид

$$p(\mathbf{w}|\mathfrak{D}, \alpha, \beta, f) = \frac{p(\mathfrak{D}|\mathbf{w}, \beta, f)p(\mathbf{w}|\alpha, f)}{p(\mathfrak{D}|\alpha, \beta, f)}.$$

а функция ошибки —

$$S(\mathbf{w}) = \frac{1}{2}\alpha\|\mathbf{w}\|^2 + \frac{1}{2}\beta\|\mathbf{y} - \mathbf{f}\|^2.$$

Параметры $\alpha$ и $\beta$ в последнем выражении играют роль регуляризирующих множителей [50, 265].

### 1.2.4. Биномиальное распределение зависимой переменной

В данной работе задача классификации рассматривается как задача логистической регрессии [164, 141, 177], то есть, как частный случай задачи восстановления регрессии (1). В данном случае регрессия и классификация различаются не более чем предположением о распределении зависимой переменной $\mathbf{y}$. При этом принимается гипотеза о биномиальном распределении зависимой переменной $y \in \{0, 1\}$:

$$y \sim \mathbf{B}(P, 1-P), \tag{24}$$

случайная величина $y$ принимает значение 0 с вероятностью $P$ и значение 1 с вероятностью $1-P$. Таким образом, функция регрессии $f(\hat{\mathbf{w}}, \mathbf{x})$ в случае биномиального распределения зависимой переменной восстанавливает регрессию, равную вероятности принадлежности вектора $\mathbf{x}$ к одному из двух классов. Функция правдоподобия реализаций многомерной случайной величины $\mathbf{y}$ имеет вид

$$p(\mathbf{y}|\mathbf{w}) = \prod_{i=1}^{m} f(\mathbf{w}, \mathbf{x}_i)^{y_i} \big(1 - f(\mathbf{w}, \mathbf{x}_i)\big)^{1-y_i}. \tag{25}$$

Модель логистической регрессии имеет вид

$$\mathbf{f}(\mathbf{w}, \mathbf{X}) = \boldsymbol{\sigma}(-\mathbf{X}\mathbf{w}) = \frac{1}{1 + \exp(-\mathbf{X}\mathbf{w})}. \tag{26}$$

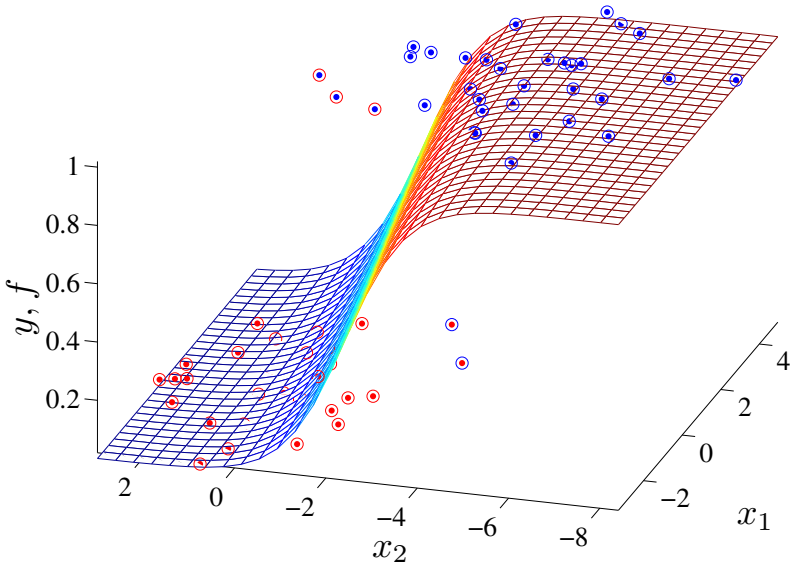



Рис. 4. Логистическая регрессия от двух переменных.

Выражение (25) при гипотезе (24) имеет вид

$$E_D(\mathbf{w}) = -\sum_{i=1}^{m} \big(y_i \ln f(\mathbf{w}, \mathbf{x}_i) + (1 - y_i) \ln(1 - f(\mathbf{w}, \mathbf{x}_i))\big), \tag{27}$$

а производная этой функции по параметрам равна

$$\nabla E_D(\mathbf{w}) = \sum_{i=1}^{m} (f(\mathbf{w}, \mathbf{x}_i) - y_i)\mathbf{x}_i.$$

где $\sigma(\mathbf{X}, \mathbf{w})$ — сигмоидная функция. Перепишем функцию ошибки общего вида (23) для гипотезы (25) в виде

$$S(\mathbf{w}) = \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{ML}})^{\intercal}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{ML}}) - \sum_{i=1}^{m}\big(y_i \ln f(\mathbf{w}, \mathbf{x}_i) + (1 - y_i)\ln(1 - f(\mathbf{w}, \mathbf{x}_i))\big).$$

На рисунке 4 синими и красными точками показана выборка, состоящая из двух классов. Восстановленная регрессия показана поверхностью $f(\hat{\mathbf{w}}, \mathbf{x})$, значение которой равно вероятности принадлежности элемента выборки $\mathbf{x}$ к одному из двух классов. Прогнозируемая принадлежность показана окружностями. При несовпадении цветов точки и окружности следует говорить об ошибке классификации.

### 1.2.5. Функция ошибки и гипотеза порождения данных

Задача восстановления регрессии есть задача нахождения условного матожидания. Для оценки качества решения этой задачи вводится *функция ошибки*, исходя из значения которой делается вывод о том, насколько хорошо модель приближает данные, а также насколько адекватна гипотеза порождения данных [80, 289, 293, 131].

**Определение 7.** *Функция ошибки $S(\mathbf{w})$ — функция, значение которой требуется минимизировать для получения оценок параметров $\mathbf{w}$ модели $f$, удовлетворяющих заданным требованиям.*

Например, это одно из следующих требований:

требование максимизации правдоподобия данных,

требование максимизации вероятности параметров модели,

требование максимизации правдоподобия самой регрессионной модели,

требования состоятельности, несмещенности оценки параметров,

либо другие требования, определяемые решаемой задачей восстановления регрессии.

Функция ошибки назначается одним из двух способов:

1) путем байесовского вывода, тогда она определяется гипотезой порождения данных и, опционально, принятой регрессионной моделью (выше приведены примеры (23) и (27) для гипотез нормального и биномиального распределения зависимой переменной);

2) другим путем, который учитывает особенности постановки решаемой прикладной задачи.

Во втором случае несмещенность и другие статистические свойства полученных оценок параметров не исследуются. В качестве примера приведем функции ошибки, обеспечивающие выполнение требований промышленных стандартов [1] и требований к минимизации потерь при совершении торговых операций [311]. В первом примере симметричная функция ошибки имеет вид

$$S(\mathbf{w}) = \sum_{i \in \mathcal{I}} \|f(\mathbf{w}, \mathbf{x}_i) - y_i\|_1. \tag{28}$$

Во втором примере несимметричная функция ошибки, являющаяся кусочно-линейной функцией, имеет вид

$$S(\mathbf{w}) = \sum_{i\in\mathcal{I}}\sum_{s}\begin{cases} a_s + b_s(f(\mathbf{w},\mathbf{x}_i) - y_i), \text{ при } f \in (z_{s-1}, z_s); \\ 0, \text{ в противном случае.} \end{cases} \tag{29}$$

Параметры $a_s, b_s$ и концы отрезков $z_s$ выбираются согласно расчетам убытков при совершении торговых операций при условии непрерывности функции ошибки и её первой производной. В обоих случаях происходит отказ от принятия гипотезы порождения данных, и функция ошибки $S(\mathbf{w})$ оптимизируется, исходя из условий поставленной задачи. Например, при восстановлении регрессии измерений некоторых физических величин используется метод наименьших модулей [330, 351, 350], согласно которому функция ошибки задана как сумма модулей регрессионных остатков. Задача нахождения минимального значения функций вида (28) или (29) решается методами линейного программирования. В таблице 3 приведен набор функций ошибок, часто используемых при решении задач прогнозирования.

Таблица 3. Функции ошибок регрессионных моделей.

| | |
|---|---|
| Среднее арифметическое модулей остатков | $\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}\lvert\varepsilon_i\rvert$ |
| Среднее арифметическое модулей относительных остатков | $\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\lvert\frac{\varepsilon_i}{y_i}\rvert$ |
| Среднее отклонение модулей остатков | $\text{PMAD} = \sum_{i=1}^{m}\lvert\varepsilon_i\rvert\,(\sum_{i=1}^{m}\lvert y_i\rvert)^{-1}$ |
| Среднеквадратичная ошибка | $\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}\varepsilon_i^2$ |
| Корень среднеквадратичной ошибки | $\text{RMSE} = \frac{1}{\sqrt{m}}\sqrt{\sum_{i=1}^{m}\varepsilon_i^2}$ |
| Сила прогноза | $\text{SS} = 1 - \frac{\text{MSE}_{\text{forecast}}}{\text{MSE}_{\text{history}}}$ |

**Функция ошибки и разбиение выборки.** В данной работе не предполагается, что для оценки наиболее вероятных параметров модели, либо для выбора наиболее правдоподобной модели из некоторого множества требуется разбиение множества индексов $\mathcal{I}$ элементов выборки $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I}$ на обучающую и контрольную: $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$. Тем не менее, следует отметить, что при выборе моделей такое разбиение является одним из наиболее эффективных способов избежать переобучения, см. [295, 294, 192]. Поэтому ниже приведен ряд примеров эвристических функций ошибок, предложенных авторами метода группового учета аргументов. Данные функции называются критериями. Значительная их часть опубликована на сайте [2].

Используемые в этом подразделе обозначения $\mathbf{X}_\mathcal{C}, \mathbf{y}_\mathcal{C}, \mathbf{w}_\mathcal{L}$ означают, что значения переменных $\mathbf{X}, \mathbf{y}, \mathbf{w}$ фиксированы, в выборку $(\mathbf{X}_\mathcal{C}, \mathbf{y}_\mathcal{C})$ вошли только объекты с индексами из множества $\mathcal{C} \in \mathcal{I} \neq \emptyset$, а оцека вектора параметров $\mathbf{w}_\mathcal{L}$ получена с использованием выборки, состоящей из элементов с индексами из множества $\mathcal{L} \subset \mathcal{I} \neq \emptyset$:

$$\mathbf{w}_\mathcal{L} = \arg\min_{i\in\mathcal{L}\subset\mathcal{I}} S(\mathbf{w}|\mathbf{X}_\mathcal{L}, \mathbf{y}_\mathcal{L}, f).$$

При этом считается, что множество индексов $\mathcal{I}$ элементов выборки разбито на подмножества

$$\mathcal{I} = \mathcal{L} \sqcup \mathcal{C} \sqcup \mathcal{V},$$

в котором $\mathcal{L}$ — обучающая выборка, $\mathcal{C}$ — котрольная выборка, $\mathcal{V}$ — валидационная выборка. Последняя в ряде задач может быть пустой.

Метод группового учета аргументов [307, 308, 309] использует внутренний и внешний критерий, так как при оценке параметров моделей и при выборе моделей используются разные элементы выборки. *Внутренний критерий* используется для оценки параметров: их значения оцениваются на подвыборке элементов с индексами из $\mathcal{L}$. Выбор моделей производится с помощью *внешнего критерия*, значение которого вычисляется на множестве $\mathcal{C}$. При выборе минимум внешнего критерия означает, что модель, доставляющая такой минимум, является искомой.

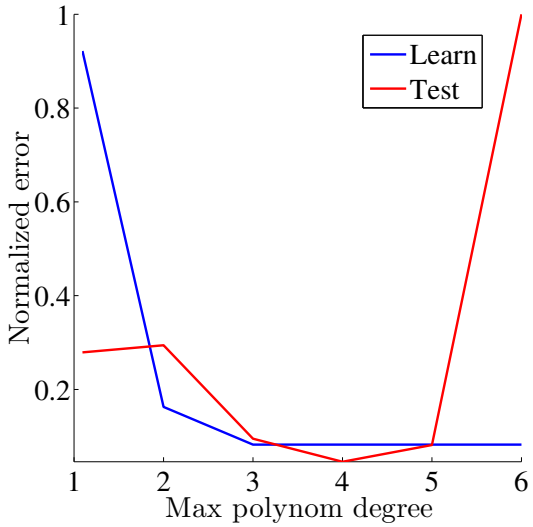



Рис. 5. Внешний и внутренний критерии при различных значениях структурной сложности.

*Критерий регулярности* $S_{\Delta^2}$ равен норме разности вектора значений зависимой переменной и вектора значений функции регрессии на тестовой подвыборке $\mathcal{C}$ при параметрах, оцененных на обучающей подвыборке $\mathcal{L}$.

$$S_{\Delta^2_{\mathcal{C}}} = \|\mathbf{y}_{\mathcal{C}} - \mathbf{X}_{\mathcal{C}}\mathbf{w}_{\mathcal{L}}\|^2,$$

где

$$\mathbf{w}_{\mathcal{L}} = (\mathbf{X}_{\mathcal{L}}^{\mathsf{T}}\mathbf{X}_{\mathcal{L}})^{-1}(\mathbf{X}_{\mathcal{L}}^{\mathsf{T}}\mathbf{y}_{\mathcal{L}}).$$

Этот критерий может быть нормирован выражениями $\|\mathbf{y}_{\mathcal{L}}\|^2$ или $\|\mathbf{y}_{\mathcal{L}} - \mathrm{mean}(\mathbf{y}_{\mathcal{L}})\|^2$.

*Критерий предсказательной способности* — модификация критерия регулярности для задач прогнозирования. Этот критерий включает среднеквадратичную ошибку для валидационной выборки $\mathcal{V}$, которая не используется ни при оценке параметров, ни при выборе модели. В этом случае выборка делится на три части. Критерий предсказательной способности имеет вид

$$S_{\Delta^2_{\mathcal{V}}} = \frac{\|\mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{V}}\|^2}{\|\mathbf{y}_{\mathcal{I}} - \mathrm{mean}(\mathbf{y}_{\mathcal{I}})\|^2}.$$

*Критерий минимального смещения* или *критерий непротиворечивости*: модель, которая имеет на обучающей и на контрольной выборках различные векторы невязок, называется противоречивой. Критерий задан разностью между значениями функции регрессии, вычисленными на двух различных выборках, заданных множествами $\mathcal{L}$ и $\mathcal{C}$ и требует, чтобы оценки параметров, вычисленные на этих выборках, различались минимально. Он имеет вид:

$$S_{\eta^2_{\text{bs}}} = \|\mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{C}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{L}}\|^2,$$

модификация:

$$S_{\eta^2_{\text{a}}} = \|\mathbf{w}_{\mathcal{C}} - \mathbf{w}_{\mathcal{L}}\|^2,$$

где $\mathbf{w}_{\mathcal{C}}$ и $\mathbf{w}_{\mathcal{L}}$ — векторы параметров, полученные с использованием подвыборок $\mathcal{C}$ и $\mathcal{L}$.

*Критерий иммунитета к шуму* имеет вид

$$\begin{aligned} S_{V^2} =& (\mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{C}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{I}})^{\intercal}(\mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{L}}) = \\ & (\mathbf{w}_{\mathcal{C}} - \mathbf{w}_{\mathcal{I}})^{\intercal}\mathbf{X}_{\mathcal{I}}^{\intercal}\mathbf{X}_{\mathcal{I}}(\mathbf{w}_{\mathcal{I}} - \mathbf{w}_{\mathcal{L}}), \end{aligned}$$

где $\mathbf{w}_{\mathcal{I}}$ — вектор параметров, полученный с использованием полной выборке $\mathcal{I}$. Утверждается [307], что с помощью этого критерия в сильно зашумленных данных можно найти скрытые закономерности.

*Комбинированный критерий* позволяет использовать при выборе моделей линейную комбинацию нескольких критериев. Комбинированный критерий

$$S_{\kappa^2} = \sum_{k=1}^{K} v_k S_k, \quad \text{при условии} \quad \sum_{k=1}^{K} v_k = 1.$$

Здесь $S_k$ — принятые на рассмотрение критерии, а $v_k$ — веса этих критериев, назначенные в начале вычислительного эксперимента.

Используются также нормализованные значения критериев. При этом предыдущая формула имеет вид

$$S_{\kappa^2} = \sum_{i=1}^{K} v_k \frac{S_k}{\max\limits_{f \in \mathfrak{F}}(S_k)}.$$

Максимальное значение критерия $\max(S_k)$ берется по вычисленным значениям критериев $S_k(f)$ для всех порожденных моделей $f \in \mathfrak{F}$.

## 1.3. Задачи регрессионного анализа

Задача восстановления регрессии (1) имеет несколько разных постановок, каждую их которых можно условно отнести к одному из следующих типов:

1) задачи оценки параметров модели,

2) задачи выбора признаков или объектов регрессионной выборки,

3) задачи выбора регрессионных моделей,

4) задачи проверки гипотезы порождения данных.

Предполагается, что функция ошибки $S(\mathbf{w})$ задана гипотезой порождения данных. При задании функции ошибки используется байесовский вывод. Предполагается, что зависимая переменная имеет распределение из экспоненциального семейства (13), например, нормальное (15) или биномиальное (24). В гипотезу также включены условия взаимозависимости элементов целевого вектора $\mathbf{y}$ как многомерной случайной величины, например, условие гомоскедастичности (6), независимости (7) или гетероскедастичность (8) — как отсутствие этих условий.

Функция ошибки может быть также определена исходя из постановки задачи, например, в случаях (28) или (29), когда её вид определен задачей минимизации потерь.

### 1.3.1. Оценка параметров модели

**Задача 1.** *Задана выборка* $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$, *функция ошибки модели* $S$ *и модель — параметрическое семейство функций* $f(\mathbf{w}, \mathbf{x})$. *Требуется найти такие параметры* $\mathbf{w}$ *модели, которые бы доставляли минимум функции ошибки*

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\mathfrak{D}, f). \tag{30}$$

В выражении (30) справа от вертикальной черты указаны фиксированные значения переменных, что читается: «при заданной выборке $\mathfrak{D}$ и модели $f$», аналогично обозначению, принятому для записи условной вероятности. Далее предполагается, что запись $S(\mathbf{w})$ эквивалентна записи $S(\mathbf{w}|\mathfrak{D}, f)$, если специально не оговорено иное.

Например, задан критерий качества линейной регрессионной модели при предположениях (15) и (6). Также предполагается, что выполнены условия (37). Требуется найти параметры — весовые коэффициенты, доставляющие минимальное значение функции ошибки — квадрату евклидовой нормы вектора невязок $S(\mathbf{w}) = \sum_{i\in\mathcal{I}}\big(y_i - f(\mathbf{w}, \mathbf{x}_i)\big)^2 \to \min$. Ряд примеров, где $f$ — фиксированная нелинейная регрессионная модель рассмотрен в [243].

Функция ошибки, определенная посредством логарифмической функции правдоподобия, как в (16) и (17),

$$S(\mathbf{w}) = E_{\mathfrak{D}} = -\ln\big(p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, f)\big),$$

обеспечивает максимизацию правдоподобия параметров. Параметры, найденные минимизацией этой функции ошибок, называются наиболее правдоподобными, а задача (30) имеет вид

$$\mathbf{w}_{\text{ML}} = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\mathfrak{D}, \mathbf{B}, f).$$

Параметры, найденные минимизацией функции ошибок, заданной апостериорным распределением (22), называются наиболее вероятными, а задача (30) имеет вид

$$\mathbf{w}_{\text{MP}} = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, f).$$

При этом предполагается, что обратные ковариационные матрицы $\mathbf{A}, \mathbf{B}$ заданы, или оценивание уже выполнено.

### 1.3.2. Выбор оптимальной модели

Для выбора модели из некоторого множества допустимых моделей требуется ввести понятие *сложности модели* [118]. Различают следующие типы сложности:

1) обобщающая способность модели, оцениваемую с использованием скользящего контроля [192];

2) статистическая сложность модели, например, Mallow's $C_p$ [197], AIC [15], BIC [43]

3) минимальная длина описания модели, оцениваемая с использованием оценок правдоподобия модели [262, 264],

4) структурная сложность модели [273, 272, 106], зависящая от вида суперпозиции элементарных функций, которые задают модель.

Связано это с тем, что при выборе модели без учета сложности [266], будет выбрана наиболее сложная модель. Поставим задачу выбора моделей для случая обобщенно-линейных моделей. При этом число параметров и число признаков модели будут равны (равенство числа признаков и параметров может не быть в случае нелинейных моделей). Задача выбора модели тогда сводится к выбору признаков, то есть к поиску такого множества индексов признаков $\mathcal{A} \subseteq \mathcal{J}$, которое доставит оптимальное значение критерию сложности модели.

При использовании скользящего контроля, критерии которого описаны в предыдущем разделе, задача выбора модели ставится следующим образом.

**Задача 2.** *Задана выборка* $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$, *где множество векторов свободных переменных* $\{\mathbf{x} = [x_1, \ldots, x_j, \ldots, x_n]\}$, *проиндексировано* $j \in \mathcal{J} = \{1, \ldots, n\}$. *Задано разбиение скользящего контроля множества индексов элементов выборки* $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$. *Задана функция ошибки* $S$ *и модель — параметрическое семейство функций* $f(\mathbf{w}, \mathbf{x}) = \mu(\mathbf{w}^\mathsf{T}\mathbf{x})$, *где* $\mu$ *— функция связи* (14). *Требуется найти такое подмножество индексов* $\mathcal{A} \subseteq \mathcal{J}$, *которое бы доставляло минимум функции:*

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(f_\mathcal{A} | \hat{\mathbf{w}}, \mathfrak{D}_\mathcal{C}) \tag{31}$$

*на подмножестве* $\mathfrak{D}_\mathcal{C}$ *разбиении выборки* $\mathfrak{D}$, *определенном множеством индексов* $\mathcal{C}$. *При этом параметры* $\hat{\mathbf{w}}$ *модели должны доставлять минимум функции:*

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w} | \mathfrak{D}_\mathcal{L}, f_\mathcal{A}) \tag{32}$$

*на подмножестве* $\mathfrak{D}_\mathcal{L}$ *разбиении выборки, определенном множеством индексов* $\mathcal{L}$. *Здесь* $f_\mathcal{A}$ *обозначает обобщенно-линейную модель* $f = \mu(\mathbf{w}_\mathcal{A}^\mathsf{T}\mathbf{x}_\mathcal{A})$, *включающую только столбцы матрицы* $\mathbf{X}$ *с индексами из множества* $\mathcal{A}$.

Нелинейная модель не может быть однозначно задана множеством $\mathcal{A}$ активных признаков. Поэтому для задания модели используются правила индуктивного порождения моделей детально определенные в следующем разделе. Они позволяют однозначно индексировать модели $f$ из множества моделей $\mathcal{F}$.

**Задача 3.** *Задана выборка* $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$ *и разбиение скользящего контроля* $i \in \mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$. *Задано множество порождающих функций* $\mathcal{G} = \{g_1, \ldots, g_n\}$. *Заданы правила индуктивного порождения множества моделей* $\mathfrak{F} = \{f_r\}$, *индексированных счетным множеством* $\mathcal{R} \ni r$. *Требуется найти такую модель* $f_{\hat{r}}$, *которая бы доставляла минимум функции*

$$\hat{r} = \arg\min_{r \in \mathcal{R}} S(f_r|\hat{\mathbf{w}}, \mathfrak{D}_{\mathcal{C}}) \tag{33}$$

*при условии оценки оптимальных параметров* $\hat{\mathbf{w}}$ *решением задачи* (32).

### 1.3.3. Оценка ковариационных матриц зависимой переменной и параметров

**Задача 4.** *Задана выборка* $\mathfrak{D}$, *гипотеза порождения данных* $H$, *соответствующая ей функция ошибки* $S(\mathbf{w})$ *и модель* $f(\mathbf{w}, \mathbf{x})$. *Задан вектор* $\hat{\mathbf{w}}$ *оптимальных параметров модели. Требуется оценить обратные ковариационные матрицы* $\mathbf{A}, \mathbf{B}$ *для случаев* (6)–(11), *максимизируя правдоподобие модели:*

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \arg\max_{\mathbf{A} \in \mathbb{R}^{n^2}, \mathbf{B} \in \mathbb{R}^{m^2}} \int p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, f) p(\mathbf{w}|\mathbf{A}, f) d\mathbf{w}.$$

### 1.3.4. Совместный выбор объектов и признаков в обобщенно-линейных моделях

Обобщенно-линейная модель $f$ однозначно задается активным множеством индексов признаков $\mathcal{A} \subseteq \mathcal{J}$. Функция связи задана гипотезой порождения данных. Предполагая частичную гомоскедастичность выборки (например, среди объектов встречаются выбросы, которые должны быть исключены из рассмотрения), зададим «фильтрованную» выборку (другими словами — активное множество объектов) индексами из множества $\mathcal{B} \subseteq \mathcal{I}$. Обозначим множество векторов $\{\mathbf{x}_i | i \in \mathcal{B}\}$ как $\mathbf{x}_{\mathcal{B}}$. Задача выбора модели имеет вид

$$(\hat{\mathcal{A}}, \hat{\mathcal{B}}) = \arg\max_{\mathcal{A} \subseteq \mathcal{J}, \mathcal{B} \subseteq \mathcal{I}} \mathcal{E}\big(f_{\mathcal{A}}(\mathbf{w}, \mathbf{x}_{\mathcal{B}})\big), \tag{34}$$

где функция правдоподобия модели равна интегралу по пространству её параметров

$$\mathcal{E}(f|\mathfrak{D}) = \int\limits_{\mathbf{w} \in \mathbb{R}^n} p(\mathfrak{D}|\mathbf{w}, \hat{\mathbf{B}}, f_{\mathcal{A}}) p(\mathbf{w}|\hat{\mathbf{A}}, f) d\mathbf{w}. \tag{35}$$

Подынтегральное произведение включает функцию правдоподобия данных и априорное распределение параметров модели.

### 1.3.5. Выбор наиболее правдоподобной модели

Обобщим предыдущую задачу на случай существенно нелинейных моделей. При этом для простоты будем считать, что элементы-выбросы уже исключены из регрессионной выборки, то есть, $\mathcal{I} = \mathcal{B}$. Тогда задача выбора правдоподобной модели $f_r$ с индексом $r$ из множества моделей-претендентов $\mathcal{F}$ имеет следующий вид.

**Задача 5.** *Задана выборка* $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I}$. *Задано множество моделей* $\mathcal{F} = \{f_r\}$, *индексированных счетным множеством* $\mathcal{R} \ni r$. *Требуется выбрать наиболее правдоподобную модель*

$$\hat{r} = \arg\max_{r\in\mathcal{R}} p(f_r|\mathfrak{D}) = \arg\max_{r\in\mathcal{R}} \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \hat{\mathbf{B}}, f_r)p(\mathbf{w}|\mathfrak{D}, \hat{\mathbf{A}}, f_r)d\mathbf{w}$$

Заметим, что оценивать параметры модели для того, чтобы выбрать наиболее правдоподобную модель, необязательно.

При предположении о равенстве априорных вероятностей моделей:

$$p(f_i) = p(f_j), i, j \in \mathcal{R}, \quad \text{и} \sum_{i\in\mathcal{R}} p(f_i) = 1,$$

задача выбора наиболее правдоподобной модели становится эквивалентной задаче выбора наиболее вероятной модели. Задача оценивания апостериорного распределения или априорной функции вероятности моделей в данной работе не рассматривается.

### 1.3.6. Выбор смеси моделей

В случае, когда невозможно получить адекватную функцию регрессии в связи со сложностью регрессионной выборки, решается задача восстановления регрессии с помощью нескольких регрессионных моделей. При этом некоторому элементу выборки может быть поставлена в соответствие либо одна, либо несколько моделей.

**Определение 8.** *Многоуровневой моделью* $\mathfrak{f}$ *называется набор моделей* $\mathfrak{f} = \{f_k(\mathbf{w}_k, \mathbf{x}_{\mathcal{B}_k})|f \in \mathfrak{F}\}$, $k = 1, \ldots, K$, *такой, что*

$$f_k : \mathbb{W}_k \times \mathbb{X}_{\mathcal{B}_k} \to \mathcal{Y}_{\mathcal{B}_k},$$

*при разбиении*

$$\mathcal{I} = \bigsqcup_{k\in\{1,\ldots,K\}} \mathcal{B}_k.$$

**Задача 6.** *Задана выборка* $\mathfrak{D}$, *гипотеза порождения данных* $H$, *набор моделей* $\mathfrak{F} \ni f_k$ *и распределения* $p(\mathfrak{D}|\mathbf{w}, \mathbf{B}, f_k), p(\mathbf{w}|\mathbf{A}, f_k)$. *Требуется найти разбиение* $\mathcal{I} = \bigsqcup_{k\in\{1,\ldots,K\}} \mathcal{B}_k$, *которое доставляло бы максимум произведению функций правдоподобия соответствующих моделей:*

$$(\hat{\mathcal{B}}_1, \ldots, \hat{\mathcal{B}}_K) = \arg\max_{\mathcal{B}_1\sqcup\cdots\sqcup\mathcal{B}_K=\mathcal{I}} \prod_{k\in\{1,\ldots,K\}} \mathcal{E}(f_k|\mathfrak{D}).$$

*Правдоподобие модели определено в* (35).

Частным случаем задачи разбиения выборки на несколько подмножеств является задача выбора опорных объектов [251]. При постановке этой задачи множество индексов разбивается на два подмножества, $\mathcal{B}_1 \sqcup \mathcal{B}_0 = \mathcal{I}$, причем модель определяется только на выборке с индексами объектов $\mathcal{B}_1$. Эти объекты считаются опорными. Объекты с индексами $\mathcal{B}_0$ при решении задачи не рассматриваются.

### 1.3.7. Нахождение инвариантов моделей

Данная задача возникает при прогнозировании квазипериодических временных рядов. При нахождении многоуровневой модели может возникнуть ситуация, когда две или несколько моделей оказываются «похожими». Предлагается сократить число моделей за счет объединения элементов выборки, которые к ним относятся, иначе — найти функцию $h : i, j \to k,$ для некоторых $i, j \in \{1, \ldots, K\},$ таких, что $\mathcal{B}_k = \mathcal{B}_i \cup \mathcal{B}_j \to \mathcal{B}$. В качестве функции расстояния между моделями предлагается использовать расстояние Дженсена-Шеннона, вычисляемое как расстояние между апостериорными распределениями моделей.

**Задача 7.** *Задано разбиение выборки $\mathcal{I} = \hat{\mathcal{B}}_1 \sqcup \cdots \sqcup \hat{\mathcal{B}}_K$. Требуется найти функцию, отображающую $\{1, \ldots, K\} \to \{1, \ldots, P\}$.*

### 1.3.8. Проверка гипотезы порождения данных

Выше предполагалось, что гипотеза порождения данных определяет функцию ошибки и, в конечном итоге, выбранную модель. Однако, после получения оптимальной модели необходимо выполнить анализ регрессионных остатков, целью которого является возможное отвержение принятой ранее гипотезы.

**Задача 8.** *Заданы выборка $\mathfrak{D}$ и функция регрессии $f(\hat{\mathbf{w}}, \mathbf{x})$, то есть, задан вектор регрессионных остатков $\boldsymbol{\varepsilon} = \mathbf{f} - \mathbf{y}$. Задан набор гипотез порождения данных $\mathcal{H} = \{H(\boldsymbol{\theta})\}$ — функций распределения многомерной случайной величины $\mathbf{y}$. Требуется оценить параметры каждой функции распределения и выбрать наиболее адекватную гипотезу порождения данных.*

Оценки параметров при выборе гипотезы порождения данных должны быть несмещенные и состоятельные [53]. Оценка $\hat{\theta}(x_1, \ldots, x_n)$ параметра $\theta$ называется *несмещенной*, если $\mathbb{E}\hat{\theta}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = \theta$ для всех $\theta \in \Theta$. Смещением называется разность

$$b(\theta) = \mathbb{E}\hat{\theta} - \theta.$$

Оценка $\hat{\theta}(x_1, \ldots, x_n)$ параметра $\theta$ называется *состоятельной*, если для всех $\theta \in \Theta$ последовательность

$$\hat{\theta}_n = \hat{\theta}(\mathbf{X}_1, \ldots, \mathbf{X}_n) \xrightarrow{P} \theta \quad \text{при } n \to \infty.$$

Символ $\xrightarrow{P}$ означает сходимость по вероятности: для любого $\epsilon > 0$ вероятность $P(|\hat{\theta}_n - \theta| > \epsilon) \to 0$ при $n \to \infty$. Предлагаемый состав методов анализа регрессионных остатков приведен ниже.

Математические объекты, упомянутые в вышеперечисленных задачах сведены в таблицу 4. Знак прочерка «–» означает, что соответствующий объект в задаче не используется, либо вычисляется в ходе решения и используется как промежуточный результат.

## 1.4. Оценка параметров моделей

Оценка параметров моделей является одной из наиболее часто решаемых задач регрессионного анализа. Способ получения оптимального решения $\mathbf{w}_0$ может определяться видом оптимизируемой функции [54, 51, 238] ошибок $S(\mathbf{w})$, но также решение может быть получено

Таблица 4. Сводная таблица задач, решаемых при восстановлении регрессии.

| | Задача | Модель $f$ | Параметры $\mathbf{w}$ | Гиперпараметры $\mathbf{A}, \mathbf{B}$ | Признаки $\mathcal{A}$ | Объекты $\mathcal{B}$ |
|---|---|---|---|---|---|---|
| 1 | Оценка параметров | задана | найти | – | заданы | заданы |
| 2 | Выбор оптимальной модели | найти | заданы | – | заданы | заданы |
| 3 | Оценка ковариационных матриц | задана | – | найти | заданы | заданы |
| 4 | Выбор объектов и признаков | задана | – | заданы | найти подмножество | найти подмножество |
| 5 | Выбор правдоподобной модели | найти | – | заданы | заданы | заданы |
| 6 | Выбор смеси моделей | задана | – | заданы | заданы | найти покрытие |
| 7 | Нахождение инвариантов | задана | – | заданы | заданы | найти разбиение |
| 8 | Оценка мощности выборки | задана | заданы | заданы | заданы | оценить число |
| 9 | Проверка гипотезы порождения данных | задана | заданы | – | – | заданы |

с помощью стохастических оптимизационных алгоритмов [283]. Последние называют также алгоритмами глобальной оптимизации, и, как показано в [280, 278, 279], эти алгоритмы могут быть более эффективны, чем алгоритмы, использующие градиентный спуск, особенно в случаях большого числа локальных экстремумов функции ошибки. На практике применяется комбинация этих алгоритмов. Далее будут рассмотрены модификации таких алгоритмов, как алгоритм Левенберга-Марквардта [179] и алгоритм регионов доверия [71].

### 1.4.1. Линейные модели

Нахождение параметров $\mathbf{w}$ линейной модели (3) при предположении о нормальном распределении (15) зависимой переменной $\mathbf{y}$ заключается в минимизации евклидовой нормы вектора регрессионных остатков

$$S(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = \|\boldsymbol{\varepsilon}\|^2. \tag{36}$$

Предполагается выполнение следующих условий: (37)

1) независимые переменные $\mathbf{x}$ не являются случайными величинами,

2) математическое ожидание $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$,

3) дисперсия $\mathsf{D}(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \mathbf{I}$ (условие гомоскедастичности),

4) при $i \neq k$ математическое ожидание $\mathsf{E}(\varepsilon_i, \varepsilon_k) = 0$,

5) $\operatorname{rank}(\mathbf{X}) = n \leqslant m$.

Эти условия называются условиями Гаусса-Маркова [301]. При этом оценки параметров модели (3) являются состоятельными и несмещенными. Оценки являются также эффективными, если $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\varepsilon}^2 \mathbf{I})$.

Требуется минимизировать евклидово расстояние от вектора $\mathbf{y}$ до вектора $\mathbf{Xw}$. Этот вектор лежит в пространстве столбцов матрицы $\mathbf{X}$, так как $\mathbf{Xw}$ — это линейная комбинация столбцов этой матрицы с коэффициентами $w_1, \ldots, w_n$. Задача оценки $\mathbf{w}$ эквивалентна задаче нахождения точки $\mathbf{p} = \mathbf{Xw}$, ближайшей к $\mathbf{y}$ и находящейся в пространстве столбцов матрицы $\mathbf{X}$. Следовательно, вектор $\mathbf{p}$ должен быть проекцией $\mathbf{y}$ на пространство столбцов, вектор регрессионных остатков $\mathbf{Xw} - \mathbf{y}$ должен быть ортогонален этому пространству. Рассмотрим произвольный вектор $\mathbf{Xv}$, ортогональный вектору регрессионных остатков $\mathbf{Xw} - \mathbf{y}$:

$$(\mathbf{Xv})^{\intercal}(\mathbf{Xw} - \mathbf{y}) = \mathbf{v}^{\intercal}(\mathbf{X}^{\intercal}\mathbf{Xw} - \mathbf{X}^{\intercal}\mathbf{y}) = 0.$$

Так как это равенство должно быть справедливо для произвольного вектора $\mathbf{v}$, то $\mathbf{X}^{\intercal}\mathbf{Xw} - \mathbf{X}^{\intercal}\mathbf{y} = 0$, см. рис. 6. Если столбцы матрицы $\mathbf{X}$ линейно независимы, то матрица $\mathbf{X}^{\intercal}\mathbf{X}$ обратима и уравнение имеет единственное решение относительно параметров

$$\mathbf{w} = (\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y}. \tag{38}$$

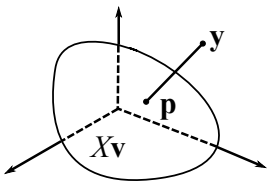



Рис. 6. Проекция вектора зависимой переменной на пространство столбцов матрицы плана.

Проекция вектора $\mathbf{y}$ на пространство столбцов матрицы $\mathbf{X}$ имеет вид

$$\mathbf{p} = \mathbf{Xw} = \mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}\mathbf{y} = \mathbf{Py}.$$

Матрица $\mathbf{P} = \mathbf{X}(\mathbf{X}^{\intercal}\mathbf{X})^{-1}\mathbf{X}^{\intercal}$ называется матрицей проектирования. Она она идемпотентна, $\mathbf{P}^2 = \mathbf{P}$, и симметрична, $\mathbf{P}^{\intercal} = \mathbf{P}$.

Используемые здесь методы нахождения оптимальных параметров моделей предполагают непрерывную дифференцируемость функции $S(\mathbf{w})$ в области $\mathbb{W} \ni \mathbf{w}$. Согласно (36),

$$S(\mathbf{w}) = (\mathbf{Xw} - \mathbf{y})^{\intercal}(\mathbf{Xw} - \mathbf{y}) = \mathbf{y}^{\intercal}\mathbf{y} - 2\mathbf{y}^{\intercal}\mathbf{Xw} + \mathbf{w}^{\intercal}\mathbf{X}^{\intercal}\mathbf{Xw}.$$

Для того, чтобы найти минимум этой функции, требуется приравнять её градиент к нулю:

$$\frac{\partial S}{\partial \mathbf{w}} = -2\mathbf{X}^{\intercal}\mathbf{y} + 2\mathbf{X}^{\intercal}\mathbf{Xw} = 0.$$

Решение этого уравнения совпадает с решением (38).

### 1.4.2. Существенно нелинейные модели

Для произвольной существенно нелинейной модели $f(\mathbf{w}, \mathbf{x})$ принята гипотеза (15) о нормальности распределения зависимой случайной величины $y$. Требуется найти такое значение вектора параметров $\mathbf{w}$, которое бы доставляло локальный минимум функции $S(\mathbf{w})$, заданной выражением (36). Принятие этой функции ошибки обусловлено предположением о возможности такой линеаризации

$$\mathbf{y} - \mathbf{f}(\mathbf{w} - \mathbf{w}_0, \mathbf{X}) = \mathbf{J}(\mathbf{w} - \mathbf{w}_0)$$

существенно нелинейной модели $f$,

$$\mathbf{J} = \left[\frac{\partial f(\mathbf{w}, \mathbf{x}_i)}{\partial w_j}\right], i \in \mathcal{I}, j \in \mathcal{J}, \tag{39}$$

в окрестности вектора параметров $\mathbf{w}_0$, котоая удовлетворяла бы условиям (37).

Так как функция $S(\mathbf{w})$ имеет локальный минимум, но в общем случае этот минимум не является единственным [301, 302], то предлагается назначить начальное значение вектора параметров $\mathbf{w}_0$, а затем найти последовательность приближений $\mathbf{w}_k$ вектора параметров к оптимальному вектору $\hat{\mathbf{w}}$ по шагам:

$$\hat{\mathbf{w}} \approx \mathbf{w}_{k+1} = \mathbf{w}_k + \Delta\mathbf{w}_k.$$

Здесь индекс $k$ вектора параметров обозначает номер итерации, $\Delta\mathbf{w}_k$ — вектор приращения, разность векторов параметров на двух последовательных шагах.

Для оценки приращения $\Delta\mathbf{w}_k$ используется линейное приближение функции

$$\mathbf{f}(\mathbf{w}_{k+1}, \mathbf{X}) = \mathbf{f}(\mathbf{w}_k, \mathbf{X}) + \mathbf{J}\Delta\mathbf{w}_k, \tag{40}$$

где $\mathbf{J}$ — матрица Якоби (39) вектор-функции $\mathbf{f}(\mathbf{w}, \mathbf{X})$ в точке $\mathbf{w}_k$.

Приращение $\Delta\mathbf{w}_k$ в точке $\mathbf{w}$, доставляющее минимум $S(\mathbf{w})$, равно нулю. Поэтому для нахождения следующего значения приращения $\Delta\mathbf{w}$ приравняем к нулю вектор частных производных $S(\mathbf{w})$ по $\mathbf{w}$. Для этого представим выражение (36) в виде

$$S(\mathbf{w}) = \|\mathbf{y} - \mathbf{f}(\mathbf{w} + \Delta\mathbf{w}_k)\|^2 = \mathbf{f}^\intercal(\mathbf{w} + \Delta\mathbf{w}_k, \mathbf{X})\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}_k, \mathbf{X}) - 2\mathbf{y}^\intercal\mathbf{f}(\mathbf{w} + \Delta\mathbf{w}_k, \mathbf{X}) + \mathbf{y}^\intercal\mathbf{y}$$

продифференцируем и приравняем к нулю:

$$\frac{\partial S}{\partial \mathbf{w}_k} = (\mathbf{J}^\intercal\mathbf{J})\Delta\mathbf{w} - \mathbf{J}^\intercal\big(\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\big) = 0.$$

Таким образом, чтобы найти значение $\Delta\mathbf{w}_k$, нужно решить систему линейных уравнений

$$\Delta\mathbf{w}_k = (\mathbf{J}^\intercal\mathbf{J})^{-1}\mathbf{J}^\intercal\big(\mathbf{y} - \mathbf{f}(\mathbf{w}_k, \mathbf{X})\big). \tag{41}$$

В том случае, когда функция ошибки $S(\mathbf{w})$ задана как взвешенная сумма квадратов остатков,

$$S(\mathbf{w}) = \big(\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\big)^\intercal\mathbf{W}\big(\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\big),$$

где $\mathbf{W}$ — диагональная матрица с неотрицательными элементами на диагонали, уравнение (41) будет иметь вид

$$\Delta\mathbf{w}_k = (\mathbf{J}^\intercal\mathbf{W}\mathbf{J})^\intercal\mathbf{J}^\intercal\mathbf{W}\big(\mathbf{y} - \mathbf{f}(\mathbf{w}_k, \mathbf{X})\big).$$

Так как число обусловленности матрицы $\mathbf{J}^\intercal\mathbf{J}$ есть квадрат числа обусловленности матрицы $\mathbf{J}$, то матрица $\mathbf{J}^\intercal\mathbf{J}$ может оказаться существенно вырожденной. Проблема большого числа обусловленности матрицы $\mathbf{J}^\intercal\mathbf{J}$, возникающая при итеративном нахождении параметров существенно-нелинейных моделей, решается методами регуляризации [265, 349, 310, 348]. При этом вводится параметр регуляризации $\lambda \geqslant 0$:

$$\Delta\mathbf{w} = (\mathbf{J}^\intercal\mathbf{J} + \lambda\mathbf{I})^{-1}\mathbf{J}^\intercal\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big).$$

Параметр $\lambda$ назначается на каждой итерации алгоритма. Если значение ошибки $S$ убывает быстро, то можно выбрать близкое к нулю значение $\lambda$ и свести этот алгоритм к алгоритму Гаусса-Ньютона [179].

Итерации прекращаются в том случае, если приращение $\Delta\mathbf{w}$ в последующей итерации меньше заданного значения, либо если параметры $\mathbf{w}$ доставляют ошибку $S(\mathbf{w})$, меньшую заданной величины. Значение вектора $\mathbf{w}$ на последней итерации считается искомым.

Недостаток алгоритма — значительное увеличение параметра $\lambda$ при плохой скорости сходимости. При этом обращение матрицы $(\mathbf{J}^\intercal\mathbf{J} + \lambda\mathbf{I})$ сводится к обращению её второго слагаемого. Этот недостаток можно устранить, используя диагональ матрицы $\mathbf{J}^\intercal\mathbf{J}$ в качестве регуляризующего слагаемого:

$$\Delta\mathbf{w} = \big(\mathbf{J}^\intercal\mathbf{J} + \lambda\mathrm{diag}(\mathbf{J}^\intercal\mathbf{J})\big)^{-1}\mathbf{J}^\intercal\big(\mathbf{y} - \mathbf{f}(\mathbf{w})\big). \tag{42}$$

### 1.4.3. Оптимизация целевой функции общего вида

Приведем вариант алгоритма, описанного в предыдущем разделе, который минимизирует не функцию ошибки (36), функцию ошибки (23) общего вида, которая включает матрицы гиперпараметров $\mathbf{A}, \mathbf{B}$. Как и ранее, используем пошаговое линейное приближение (40), в котором строка матрицы Якоби $\mathbf{J}_i = \frac{\partial f(\mathbf{w},\mathbf{x}_i)}{\partial \mathbf{w}}$. Разложение целевой функции в ряд Тейлора имеет вид

$$S(\mathbf{w}+\Delta\mathbf{w}) \approx \frac{1}{2}(\mathbf{w}+\Delta\mathbf{w})^\intercal\mathbf{A}(\mathbf{w}+\Delta\mathbf{w}) + \frac{1}{2}\left(\mathbf{y} - \mathbf{f}(\mathbf{w},\mathbf{X}) - \mathbf{J}\Delta\mathbf{w}\right)^\intercal\mathbf{B}\left(\mathbf{y} - \mathbf{f}(\mathbf{w},\mathbf{X}) - \mathbf{J}\Delta\mathbf{w}\right).$$

Отбросив члены второго порядка, получим:

$$S(\mathbf{w}+\Delta\mathbf{w}) \approx \frac{1}{2}\mathbf{w}^\intercal\mathbf{A}\mathbf{w} + \mathbf{w}^\intercal\mathbf{A}\Delta\mathbf{w} + \left(\mathbf{y} - \mathbf{f}(\mathbf{w},\mathbf{X}) - \mathbf{J}\Delta\mathbf{w}\right)^\intercal\mathbf{B}\left(\mathbf{y} - \mathbf{f}(\mathbf{w},\mathbf{X}) - \mathbf{J}\Delta\mathbf{w}\right).$$

Для нахождения минимума функции $S(\mathbf{w}+\Delta\mathbf{w})$ приравняем к нулю её градиент по $\Delta\mathbf{w}$:

$$\frac{\partial S(\mathbf{w}+\Delta\mathbf{w})}{\partial\Delta\mathbf{w}} = \mathbf{w}^\intercal\mathbf{A} - \beta\mathbf{J}^\intercal(\mathbf{y} - f(\mathbf{w},\mathbf{x}) - \mathbf{J}\Delta\mathbf{w}) = 0,$$

откуда получим решение нормального уравнения для функции ошибки общего вида

$$\Delta\mathbf{w} = (\mathbf{J}^\intercal\mathbf{J})^{-1}\left(\mathbf{J}^\intercal\big(\mathbf{y} - f(\mathbf{w},\mathbf{x})\big) - \frac{1}{\beta}\mathbf{A}^{-1}\mathbf{w}\right).$$

Так как число обусловленности матрицы $(\mathbf{J}^\intercal\mathbf{J})^{-1}$ растет при приближении функции $S(\mathbf{w})$ к минимальному значению, как в случае (42) используем регуляризирующий параметр $\lambda$.

Сформулируем вышеприведенные результаты в виде теоремы.

**Теорема 1.** *Для гипотезы нормального распределения зависимой переменной при фиксированных ковариационных матрицах* $\mathbf{A}^{-1}, \mathbf{B}^{-1}$ *итерационный алгоритм оценки параметров*

$$\Delta\mathbf{w}_{k+1} = (\mathbf{J}^\intercal\mathbf{J})^{-1}\left(\mathbf{J}^\intercal\big(\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\big) - \frac{1}{\beta}\mathbf{A}^{-1}\mathbf{w}_k\right)$$

*доставляет локальный минимум функции ошибки общего вида* $S(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, \mathbf{f})$ *при сходимости последовательности векторов* $\mathbf{w}_k$.

**Замечание 1.** *Итерационный алгоритм* $\mathbf{w}_{k+1} = \Delta\mathbf{w}_{k+1} + \mathbf{w}_k$ *предполагает известное начальное приближение* $\mathbf{w}_0$. *Последовательность* $\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2$ *монотонно убывает с увеличением номера шага* $k$.

### 1.4.4. Оценка параметров функции ошибки общего вида методом сопряженных градиентов

Если известны значения гиперпараметров $\mathbf{A}, \mathbf{B}$ для нелинейной регрессионной модели, то можно использовать алгоритм Левенберга-Марквардта для оценки вектора параметров модели $\mathbf{w}$. Пусть задано некоторое приближение для значений параметров модели $\mathbf{w}$. Функция ошибки имеет вид:

$$S = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^\intercal\mathbf{A}(\mathbf{w} + \Delta\mathbf{w}) + \frac{1}{2}\big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^\intercal\mathbf{B}\big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big). \tag{43}$$

Для минимизации функции ошибки воспользуемся алгоритмом Левенберга-Марквардта, который предназначен для оптимизации параметров нелинейных регрессионных моделей. Алгоритм заключается в последовательном приближении заданных начальных значений параметров к искомому локальному оптимуму и является обобщением метода сопряжённых градиентов и алгоритма Ньютона-Гаусса.

На первой итерации алгоритма задаётся начальное приближение для $\mathbf{w}$. Приращение $\Delta\mathbf{w}$ в точке оптимума для функции ошибки (43) равно нулю. Поэтому для нахождения экстремума приравняем вектор частных производных $S$ по $\mathbf{w}$ к нулю. Для этого представим $S$ в виде двух слагаемых:

$$S_1 = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^\intercal\mathbf{A}(\mathbf{w} + \Delta\mathbf{w}), \tag{44}$$

$$S_2 = \frac{1}{2}\big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^T\mathbf{B}\big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big). \tag{45}$$

После дифференцирования получим следующие выражения:

$$\frac{\partial S_1}{\partial\mathbf{w}} = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^\intercal(\mathbf{A} + \mathbf{A}^\intercal),$$

$$\frac{\partial S_2}{\partial\mathbf{w}} = \frac{1}{2}\big[\big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^\intercal\mathbf{B}^\intercal\mathbf{X} + \big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^\intercal\mathbf{B}\mathbf{X}\big].$$

Таким образом, чтобы найти приращение $\Delta\mathbf{w}$ необходимо решить систему линейных уравнений:

$$\nabla S = \frac{1}{2}(\mathbf{w} + \Delta\mathbf{w})^\intercal(\mathbf{A} + \mathbf{A}^\intercal) + \frac{1}{2}\big[\big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^\intercal\mathbf{B}^\intercal\mathbf{X} + \big(\mathbf{X}(\mathbf{w} + \Delta\mathbf{w}) - \mathbf{y}\big)^\intercal\mathbf{B}\mathbf{X}\big] = 0,$$

$$\Delta\mathbf{w} = \big[\big(\mathbf{A} + \mathbf{A}^\intercal + \mathbf{X}^\intercal(\mathbf{B}^\intercal + \mathbf{B})\mathbf{X}\big)^{-1}\big]^\intercal\big(-\mathbf{w}^\intercal(\mathbf{A} + \mathbf{A}^\intercal) + (\mathbf{y} - \mathbf{X}\mathbf{w})^\intercal(\mathbf{B}^\intercal + \mathbf{B})\mathbf{X}\big)^\intercal.$$

Так как матрицы $\mathbf{A}$, $\mathbf{B}$ — симметричные, положительно определенные матрицы ковариации, то эта система эквивалентна:

$$\Delta\mathbf{w} = \left[\left(\mathbf{A} + \mathbf{X}^\intercal\mathbf{B}\mathbf{X}\right)^{-1}\right]^\intercal\left(-\mathbf{w}^\intercal\mathbf{A} + (\mathbf{y} - \mathbf{X}\mathbf{w})^\intercal\mathbf{B}\mathbf{X}\right)^\intercal.$$

То есть,

$$\Delta\mathbf{w} = \left(\mathbf{X}^\intercal\mathbf{B}\mathbf{X} + \mathbf{A}\right)^{-1}\mathbf{X}^\intercal\mathbf{B}^\intercal\mathbf{y} - \mathbf{w}.$$

Вышеописанная процедура останавливается, в том случае, если приращение $\Delta\mathbf{w}$ в последующей итерации меньше заданного значения, либо если параметры $\mathbf{w}$ доставляют ошибку $S$ меньшую заданной величины. Значение вектора $\mathbf{w}$ на последней итерации считается искомым.

### 1.4.5. Обобщенно-линейные модели

Пусть целевая функция $S(\mathbf{w})$ задана в виде (27), согласно гипотезе (24). Используя метод Ньютона-Рафсона [94], рассмотрим $(k+1)$-й шаг минимизации целевой функции $S(\mathbf{w})$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mathbf{H}^{-1}\nabla S(\mathbf{w}), \tag{46}$$

где $\mathbf{H}$ — матрица Гессе, элементы которой являются вторыми производными целевой функции по параметрам модели

$$\mathbf{H} = \left[\frac{\partial^2 S(\mathbf{w})}{\partial w_i \partial w_j}\right], \quad i, j \in \mathcal{J}.$$

Применим этот метод к модели линейной регрессии (3) считая, зависимая переменная $y$ является нормально распределенной многомерной случайной величиной (15). Тогда первые и вторые производные функции ошибки $S(\mathbf{w})$ будут иметь вид

$$\nabla S(\mathbf{w}) = \frac{\partial S(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{X}^\intercal\mathbf{X}\mathbf{w} - \mathbf{X}^\intercal\mathbf{y} = \sum_{i=1}^{m}\left(\mathbf{w}^\intercal\mathbf{x}_i - y^i\right)$$

и

$$\mathbf{H} = \nabla\nabla S(\mathbf{w}) = \mathbf{X}^\intercal\mathbf{X} = \sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^\intercal.$$

Перепишем $(k+1)$-й шаг итерации (46) в виде

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\left(\mathbf{X}^\intercal\mathbf{X}\mathbf{w}_k - \mathbf{X}^\intercal\mathbf{y}\right) = \left(\mathbf{X}^\intercal\mathbf{X}\right)^{-1}\mathbf{X}^\intercal\mathbf{y}.$$

Полученное решение эквивалентно решению методом наименьших квадратов (38).

Применим этот метод, принимая гипотезу распределения данных (24) для модели логистической регрессии (26). Дифференцируя однократно и двукратно функцию ошибки $S(\mathbf{w})$ по элементам вектора параметров $\mathbf{w}$, получим её градиент

$$\nabla S(\mathbf{w}) = \mathbf{X}^\intercal(\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}) = \sum_{i=1}^{m}(f_i(\mathbf{w}, \mathbf{x}_i) - y_i)\mathbf{x}_i$$

и гессиан

$$\mathbf{H} = \nabla\nabla S(\mathbf{w}) = \mathbf{X}^\intercal\mathbf{B}\mathbf{X} = \sum_{i=1}^{m} f(\mathbf{w}, \mathbf{x}_i)\big(1 - f(\mathbf{w}, \mathbf{x}_i)\big)\mathbf{x}_i\mathbf{x}_i^\intercal.$$

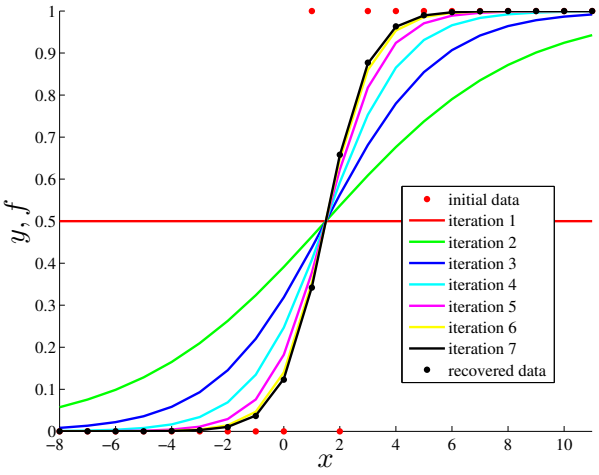



Рис. 7. Сходимость параметров логистической регрессионной модели.

В последнем выражении присутствует диагональная матрица весовых коэффициентов $\mathbf{B} = \operatorname{diag}[\beta_1, \ldots, \beta_m]$ размера $(m \times m)$ с элементами

$$\beta_i = f(\mathbf{w}, \mathbf{x}_i)\big(1 - f(\mathbf{w}, \mathbf{x}_i)\big).$$

Так как гессиан зависит в этом случае от матрицы $\mathbf{B}$, элементы которой, в свою очередь, содержат параметры модели, то целевая функция $S(\mathbf{w})$ не является теперь квадратичной. Так как модель $f(\mathbf{w}, \mathbf{x})$ вида (26) принимает значения на интервале $(0, 1)$, то для произвольного вектора $\mathbf{u}$ справедливо неравенство

$$\mathbf{u}^\intercal \mathbf{H} \mathbf{u} > 0.$$

Следовательно, целевая функция $S$ является выпуклой функцией аргумента $\mathbf{w}$ и имеет единственный минимум.

Процедура Ньютона-Рафсона имеет вид

$$\mathbf{w}_{k+1} = \mathbf{w}_k - (\mathbf{X}^\intercal \mathbf{B} \mathbf{X})^{-1} \mathbf{X}^\intercal (\mathbf{f} - \mathbf{y}) = (\mathbf{X}^\intercal \mathbf{B} \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{B} \big(\mathbf{X}\mathbf{w}_k - \mathbf{B}^{-1}(\mathbf{f} - \mathbf{y})\big).$$

Элементы диагональной матрицы $\mathbf{B}$ интерпретируются как дисперсии элементов вектора $\mathbf{y}$ — многомерной случайной величины. При этом условное математическое ожидание зависимой переменной

$$\mathsf{E}(y|\mathbf{x}) = \sigma(\mathbf{x}) = f(\mathbf{w}, \mathbf{x})$$

и дисперсия

$$\operatorname{var}(y|\mathbf{x}) = \mathsf{E}(y^2|\mathbf{x}) - E^2(y|\mathbf{x}) = \sigma(\mathbf{x}) - \sigma^2(\mathbf{x}) = f(\mathbf{w}, \mathbf{x})\big(1 - f(\mathbf{w}, \mathbf{x})\big).$$

Заметим, что для данной гипотезы порождения данных значение зависимой переменной $y = y^2$, так как $y \in \{0, 1\}$. Так как матрица $\mathbf{B}$ является диагональной, предполагается, что элементы многомерной случайной величины $\mathbf{y}$ некореллированы. При биномиальном распределении зависимой переменной и нормальном распределении параметров выражение (22), числитель которого соответствует функции ошибки $S(\mathbf{w})$, процедура оценки параметров моделей имеет

вид

$$\begin{aligned}\mathbf{w}_{k+1} =& \mathbf{w}_k - (\mathbf{X}^\intercal\mathbf{B}\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{f}-\mathbf{y}) + \frac{1}{2}\mathbf{w}_k^\intercal\mathbf{A}\mathbf{w}_k = \\ & (\mathbf{X}^\intercal\mathbf{B}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{B}\big(\mathbf{X}\mathbf{w}_k - \mathbf{B}^{-1}(\mathbf{f}-\mathbf{y})\big) + \frac{1}{2}\mathbf{w}_k^\intercal\mathbf{A}\mathbf{w}_k.\end{aligned} \tag{47}$$

Рис. 7 иллюстрирует сходимость вышеописанного алгоритма. На оси абсцисс показана единственная свободная переменная, на оси ординат — значение функции регрессии $f$ и зависимой переменной $y$.

Перепишем полученные результаты в виде теоремы.

**Теорема 2.** *Для гипотезы нормального распределения зависимой переменной вариант: биномиального при фиксированных ковариационных матриц* $\mathbf{A}^{-1}, \mathbf{B}^{-1}$ *итерационный алгоритм оценки параметров обобщенно-линейной модели*

$$\Delta\mathbf{w}_{k+1} = \left(\mathbf{X}^\intercal\mathbf{B}\mathbf{X} + \mathbf{A}\right)^{-1}\mathbf{X}^\intercal\mathbf{B}^\intercal\mathbf{y} - \mathbf{w}_k, \quad \textit{вариант:}$$

$$\Delta\mathbf{w}_{k+1} = (\mathbf{X}^\intercal\mathbf{B}\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{B}\big(\mathbf{X}\mathbf{w}_k - \mathbf{B}^{-1}(\mathbf{f}-\mathbf{y})\big) + \frac{1}{2}\mathbf{w}_k^\intercal\mathbf{A}\mathbf{w}_k$$

*доставляет локальный минимум функции ошибки общего вида* $S(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, \mathbf{f})$ *при сходимости последовательности векторов* $\mathbf{w}_k$.

### 1.4.6. Оптимизация многокритериальной функции ошибок

При выборе регрессионных моделей в ряде задач требуется использовать несколько функций ошибок или критериев качества моделей. В качестве примера приведем задачу кредитного скоринга, которая ставится как задача логистической регрессии и предполагает, что оценка параметров получена в предположении о биномиальном распределении зависимой переменной. Одновременно, стандарт «Basel-II» [3] выдвигает ряд дополнительных критериев и требований к моделям-претендентам. В таких случаях для выбора строится Парето-оптимальный фронт, называемый также оболочкой Эджворда-Парето [202, 99], включающий в качестве элементов недоминируемое геометрическое место точек.

Для решения задачи восстановления регрессии задается множество критериев, условиям оптимальности которых должна удовлетворять модель. Отыскиваются векторы, принадлежащие Парето-оптимальному фронту множества всех векторов, соответствующих порожденным моделям [320, 319, 91]. Поставим оптимизационную задачу оценки Парето-оптимального фронта в многокритериальной оптимизации. Рассмотрим невыпуклую немонотонную вектор-функцию $\boldsymbol{\zeta} : \mathcal{W} \to \mathcal{S}$, переводящую односвязную область $\mathcal{W} \subseteq \mathbb{R}^n$ в область $\mathcal{S} \subseteq \mathbb{R}^p$. Множество $\mathcal{W}$ будем называть множеством возможных решений, а множество $\mathcal{S}$ — множеством достижимых значений критериальных векторов

$$\mathcal{S} = \{\mathbf{s} : \mathbf{s} = \boldsymbol{\zeta}(\mathbf{w}), \mathbf{w} \in \mathcal{W}\}.$$

Направления желательного изменения критериев заданы следующим образом. Задано отношение доминирования на множестве $\mathcal{S}$ такое, что вектор $\mathbf{s}_1 \in \mathcal{S}$ доминирует вектор $\mathbf{s}_2 \in \mathcal{S}$,

$$\mathbf{s}_1 \succ \mathbf{s}_2, \text{если } s_{11} \geqslant s_{21}, ..., s_{1p} \geqslant s_{2p},$$

при условии, что векторы $\mathbf{s}_1 = [s_{11}, ..., s_{1p}]^\intercal$ и $\mathbf{s}_2 = [s_{21}, ..., s_{2p}]^\intercal$ не совпадают.

Множество $\mathcal{S}^*(\boldsymbol{\zeta})$ недоминированных значений целевой функции $\boldsymbol{\zeta}$ называется *Парето-оптимальным фронтом*

$$\mathcal{S}^* = \{\mathbf{s}^* : \nexists \boldsymbol{\zeta}(\mathbf{w}) \succ \mathbf{s}^*, \mathbf{w} \in \mathcal{W}\}.$$

Множество векторов $\mathcal{W}^* = \{\mathbf{w}^*\}$ из прообраза отображения $\boldsymbol{\zeta} : \mathcal{W} \to \mathcal{S}$ будем называть *Парето-оптимальным множеством*, если образ каждого из этих векторов принадлежит Парето-оптимальному фронту:

$$\mathcal{W}^* = \{\mathbf{w}^* \in \mathcal{W} : \boldsymbol{\zeta}(\mathbf{w}^*) \in \mathcal{S}^*\}.$$

Одним из способов сведения задачи многокритериальной оптимизации к однокритериальной является свертка критериев. Сверткой $\varphi$ вектор-функции $\boldsymbol{\zeta}$ будем называть взвешенную сумму

$$\varphi\big(\mathbf{v}, \boldsymbol{\zeta}(\mathbf{w})\big) = \sum_{i=1}^{p} v_i \zeta_i(\mathbf{w}), \quad \text{где} \quad \sum_{i=1}^{p} v_i = 1.$$

Оптимизационная задача имеет вид

$$\hat{\mathbf{w}} = \underset{\mathbf{w}\in\mathcal{W}, \mathbf{v}\in\{\|\mathbf{v}\|_1=1\}}{\arg\max} \varphi\big(\mathbf{v}, \boldsymbol{\zeta}(\mathbf{w})\big).$$

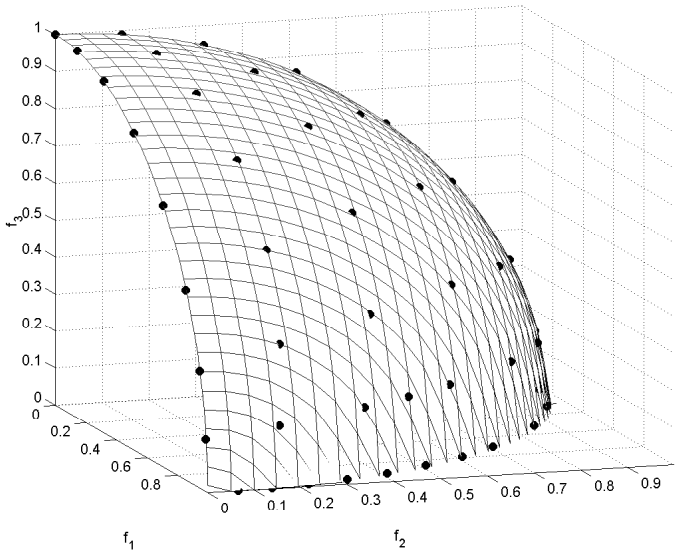



Рис. 8. Пример полученного Парето-оптимального фронта.

Парето-оптимальный фронт $\mathcal{S}^*$ также можно получить, максимизируя один из интегральных критериев его качества[271, 77, 161] , описанных ниже.

**Критерии качества найденного Парето-оптимального фронта** включают критерии двух типов: близость к истинному фронту и разнородность или разброс решений в пространстве. Критерии близости:

1) число найденных решений $K_{\text{conv}}$ (в частных случаях предполагается, что множество найденных решений счетно), расстояние от которых до предполагаемого или известного фронта не превышает заданное;

2) отношение числа недоминируемых точек фронта $K_{\text{nondom}}$ к количеству точек;

3) средняя сходимость к фронту $\mu_{\text{AC}}$.

Используется следующая модификация последнего критерия:

$$\mu_{\text{AC}} = \frac{1}{K}\sum_{i=1}^{K} d_i,$$

где $K$ — число точек, принадлежащих фронту, а $d_i$ — евклидово расстояние от $i$-й точки до известной границы фронта.

Критерии разнородности:

1) критерий разнообразия Шотта $\mu_{\text{SS}}$,

2) критерий отношений объемов $\mu_{\text{VR}}$,

3) критерий разности объемов $\mu_{\text{VD}}$,

4) критерий разнородности $\mu_{\text{DD}}$.

Критерий разнообразия Шотта равен

$$\mu_{\text{SS}} = \sqrt{\frac{1}{K-1}\sum_{i=1}^{K}(d_i - \bar{d})^2},$$

где

$$d_i = \min_k \sum_{j=1}^{p} |s_{ij} - s_{kj}|, \quad i,k \in \{1,...,K\}$$

является манхеттенским расстоянием между точками $s_i, s_k$ найденного фронта, а $\bar{d}$ — среднее расстояние $d_1, ..., d_K$.

Критерий отношений объемов равен

$$\mu_{\text{VR}} = \frac{h}{\mathbf{H}},$$

где

$$h = \prod_{j=1}^{p}\left(\max_{i\in\{1,...,K\}} s_{ij} - \min_{k\in\{1,...,K\}} s_{kj}\right),$$

и $\mathbf{H}$ — объемы минимальных кубов размерности $p$, соответственно включающих полученный и истинный фронты.

Критерий разности объемов — относительный объем области, доминируемой точками полученного фронта. Обозначим буквами $C(\mathbf{s}_1), \ldots, C(\mathbf{s}_K)$ положительные конусы с вершинами в точках $\mathbf{s}_1, \ldots, \mathbf{s}_K$ фронта, включающие геометрическое место точек, доминирующее данные точки. Пусть $\mathbf{r}$ — заданная опорная точка, $R$ — отрицательный конус, геометрическое место точек, доминируемое опорной точкой.

Критерий разности объемов задан как

$$\mu_{\text{VD}} = \frac{\text{vol}\left(\bigcup_{i=1}^{K}(C(\mathbf{s}_i)\cap R)\right)}{\text{vol}\left(\bigcup_{\mathbf{s}\in Z^*}(C(s)\cap R)\right)}.$$

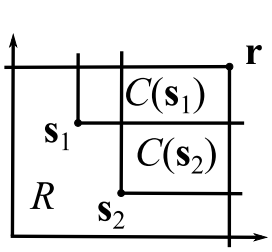



Рис. 9. Максимизация объема, заданного опорной точкой при поиске Парето-оптимального фронта.

Здесь vol(·) — объем объединения конусов, а $Z^*$ — множество всех точек в истинном фронте.

Рисунок 9 иллюстрирует процедуру вычисления этого критерия. Точки $\mathbf{s}_1, \mathbf{s}_2$ на рисунке задают конусы $C(\mathbf{s}_1), C(\mathbf{s}_2)$. Опорная точка $\mathbf{r}$ задает отрицательный конус $R$. Заштрихованная область соответствует объединению конусов:

$$\big(C(\mathbf{s}_1) \cap R\big) \cup \big(C(\mathbf{s}_2) \cap R\big).$$

Искомое значение критерия определяется объемом этой области в $p$-мерном пространстве.

## 1.5. Ограничения, накладываемые на множество моделей

После оценки параметров выбранной регрессионной модели встает вопрос о её статистических свойствах. При этом кроме требований и ограничений заданных прикладной задачей исследуется, во-первых, качество полученной функции регрессии и, во-вторых, её устойчивость относительно возмущения параметров. В первом случае выполняется анализ регрессионных остатков, во втором случае исследуется мультиколлинеарность признаков и вырожденность пространства параметров.

### 1.5.1. Анализ регрессионных остатков

Требование соответствия вектора регрессионных остатков $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{f}$ принятой гипотезе порождения данных не только задает функцию ошибки, но и влечет ряд дополнительных условий, проверка которых называется *анализом регрессионных остатков*.

Анализ регрессионных остатков заключается в проверке следующих гипотез:

1) что матожидание регрессионных остатков равно нулю,

$$\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \tag{48}$$

2) дисперсия регрессионных остатков постоянна и не зависит от переменной $\mathbf{x}$,

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}, \tag{49}$$

3) что регрессионные остатки распределены нормально,

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{50}$$

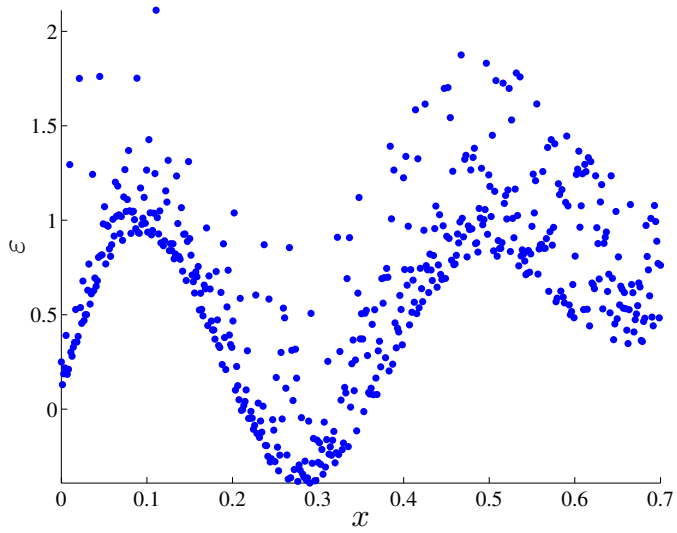



Рис. 10. Гетероскедактичные регрессионные остатки с ненулевым средним.

где $\boldsymbol{\varepsilon}$ — вектор регрессионных остатков некоторой модели. На рисунке 10 представлен пример регрессионных остатков, не удовлетворяющих ни одному из условий (48), (49) и (50).

Гипотеза (48) проверяется критерием знаков. Гипотеза гомоскедастичности [18, 248, 90] или постоянства дисперсий (49) проверяется тестом Ансари-Брэдли и критерием Голдфелда-Кванта. Так как тест Ансари-Брэдли проверяет равенство дисперсий двух выборок, предлагается разбить выборку регрессионных остатков на две подвыборки несколько раз. Независимость (49) проверяется статистикой Дарбина-Ватсона. Нормальность распределения (50) проверяется критерием согласия $\chi^2$, который сравнивает распределение остатков с эталонным нормальным распределением, параметры которого вычислены по регрессионным остаткам. Вышеперечисленные тесты и статистики подробно рассмотрены в [303, 321, 141, 312, 81].

При отвержении теста гомоскедастичности рекомендуется использовать один из нижеперечисленных тестов гетероскедаксичности.

**Тест Уайта.** Предположим, что гетероскедастичность модели вызвана неявной зависимостью дисперсий ошибок от признаков. Примем гипотезу $H_0$ без каких-либо предположений о структуре гетероскедастичности. Сначала применим к исходной модели метод наименьших квадратов и найдем вектор регрессионных остатков $\boldsymbol{\varepsilon}$. Затем восстановим регрессию квадратов этих остатков $\boldsymbol{\varepsilon}^2$ на все признаки, их квадраты, попарные произведения и константу. Тогда при неотвержении гипотезы $H_0$ величина $mR^2$ асимптотически имеет распределение $\chi^2(N-1)$, где $m$ — число элементов выборки, $N$ — число признаков второй регрессии а $R^2$ — коэффициент детерминации

$$R^2 \stackrel{\text{def}}{=} 1 - \frac{\|\boldsymbol{\varepsilon}\|^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}, \quad \text{здесь} \quad \bar{y} = \frac{1}{m}\sum_{i\in\mathcal{I}} y_i.$$

Недостаток этого теста в том, что если гипотеза $H_0$ отвергается, получить зависимость дисперсии регрессионных остатков от независимых переменных невозможно.

**Тест Голдфелда-Кванта.** Применяется, когда есть предположение о прямой зависимости дисперсии ошибок от одного признака $\boldsymbol{\chi}_j$. Упорядочим множество индексов $\mathcal{I} \ni i$ по убыванию признака $\boldsymbol{\chi}_j$, исключим $d$ средних наблюдений (пусть $d \approx 3^{-1}m$), восстановим две независимые регрессии для первых $d$ наблюдений и последних $d$ наблюдений получим регрессионные остатки $\boldsymbol{\varepsilon}_1$ и $\boldsymbol{\varepsilon}_2$. Далее вычислим статистику Фишера

$$F = \frac{\boldsymbol{\varepsilon}_1^\intercal \boldsymbol{\varepsilon}_1}{\boldsymbol{\varepsilon}_2^\intercal \boldsymbol{\varepsilon}_2}.$$

Если верна гипотеза $H_0$, то $F$ имеет распределение Фишера. Большая величина этой статистики означает, что гипотеза $H_0$ отвергается.

**Критерий Дарбина-Ватсона.** Если выборочная регрессия $\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{w}}, \mathbf{X})$ описывает истинную зависимость между $\mathbf{y}$ и $\mathbf{X}$, то регрессионные остатки $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_m]^\intercal$ должны быть независимыми, что проверяется при помощи коэффициента корреляции Дарбина-Ватсона

$$D = \|\boldsymbol{\varepsilon}\|^{-2} \sum_{i=2}^{m} (\varepsilon_i - \varepsilon_{i-1})^2.$$

При $D > D_1(\tau)$ или $D > 4 - D_1(\tau)$ с достоверностью $\tau$ принимается гипотеза о наличии соответственно отрицательной или положительной корреляции регрессионных остатков. При $D_2(\tau) > D > D_1(\tau)$ или $4 - D_1(\tau) > D > 4 - D_2(\tau)$ критерий не позволяет выявить наличие или отсутствии корреляции регрессионных остатков. При $D_2(\tau) < D < 4 - D_2(\tau)$ гипотеза корреляции регрессионных остатков отклоняется. Критические значения $D_1(\tau), D_2(\tau)$ для различных $\tau$ заданы.

В статистический отчет об анализе регрессионных остатков вместе со значением функции ошибки и результатами, полученными при тестировании вышеперечисленных гипотез входят также значения следующих критериев:

1) квадрат остаточной дисперсии:

$$\sigma^2_{\text{res}} = \frac{1}{m} \sum_{i \in \mathcal{I}} \big(y_i - f(\hat{\mathbf{w}}, \mathbf{x}_i)\big)^2.$$

2) квадрат дисперсии зависимой переменной:

$$\sigma^2_y = \frac{1}{m} \sum_{i \in \mathcal{I}} (y_i - \bar{y})^2,$$

где $\bar{y} = \frac{1}{m} \sum_{i \in \mathcal{I}} y_i$ — среднее значение элементов вектора $\mathbf{y}$.

3) коэффициент детерминации:

$$R^2 = 1 - \frac{m\sigma^2_{\text{res}}}{\sigma^2_y}.$$

В частности, если этот коэффициент окажется больше 0.95, то линейная регрессионная модель считается адекватной экспериментальным данным, иначе — неадекватной.

### 1.5.2. Адекватность регрессионной модели

«Адекватность модели — соответствие модели моделируемому объекту или процессу. Адекватность — в какой-то мере условное понятие, так как полного соответствия модели реальному объекту быть не может, иначе это была бы не модель, а сам объект. При моделировании имеется в виду адекватность не вообще, а по тем свойствам модели, которые для исследования считаются существенными» [318].

*Адекватной регрессионной моделью* [53] называется модель, удовлетворяющая условию максимума критерия Мэллоуза $C_p$,

$$C_p = \frac{\|\mathbf{y} - \mathbf{f}(\hat{\mathbf{w}}_{\mathcal{A}}, \mathbf{X}_{\mathcal{A}})\|^2}{\|\mathbf{y} - \mathbf{f}(\hat{\mathbf{w}}_{\mathcal{J}}, \mathbf{X}_{\mathcal{J}})\|^2} - |\mathbf{I}| + 2|\mathcal{A}|.$$

Другими словами, адекватной регрессионной моделью, называется модель, имеющая оптимальную сложность, определенную с помощью данного критерия [197]. Нижние индексы $\mathcal{A}, \mathcal{J}$ задают наборы признаков, на которых получены оценки параметров $\hat{\mathbf{w}}_{\mathcal{A}}, \hat{\mathbf{w}}_{\mathcal{J}}$.

Предположим, что модель адекватна и рассмотрим основные статистические свойства переменных, которые она включает [292]. Оценки параметров $\hat{\mathbf{w}} = \mathbf{w}_{\text{ML}}$ модели являются несмещенными, найдем их матожидание следующим образом

$$\mathsf{E}(\hat{\mathbf{w}}) = \mathsf{E}\big((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}\big) = \mathsf{E}\left((\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon})\right) = \mathbf{w}, \tag{51}$$

так как $\mathsf{E}(\boldsymbol{\varepsilon}) = 0$ и $(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X} = \mathbf{I}$. Ковариационная матрица многомерной случайной величины $\mathbf{w}$ имеет вид

$$\text{Cov}(\mathbf{w}) = \mathsf{E}\left((\mathbf{w} - \mathsf{E}\mathbf{w})(\mathbf{w} - \mathsf{E}\mathbf{w})^\intercal\right) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1},$$

так как $\mathsf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\intercal) = \mathbf{I}\sigma^2$. То есть матрица $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ является матрицей оценок ковариаций элементов вектора параметров $\mathbf{w}$.

Оценка вектора зависимых переменных $\mathbf{y}$ находится с помощью оценки вектора параметров $\mathbf{w}$. Принимая линейную модель $\mathsf{E}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\mathbf{w}$, получаем оценку $\hat{\mathbf{y}} = \mathsf{E}(\mathbf{y}|\mathbf{X}) = \mathbf{X}\hat{\mathbf{w}}$. При этом дисперсия зависимой переменной $\mathsf{E}\big(\mathsf{E}(\mathbf{y}|\mathbf{X})\big)$ — ковариационная матрица многомерной случайной величины $\mathbf{y}$, которая определяется выражением

$$\text{Cov}(\mathbf{y}) = \text{Cov}(\mathbf{X}\mathbf{w}) = \mathbf{X}\text{Cov}(\mathbf{w})\mathbf{X}^\intercal = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\sigma^2.$$

Вектор регрессионных остатков $\boldsymbol{\varepsilon}$, получаемый в результате оценки параметров, определяется выражением

$$\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right)\mathbf{y}, \tag{52}$$

где матрица $(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal) = \mathbf{P}$ симметрична и идемпотентна: $\mathbf{P}^\intercal = \mathbf{P}$ и $\mathbf{P}^2 = \mathbf{P}$.

Рассмотрим свойства суммы квадратов регрессионных остатков

$$\text{SSE} = \boldsymbol{\varepsilon}^\intercal\boldsymbol{\varepsilon}. \tag{53}$$

Подставляя выражение (52) в (53) и учитывая, что

$$\hat{\mathbf{w}}^\intercal = \mathbf{y}^\intercal\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1},$$

получаем

$$\mathrm{SSE} = \mathbf{y}^\intercal \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right)\mathbf{y} = \mathbf{y}^\intercal\mathbf{y} - \hat{\mathbf{w}}^\intercal\mathbf{X}^\intercal\mathbf{y}. \tag{54}$$

Здесь SSE записана как квадратичная форма вектора $\mathbf{y}$. Так как мы предполагаем, что

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}), \tag{55}$$

то математическое ожидание SSE имеет вид [168, 340].

Так как наиболее правдоподобная оценка параметров $\mathbf{w}$ при предположении (55) имеет вид $\mathbf{w} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$, и функция ошибки в этом случае задана как $S(\mathbf{w}) = \mathrm{SSE}$, то математическое ожидание суммы квадратов регрессионных остатков имеет вид

$$\mathsf{E}(S(\mathbf{w})) = \mathrm{tr}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right)\mathbf{I}\sigma^2 + (\mathbf{X}\mathbf{w})^\intercal\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right)\mathbf{X}\mathbf{w}. \tag{56}$$

Так как след идемпотентной матрицы (в данном случае это матрица Мура-Пенроуза $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$), то

$$\begin{aligned}\mathsf{E}(S(\mathbf{w})) = \mathrm{rank}\left((\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right)\mathbf{I}\sigma^2 &= \left(m - \mathrm{rank}(\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\right)\sigma^2 \\ &= (m - \mathrm{rank}(\mathbf{X}))\,\sigma^2 = (m-n)\sigma^2,\end{aligned}$$

здесь $m$ — число элементов выборки и строк матрицы $\mathbf{X}$. Если матрица плана $\mathbf{X}$ не содержит коллинеарных столбцов и её ранг $\mathrm{rank}(\mathbf{X}) = n$, то несмещенной оценкой $\sigma^2$ является оценка

$$\hat{\sigma}^2 = \frac{S(\mathbf{w})}{m-n}.$$

Оценки $\hat{\mathbf{w}}$ и $\sigma^2$ являются независимыми [242]. Для того, чтобы это показать, рассмотрим выражения (51) и (56) с учетом предположения (55) и проверим равенство нулю следующего выражения:

$$(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal(\sigma^2\mathbf{I})\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right) = \mathbf{0}.$$

Статистика $\frac{\mathrm{SSE}}{\sigma^2}$ имеет $\chi^2$-распределение. Из выражения (54), SSE является квадратичной формой от вектора $\mathbf{y}$, поэтому $\mathrm{SSE} = \mathbf{y}^\intercal P\mathbf{y}$, где матрица $P = \mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. Так как $\mathrm{SSE}/\sigma^2 = \mathbf{y}^\intercal(\frac{1}{\sigma^2}P)\mathbf{y}$, где матрица $P$ идемпотентна и $\mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, то матрица $\frac{1}{\sigma^2}P\sigma^2\mathbf{I}$ тоже идемпотентна. Таким образом, при предположении (55) матрица $\mathbf{y}^\intercal P\mathbf{y}$ имеет нецентральное распределение

$$\mathbf{y}^\intercal P\mathbf{y} \sim \chi^2\left(\mathrm{rank}(P), \frac{1}{2}(\mathbf{X}\mathbf{w})^\intercal P\mathbf{X}\mathbf{w}\right)$$

и, следовательно,

$$\frac{1}{\sigma^2}\mathrm{SSE} \sim \chi^2\left(\mathrm{rank}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right), \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w})^\intercal\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\right)\mathbf{X}\mathbf{w}\right),$$

что приводимо к виду $\frac{1}{\sigma^2}\mathrm{SSE} \sim \chi^2_{m-n}$, где $\mathrm{rank}(\mathbf{X}) = n$. В результате имеем следующее распределение:

$$\frac{\hat{\sigma}^2}{\sigma^2}(n-p) \sim \chi^2_{m-n}.$$

Мы показали, что $\frac{1}{\sigma^2}\mathrm{SSE}$ имеет центральное $\chi^2$-распределение. Покажем, что SSR имеет нецентральное $\chi^2$-распределение, независимое от SSE. С помощью этих двух статистик получим $F$-статистику, имеющую нецентральное $F$-распределение.

Таблица 5. Анализ дисперсии регрессионных остатков.

| Источник дисперсии | Сумма квадратов регрессионных остатков | Число степеней свободы | Средние квадраты | $F$-статистика |
|---|---|---|---|---|
| Регрессия | $\mathrm{SSR} = \mathbf{w}^\intercal \mathbf{X}^\intercal \mathbf{y}$ | $n$ | $\mathrm{MSR} = \frac{1}{n}\mathrm{SSR}$ | $F_R = \frac{\mathrm{MSR}}{\mathrm{MSE}}$ |
| Остатки | $\mathrm{SSE} = \mathbf{y}^\intercal \mathbf{y} - \mathbf{w}^\intercal \mathbf{X}^\intercal \mathbf{y}$ | $m - n$ | $\mathrm{MSE} = \frac{1}{(m-n)}\mathrm{SSE}$ | |
| Общая дисперсия | $\mathrm{SST} = \mathbf{y}^\intercal \mathbf{y}$ | $n$ | | |

**Составные части суммы квадратов невязок.** Представим полную сумму квадратов невязок как

$$\mathrm{SSR} = \mathrm{SST} - \mathrm{SSE} = \mathbf{w}^\intercal \mathbf{X}^\intercal \mathbf{y},$$

где $\mathrm{SST} = \mathbf{y}^\intercal \mathbf{y}$. Из последнего выражения получаем $\mathrm{SSR} = \mathbf{y}^\intercal \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{y}$. Используемая здесь матрица $\mathbf{Q} = \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal$ идемпотентна; произведение $\mathbf{Q}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal\right)$ равно нулю. Следовательно, при предположении (55), квадратичные формы $\mathbf{y}^\intercal P\mathbf{y}$ и $\mathbf{y}^\intercal \mathbf{Q}\mathbf{y}$ распределены независимо только тогда [242], когда

$$P(\sigma^2\mathbf{I})\mathbf{Q} = \mathbf{0} = \mathbf{Q}(\sigma^2\mathbf{I})P.$$

Выражение (57) равно матрице с нулевыми элементами,

$$\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \left(\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal\right)\sigma^2 = \left(\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal\right)\sigma^2 = \mathbf{0}, \qquad (57)$$

и, следовательно, величины SSR и SSE распределены независимо.

Статистика $\frac{1}{\sigma^2}\mathrm{SSR}$ имеет $\chi^2$-распределение. Поэтому, если $\frac{1}{\sigma^2}\mathrm{SSR} = \mathbf{y}^\intercal(\frac{1}{\sigma^2})\mathbf{Q}\mathbf{y}$, так как матрица $\mathbf{Q}$ является идемпотентной и $\mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}$, то матрица $\frac{1}{\sigma^2}\mathbf{Q}(\sigma^2\mathbf{I})$ также является идемпотентной. Следовательно, при предположении (55), случайная величина $\mathbf{y}^\intercal \mathbf{Q}\mathbf{y}$ имеет нецентральное распределение

$$\mathbf{y}^\intercal \mathbf{Q}\mathbf{y} \sim \chi^2\left(\mathrm{rank}(\mathbf{Q}), \frac{1}{2}(\mathbf{X}\mathbf{w})^\intercal \mathbf{Q}\mathbf{X}\mathbf{w}\right),$$

и

$$\frac{1}{\sigma^2}\mathrm{SSR} \sim \chi^2\left(\mathrm{rank}\left(\mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal\right), \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w})^\intercal \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal \mathbf{X}\mathbf{w}\right),$$

приводимая к виду

$$\frac{1}{\sigma^2}\mathrm{SSR} \sim \chi^2(p, \frac{1}{2\sigma^2}\mathbf{w}^\intercal \mathbf{X}^\intercal \mathbf{X}\mathbf{w}).$$

Построим на полученных результатах нецентральное $F$-распределение, получим из таблицы 5

$$F_R = \frac{\frac{1}{n}\mathrm{SSR}}{\frac{1}{m-n}\mathrm{SSE}} \sim F(n, m-n, \frac{1}{2\sigma^2}\mathbf{w}^\intercal \mathbf{X}^\intercal \mathbf{X}\mathbf{w}).$$

**Анализ дисперсии регрессионных остатков.** Представим выведенные ранее $F$-статистики в виде таблицы анализа дисперсии, см. таблицу 5. В таблице представлены суммы квадратов регрессионных остатков, соответствующие им степени свободы соответствующих $\chi^2$-распределений, а также значения самих $F$-статистик.

Опишем некоторые свойства распределения регрессионных остатков. При симметричной и идемпотентной матрице $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ остатки представимы в виде $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\mathbf{w} = \mathbf{y} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\,\mathbf{y} = \mathbf{P}\mathbf{y}$. Отсюда, ожидаемые величины остатков $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathsf{E}(\mathbf{P}\mathbf{y}) = \mathbf{P}\mathbf{X}\mathbf{w} = (\mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal)\,\mathbf{X}\mathbf{w} = \mathbf{0}$, поскольку $\mathbf{P}\mathbf{X} = \mathbf{0}$, и ковариационная матрица $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{Cov}(\mathbf{P}\mathbf{y}) = \mathbf{P}^2\sigma^2 = \mathbf{P}\sigma^2$.

**Проверка гипотез о матожидании параметров.** Статистика $F_R$ в таблице 5 имеет нецентральное $F$-распределение с параметром нецентральности $\frac{1}{2\sigma^2}\mathbf{w}^\intercal\mathbf{X}^\intercal\mathbf{X}\mathbf{w}$. Этот параметр становится нулевым при гипотезе: $\mathsf{E}\mathbf{w} = \mathbf{0}$. Тогда статистика $F_R$ имеет центральное $F$-распределение $F_{n,m-n}$. Если при $p$-уровне значимости $F_R$ больше значений табличного распределения $F_{n,m-n}$, то нулевая гипотеза $\mathsf{E}(\mathbf{w}) = \mathbf{0}$ отклоняется. Если величина когда $F_R$ для некоторой модели $\mathsf{E}(\mathbf{y}) = \mathbf{X}\mathbf{w}$ является значимой, то модель объясняет значительную часть дисперсии переменной $\mathbf{y}$. Если при $p$-уровне значимости статистика $F_Z$ больше табличного значения $F_{1,m-1}$, то нулевая гипотеза $\mathsf{E}(\mathbf{w}) = \mathbf{0}$ отклоняется.

### 1.5.3. Устойчивость моделей и мультиколлинеарность

Основная проблема, возникающая при порождении признаков, — проблема мультиколлинеарности. Термин *мультиколлинеарность* введен Р. Фришем [98] при рассмотрении линейных зависимостей между признаками. Мультиколлинеарность проявляется в сильной корреляции между двумя или более признаками $[\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_n]$ — столбцами матрицы $\mathbf{X}$, что затрудняет оценивание параметров модели. Мультиколлинеарность называют полной, если существует функциональная зависимость между всеми признаками. При этом становится невозможно однозначно оценить параметры модели. На практике встречаются случаи частичной мультиколлинеарности, когда имеется высокая степень корреляции между некоторыми признаками. Тогда решение получить можно, однако оценки параметров модели и их дисперсий могут быть неустойчивы. Увеличиваются дисперсии оценок и абсолютные значения регрессионных параметров, что усложняет их интерпретацию.

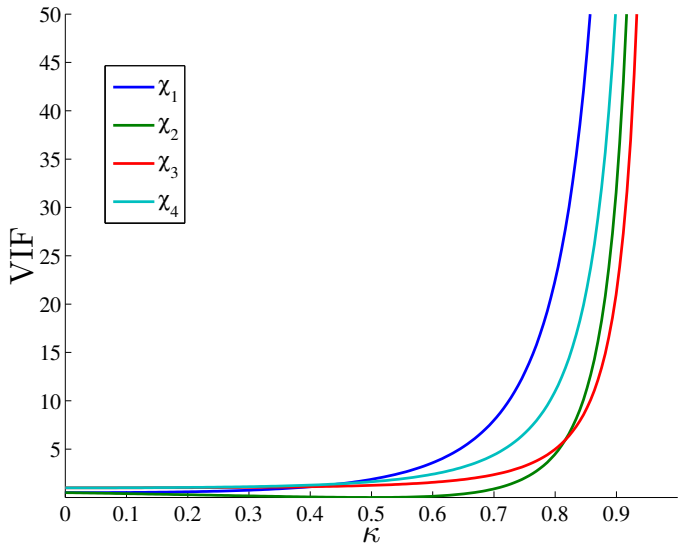



Рис. 11. Зависимость коэффициентов инфляции дисперсии от параметра $\kappa$.

Перечислим некоторые признаки мультиколлинеарности: значительные изменения в оценках параметров при добавлении или удалении признака, превышение некоторого порога абсолютным значением корреляции между признаками, близость к нулю определителя матрицы попарных корреляций $\mathrm{Cor}(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k)$ признаков. Основные способы обнаружения мультиколлинеарности — проверка корреляции между признаками [80], исследование факторов инфляции дисперсии (VIF — variance inflation factor) [200], метод Белсли [32, 101].

**Корреляция между признаками.** Пусть $\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_n]$ — матрица признаков, столбец $\boldsymbol{\chi}_j = [x_{1j}, \ldots, x_{mj}]^\intercal$ которой соответствует значениям $j$-го признака при различных измерениях. *Корреляционной матрицей* называется матрица, элементами которой являются выборочные корреляции между столбцами:

$$\mathrm{Cor}(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k) = \frac{(\boldsymbol{\chi}_j - \bar{x}_j\mathbf{1})^\intercal(\boldsymbol{\chi}_k - \bar{x}_k\mathbf{1})}{\|\boldsymbol{\chi}_j - \bar{x}_j\mathbf{1}\|\|\boldsymbol{\chi}_k - \bar{x}_k\mathbf{1}\|}, \quad j,k \in \mathbf{J}, \quad \bar{x}_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij},$$

где $\boldsymbol{\chi}_j, \boldsymbol{\chi}_k$ — столбцы $\mathbf{X}$, $\bar{x}_j, \bar{x}_k$ — средние значения соответствующих признаков, $\mathbf{1}$ — столбец из единиц, число которых равно числу признаков, см. [329]. В случае центрированных и нормированных (12) признаков $\boldsymbol{\chi}_j, \boldsymbol{\chi}_k$

$$\mathrm{Cor}(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k) = \boldsymbol{\chi}_i^\intercal \boldsymbol{\chi}_j.$$

Допустимым уровнем значимости называется минимальный уровень, вычисленный для данного значения статистики: значения коллинеарности преобразуются к $t$-статистике с $m-2$ степенями свободы в случае выполнения гипотезы неколлинеарности признаков. Если задан некоторый уровень значимости и вычисленные допустимые уровни значимости $\mathrm{Cor}(\boldsymbol{\chi}_j, \boldsymbol{\chi}_k)$ меньше заданного уровня, то считается, что мультиколлинеарность велика. Отметим, что такие значения не обязательно являются следствием мультиколлинеарности.

**Факторы инфляции дисперсии.** Для оценки мультиколлинеарности строится линейная регрессия всех признаков, кроме $j$-го, на $j$-й признак:

$$\boldsymbol{\chi}_j = \mathbf{X}_{\mathcal{A}}\mathbf{w} + \boldsymbol{\varepsilon}_j, \quad \text{где } \mathcal{A} = \{1, \ldots, n\} \setminus \{j\}.$$

Коэффициенты регрессии вычисляются с помощью метода наименьших квадратов.

Рассмотрим дисперсию регрессионных остатков $\sigma_\varepsilon^2$ при условии их гомоскедастичности

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}). \tag{58}$$

Значение фактора VIF для $j$-го параметра определим как

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

где $R_j^2$ — коэффициент детерминации, вычисленный для $j$-го признака:

$$R_j^2 = 1 - \frac{\|\boldsymbol{\varepsilon}_j\|^2}{\|\boldsymbol{\chi}_j - \bar{x}_j\mathbf{1}\|^2},$$

где $\bar{x}_j$ — среднее значение $j$-го признака. Дисперсия $j$-го параметра при этом равна

$$\sigma^2(w_j) = \mathrm{VIF}_j \frac{\sigma_\varepsilon^2}{\|\boldsymbol{\chi}_j - \mathbf{1}\bar{x}_j\|^2}.$$

Наличие мультиколлинеарности определяется по значениям VIF. Если $\mathrm{VIF} > 10$, то считается [169], что мультиколлинеарность велика. На рис. 11 показано изменение VIF для четырех признаков. Каждому значению модельного параметра $\kappa \in (0, 1)$ соответствует набор значений факторов инфляции дисперсии. При изменении параметра признаки принимают следующие значения. Векторы $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ фиксированы. Векторы $\boldsymbol{\chi}_3, \boldsymbol{\chi}_4$ приближаются соответственно к $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ так, что $\kappa = \cos(\boldsymbol{\chi}_1, \boldsymbol{\chi}_3) = \delta + \cos(\boldsymbol{\chi}_2, \boldsymbol{\chi}_4)$, где $\delta$ — некоторая небольшая константа. Видно, что при увеличении параметра $\kappa$ значение VIF может расти неограниченно. Сильная мультиколлинеарность влечет бесконечные значения VIF. Основной недостаток данного метода заключается в том, что коэффициент детерминации может принимать бесконечные значения сразу для многих значений индекса признака $j \in \mathcal{J}$.

**Обнаружение мультиколлинеарности признаков с использованием сингулярного разложения.** Рассмотрим приближенное линейное описание[352] матрицы $\mathbf{X} = [x_{ij}]$ вида

$$x_{ij} = \sum_{k=1}^{r} u_{ik}\lambda_k v_{kj} + c_{ij}, \tag{59}$$

где $i \in \{1, \ldots, m\}$ и $j \in \{1, \ldots, n\}$. Приближенное линейное описание (59) более подробно описано в [355]. Значения $u_{kj}, \lambda_k, v_{jk}$ для данного значения $k$ находятся из условия минимума выражения

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}^2 \to \min, \tag{60}$$

при ограничениях нормировки

$$\sum_{k=1}^{n} u_{ik}^2 = \sum_{k=1}^{n} v_{kj}^2 = 1 \tag{61}$$

и упорядоченности $\lambda_1 \geqslant \ldots \geqslant \lambda_r \geqslant \ldots \geqslant 0$, $i \in \{1, \ldots, m\}$ и $j \in \{1, \ldots, n\}$.

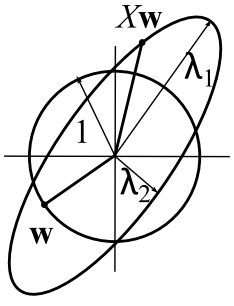



Рис. 12. Сингулярные числа матрицы плана.

Запишем выражения (59), (60) и(61) в матричных обозначениях:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^\intercal + \mathbf{C},$$
$$\mathrm{tr}(\mathbf{C}\mathbf{C}^\intercal) = \|\mathbf{C}\|^2 \to \min,$$
$$\mathbf{U}^\intercal\mathbf{U} = \mathbf{V}\mathbf{V}^\intercal = \mathbf{I},$$

где $\mathbf{U} = [u_{ik}], \mathbf{\Lambda} = \mathrm{diag}[\lambda_k], \mathbf{V} = [v_{kj}]$. Если значение $r$ достаточно велико, то $\mathbf{C} = \mathbf{0}$. Так будет заведомо при $r \geq \min\{m, n\}$. Минимальное значение $r$, при котором выполнимо равенство $\mathbf{X} = \mathbf{U\Lambda V}^\intercal$, равно рангу матрицы $\mathbf{X}$.

В работах [355, 331] утверждается, что задача минимизации выражения (60) при условии (61) эквивалентна задаче приближенного представления функции двух переменных $f(x_1, x_2)$ суммой попарных произведений $\sum_i g_i(x_1)h_i(x_2)$ функций $g_i(x)$ и $h_i(y)$ одной переменной.

Рассмотрим квадратичный алгоритм решения этой задачи. Найдем последовательно векторы $\mathbf{u}_k, \mathbf{v}_k$ и сингулярные числа $\lambda_k$ для $k = 1, \ldots, r$. В качестве этих векторов берутся нормированные значения векторов $\mathbf{a}_k$ и $\mathbf{b}_k$, соответственно

$$\underset{m\times 1}{\mathbf{u}_k} = \frac{\mathbf{a}_k}{\|\mathbf{a}_k\|} \quad \text{и} \quad \underset{1\times n}{\mathbf{v}_k} = \frac{\mathbf{b}_k}{\|\mathbf{b}k\|}.$$

Векторы $\mathbf{a}_k$ и $\mathbf{b}_k$ находятся как пределы последовательностей векторов $\{\mathbf{a}_{k_s}\}$ и $\{\mathbf{b}_{k_s}\}$, соответственно

$$\mathbf{a}_k = \lim_{s\to\infty}(\mathbf{a}_{k_s}) \ \text{ и } \mathbf{b}_k = \lim_{s\to\infty}(\mathbf{b}_{k_s}).$$

Сингулярное число $\lambda_k$ находится как произведение норм векторов

$$\lambda_k = \|\mathbf{a}_k\| \cdot \|\mathbf{b}_k\|.$$

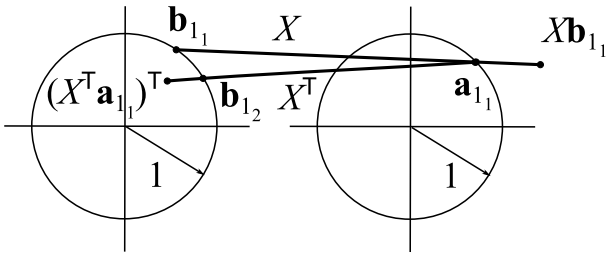



Рис. 13. Итеративная процедура оценивания сингулярных векторов.

Процедура нахождения последовательностей векторов $\mathbf{a}_{k_s}, \mathbf{b}_{k_s}$ $\mathbf{u}_k, \mathbf{v}_k$ начинается с выбора наибольшей по норме строки $\mathbf{b}_{1_1}$ матрицы $\mathbf{X}$. Для $k = 1$ формулы нахождения векторов $\mathbf{a}_{1_s}, \mathbf{b}_{1_s}$ имеют вид:

$$\underset{m\times 1}{\mathbf{a}_{1_s}} = \frac{\mathbf{X}\mathbf{b}_{1_s}^\intercal}{\mathbf{b}_{1_s}\mathbf{b}_{1_s}^\intercal}, \quad \underset{1\times n}{\mathbf{b}_{1_{s+1}}} = \frac{\mathbf{a}_{1_s}^\intercal \mathbf{X}}{\mathbf{a}_{1_s}^\intercal \mathbf{a}_{1_s}}, \quad s = 1, 2, \cdots.$$

Для вычисления векторов $\mathbf{u}_k, \mathbf{v}_k$ при $k = 2, \ldots, r$ используется вышеприведенная формула, с той разницей, что матрица $\mathbf{X}$ заменяется на скорректированную на $k$-м шаге матрицу $\mathbf{X}_{k+1} = \mathbf{X}_k - \mathbf{u}_k\lambda_k\mathbf{v}_k$. На рисунке 13 показаны две итерации, $s = 1, 2$, первого шага $k = 1$ упрощенной процедуры нахождения сингулярного разложения.

Для исследования мультиколлинеарности признаков рассмотрим сингулярное разложение матрицы $\mathbf{X}$. Пусть матрица признаков $\mathbf{X}$ имеет размерность $m \times n$, центрирована и нормирована. Выполним сингулярное разложение [299, 138, 154, 148, 334] матрицы $\mathbf{X}$,

$$\mathbf{X} = \mathbf{U\Lambda V}^\intercal, \tag{62}$$

где $\mathbf{U}$,$\mathbf{V}$ — ортогональные матрицы размерности, соответственно, $m \times n$ и $n \times n$ и $\mathbf{\Lambda}$ — диагональная матрица с сингулярными числами на диагонали, такими, что

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n > 0. \tag{63}$$

Столбцы матрицы $\mathbf{V}$ являются собственными векторами, а квадраты сингулярных чисел — собственными значениями матрицы $\mathbf{X}^\intercal\mathbf{X}$:

$$\mathbf{X}^\intercal\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{U}^\intercal\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\intercal.$$

Из предыдущего выражения получим

$$\mathbf{X}^\intercal\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2. \tag{64}$$

Найдем матрицу, псевдообратную матрице $\mathbf{X}^\intercal\mathbf{X}$ в номальном уравнении (38), используя выражение (64). Для псевдообратной матрицы $\mathbf{X}^+$, такой, что

$$\begin{aligned}
\mathbf{X}\mathbf{X}^+\mathbf{X} &= \mathbf{X}, \\
\mathbf{X}^+\mathbf{X}\mathbf{X}^+ &= \mathbf{X}^+, \\
(\mathbf{X}\mathbf{X}^+)^\intercal &= \mathbf{X}\mathbf{X}^+, \\
(\mathbf{X}^+\mathbf{X})^\intercal &= \mathbf{X}^+\mathbf{X}
\end{aligned}$$

справедливо выражение

$$\mathbf{X}^+ = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^\intercal.$$

По условию, матрица $\mathbf{X}$ невырождена. Следовательно, матрица $\mathbf{X}^\intercal\mathbf{X}$ также не является вырожденной. Числа обусловленности этих матриц относятся как

$$\varkappa(\mathbf{X}) = \frac{\lambda_1}{\lambda_n} = \sqrt{\varkappa(\mathbf{X}^\intercal\mathbf{X})},$$

то

$$(\mathbf{X}^\intercal\mathbf{X})^{-1} = \mathbf{V}^\intercal\mathbf{\Lambda}^{-2}\mathbf{V}. \tag{65}$$

Матрица $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ называется матрицей Мура-Пенроуза [34] при условии полного ранга $\mathbf{X}$. При использовании этой уравнение (38) принимает вид

$$\mathbf{w} = (\mathbf{X}^\intercal)^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{X}^+\mathbf{y}. \tag{66}$$

Используя (62), (65) и (66) получим выражение

$$\mathbf{w} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^\intercal\mathbf{y}. \tag{67}$$

Оценка ковариационной матрицы параметров $\mathbf{w}$ при условии (58) имеет вид

$$\mathrm{Cov}(\mathbf{w}) = \sigma_\varepsilon^2(\mathbf{X}^\intercal\mathbf{X})^{-1}. \tag{68}$$

Из (67), используя индексацию, принятую в (59), получим

$$w_j = \sum_{k=1}^{n} \frac{v_{kj}}{\lambda_k} \sum_{i=1}^{m} u_{ij} y_i.$$

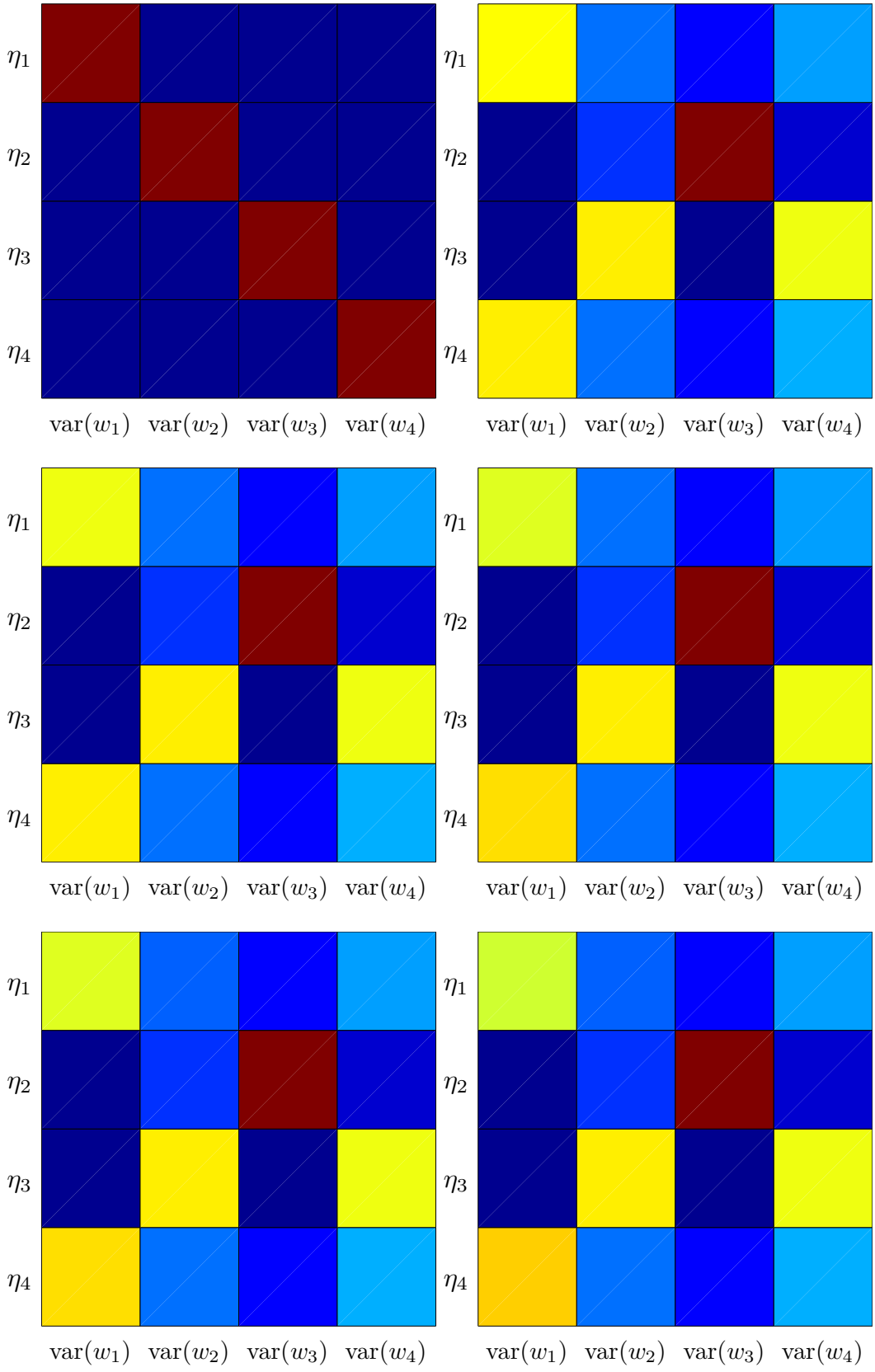



Рис. 14. Матрица значений долевых коэффициентов для различных значений параметра $\kappa$.

Таблица 6. Индексы обусловленности и дисперсии параметров.

| | $\sigma^2(w_1)$ | $\dots$ | $\sigma^2(w_j)$ | $\dots$ | $\sigma^2(w_n)$ |
|---|---|---|---|---|---|
| $\eta_1$ | $q_{11}$ | $\cdots$ | $q_{21}$ | $\cdots$ | $q_{n1}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $\eta_k$ | $q_{1k}$ | $\cdots$ | $q_{jk}$ | $\cdots$ | $q_{nk}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $\eta_n$ | $q_{1n}$ | $\cdots$ | $q_{2n}$ | $\cdots$ | $q_{nn}$ |
| $\sum\limits_{k=1}^{n} q_{jk}$ | 1 | $\dots$ | 1 | $\dots$ | 1 |

Представим ковариационную матрицу параметров (68) в виде

$$\mathrm{Cov}(\mathbf{w}) = \sigma_{\varepsilon}^2 \begin{bmatrix} \sum_{k=1}^{n} \frac{v_{1k}^2}{\lambda_k^2} & \cdots & \sum_{k=1}^{n} \frac{v_{1k}v_{nk}}{\lambda_k^2} \\ \vdots & \ddots & \vdots \\ \sum_{k=1}^{n} \frac{v_{nk}v_{1k}}{\lambda_k^2} & \cdots & \sum_{k=1}^{n} \frac{v_{nk}^2}{\lambda_k^2} \end{bmatrix}.$$

Дисперсия $j$-го элемента вектора параметров $\mathbf{w}$ имеет вид

$$\sigma^2(w_j) = \sigma_{\varepsilon}^2 \sum_{k=1}^{n} \frac{v_{jk}^2}{\lambda_k^2}.$$

Матрица корреляций вектора параметров имеет вид

$$\mathrm{Cor}(\mathbf{w}) = Q^{-\frac{1}{2}} \mathrm{Cov}(\mathbf{w}) Q^{-\frac{1}{2}},$$

где $Q = \mathrm{diag}(\mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V})$ — диагональ матрицы $\mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}$.

Назовем индексом обусловленности $\eta_k$ сингулярного разложения отношение максимального сингулярного числа к $k$-му сингулярному числу

$$\eta_k = \frac{\lambda_{\max}}{\lambda_k}. \tag{69}$$

Для обнаружения мультиколлинеарности признаков построим таблицу, в которой каждому индексу обусловленности $\eta_k$ соответствуют значения $q_{kj}$ — долевые коэффициенты, которые в сумме по индексу $k$ дают единицу. Они представляют собой части от общей величины $\sigma_{\varepsilon}^{-2}\sigma(w_j)$.

Представим диагональные элементы ковариационной матрицы параметров

$$\mathrm{Cov}(\mathbf{w}) = \sigma_{\varepsilon}^2 (\mathbf{X}^{\intercal}\mathbf{X})^{-1} = \sigma_{\varepsilon}^2 \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{\intercal}.$$

в следующем виде:

$$\begin{aligned} \sigma^2(w_j) &= \sigma_{\varepsilon}^2 \left( \frac{v_{j1}^2}{\lambda_1^2} + \cdots + \frac{v_{jk}^2}{\lambda_k^2} + \cdots + \frac{v_{jn}^2}{\lambda_n^2} \right) = \\ &= \sigma_{\varepsilon}^2 (q_{j1} + \cdots + q_{jk} + \cdots + q_{jn}) \sum_{k=1}^{n} \frac{v_{jk}^2}{\lambda_k^2} = \\ &= \sigma_{\varepsilon}^2 \sum_{k=1}^{n} q_{jk} \sum_{k=1}^{n} \frac{v_{jk}^2}{\lambda_k^2}, \end{aligned}$$

где $q_{jk}$ — отношение соответствующего слагаемого в разложении $\sigma_{\varepsilon}^{-2}\sigma^2(w_j)$ ко всей сумме. Наличие мультиколлинеарности определяется по таблице: большие величины $\eta_k$ означают, что, возможно, есть зависимость между признаками. Большие значения $q_{jk}$ в соответствующих строках относятся к признакам, между которыми эта зависимость существует.

Из определения (69) и неравенства (63) следует, что индекс обусловленности монотонно возрастает. Следовательно, наибольшие значения индекса $k$ будут соответствовать наибольшим значениям индекса обусловленности $\eta_k$. Для фильтрации мультиколлинеарных признаков выделим последние несколько элементов $\eta_{\hat{k}}, \dots \eta_n$ в отдельную группу для анализа. Предлагается найти значение $\hat{k}$ как минимум второй частной разности:

$$\hat{k} = \underset{k\in\{1,\dots,n-2\}}{\arg\min}\ (\eta_k - 2\eta_{k+1} + \eta_{k+2}).$$

Для того, чтобы найти признак $\hat{j}$, который требуется отфильтровать, предлагается решить задачу

$$\hat{j} = \underset{j\in\{1,\dots,n\}}{\arg\max} \sum_{k=n-\hat{k}+1}^{n} q_{jk}.$$

Другими словами, находится индекс признака, доставляющего максимальную сумму долевых коэффициентов $q_{jk}$, соответствующих множеству выделенных индексов обусловленности $\eta_{\hat{k}}, \dots, \eta_n$.

Рисунок 14 показывает изменение значений долевых коэффициентов, которые вычисляется по синтетическим данным. Углы между четырьмя векторами $\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_4$ — столбцами матрицы $\mathbf{X}$ изменяются так, что угол между первым и третьим столбцом, задаваемый параметром $\kappa$, уменьшается.

Основными методами устранения мультиколлинеарности являются выбор признаков, либо введение ограничений на значения параметов модели [10, 56, 231, 128]. Более подробно эти методы будут рассмотрены далее.

# 2. Порождение моделей

Задача восстановления регрессии включают в себя не только методы выбора оптимальной параметрической регрессионной модели, но и методы порождения таких моделей. Множества измеряемых признаков зачастую бывает недостаточно для построения модели удовлетворительного качества. Преблагается расширить множество признаков с помощью функциональных преобразований исходных признаков с целью повышения адекватности модели. Методы порождения имеют большую историю: в 1968 году А. Г. Ивахненко предложил метод группового учета аргументов, МГУА [309, 308, 307, 192]. Согласно этому методу регрессионная модель, доставляющая наилучшее приближение, выбирается из множества последовательно порождаемых моделей. Множество порождаемых моделей задавалось набором мономов полинома Колмогорова-Габора ограниченной степени. В частности, для построения моделей как суперпозиций функций использовались полиномиальные функции, ряды Фурье и некоторые другие функции, например многослойный персептрон, функции радиального базиса, полиномы Лагранжа, полиномы Чебышёва [300, 307]. При выборе моделей использовался скользящий контроль [162].

При порождении существенно нелинейных моделей используются методы генетического программирования и символьной регрессии [285, 156, 167, 193]. Согласно этим методам, регрессионные модели порождаются как произвольные нелинейных суперпозиции порождающих функций (англ. primitives) [193]. Для визуализации порожденных моделей используют регрессионные деревья [132, 42]. При поиске деревьев оптимальной структуры посредством генетического программирования используется их векторное представление, где число элементов вектора непостоянно и определяется структурой модели [199, 9, 110, 229].

Одним из методов для решения задачи восстановления функциональной зависимости по набору исходных данных является символьная регрессия [229, 273, 272, 229]. Джон Коза предложил реализацию этого метода с помощью аналога эволюционного алгоритма [166]. Иван Зелинка предложил дальнейшее развитие этой идеи [285], получившее название аналитического программирования. Алгоритм построения математической модели в аналитическом программировании выглядит следующим образом: задан набор элементарных функций (например, степенная функция, $+$, sin, tan и др.), из которых можно строить различные формулы. Начальный набор формул строится либо произвольным образом, либо на базе некоторых предположений эксперта. Затем на каждом шаге производится оценка каждой из формул согласно некоторой функции качества. На базе этой оценки у части формул случайным образом заменяется одна элементарная функция на другую (например, sin на cos или $+$ на $\times$), а у некоторой другой части происходит взаимный попарный обмен подвыражениями. Данный подход может быть описан в терминах эволюционного алгоритма: каждый индивид является формулой, изображенной в свою очередь в виде дерева. Тогда набор формул, существующий в определенный момент, представляет собой одно поколение. При этом хромосомы представляются поддеревьями, и, в отличие от классического генетического алгоритма, могут быть различного размера (длины). Описанный выше обмен подвыражениями представляет собой в этом случае генетическое скрещивание, замена одной элементарной функции у некоторых деревьев — мутацию. При этом возникает ряд сложностей, связанных

с областями определения и арностями элементарных функций, записанных в узлах дерева. Данный метод фактически является ненаправленным поиском и перебирает большое количество неподходящих деревьев до того момента, как приблизится к оптимуму.

Альтернативой аналитическому программированию можно считать подход обучения в глубину (Deep Learning) [36, 24]. Этот подход заключается в иерархическом представлении данных, в котором на нижнем уровне находятся сам набор данных, а на каждом уровне выше — более абстрактное его представление, которое представляет собой некую скрытую комбинацию из данных, указанных ниже. Так, например, при использовании данного метода в обработке изображений, набором данных является матрица яркости пикселей некоторого изображения, на следующем уровне — данные о выраженных геометрических закономерностях на изображении (отрезки, кривые, окружности), на более высоких уровнях иерархии — более сложные и абстрактные выявленные закономерности. В одном из основных алгоритмов, использующих данный подход, иерархия строится при помощи нейронной сети с несколькими скрытыми слоями [37]. В одном из основных методов обучения в глубину нейронная сеть обучается, получая на вход и на выход одинаковый набор данных, после чего каждый из уровней сети представляется как информация о данных на определенном уровне абстракции.

В данной работе предлагается рассмотреть метод построения математической модели, основанный на прогнозировании структуры функциональной зависимости. Предполагается, что функциональная зависимость существенно нелинейна и, аналогично описанному выше, является суперпозицией элементарных функций. При этом делается ограничение на максимальную сложность модели. Дерево суперпозиции представляется в виде матрицы. В таком виде задача сводится к задаче структурного обучения, описанной, например, в [171, 201, 150]. Методы структурного обучения решают задачу нахождения структуры или зависимости, имеющейся внутри исходных данных. Метод широко применим для синтаксического разбора предложений [149], компьютерного зрения [172].

В работе И. Зелинки [285] описаны основные проблемы, возникающие при порождении моделей методами символьной регрессии. В частности, при порождении моделей могут появиться ошибки

1) структуры модели (например, функции, которые неявно зависят от самих себя);

2) несоответствия области определения и области значения (например, функции имеют мнимые области значений);

3) неограниченный рост значений (например, возникновение деления на ноль);

4) избыточности структуры модели (например, умножение на ноль некоторой функции).

Метод, предлагаемый в данной работе, избавлен от вышеперечисленных проблем. Он заключается в следующем, см. рис. 15. Поиск моделей выполняется по итерационной схеме «порождение-выбор» в соответствии с определенными правилами порождения моделей и критерием выбора моделей. Последовательно порождаются наборы конкурирующих моделей. Каждая модель в наборе является суперпозицией элементов заданного множества

гладких параметрических функций. После построения модели каждому элементу суперпозиции ставится в соответствие гиперпараметр. Параметры и гиперпараметры модели последовательно настраиваются. Из набора выбираются наилучшие модели для последующей модификации. При модификации моделей, по значениям гиперпараметров делаются выводы о целесообразности включения того или иного элемента в модель следующего порождаемого набора.

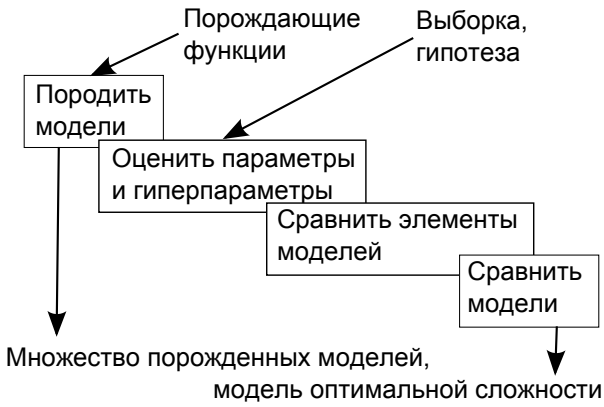



Рис. 15. Процедура индуктивного порождения и выбор моделей.

**Процедура индуктивного порождения и выбора моделей.** Рассматривается итеративная процедура порождения-выбора моделей. Основные этапы построения модели показаны на рисунке ниже.

1. Задана выборка — набор реализаций множества независимых и одной зависимой переменной. Задан набор порождающих функций.

2. Задан набор исходных моделей. Исходные модели могут быть получены в качестве произвольных суперпозиций заданных порождающих функций. Параметры и гиперпараметры моделей настраиваются в соответствии с оптимизационным методом. Тип метода зависит от суперпозиции: это может быть метод сопряженных градиентов, метод стохастической оптимизации или метод наименьших квадратов. Настройка параметров и гиперпараметров выполняется итеративно.

3. Для каждой модели оцениваются веса элементов суперпозиции. Веса зависят от значений гиперпараметров.

4. Производится выбор наилучших моделей в соответствии с функцией ошибки. Функция ошибки зависит от гипотезы порождения данных — исходных предположений о распределении зависимой переменной и параметров модели.

5. Выбранные модели модифицируются, и порождаются новые модели в соответствии с правилами порождения. Поскольку число порождаемых моделей, вообще говоря, счетно, вводятся дополнительные ограничения на правила порождения моделей. Для того, чтобы обеспечить разнообразие порождаемых моделей, используется набор порождающих функций. Информация о том, как модифицировать модели для улучшения качества, заключена

в гиперпараметрах модели. Алгоритм останавливается при достижении требуемого качества, или при достижении заданного количества порождаемых моделей.

Методы индуктивного порождения регрессионных моделей, например, метод группового учета аргументов или символьная регрессия, используют в качестве моделей-претендентов различные суперпозиции свободных переменных. В частности, МГУА использует линейные комбинации произведений свободных переменных, а символьная регрессия — их произвольные суперпозиции.

Будем считать, что термин «суперпозиция», используемый в этой работе, является синонимом терминов «формула», «выражение», «математическое выражение», используемых в работах [85, 86, 211, 229] обозначают некоторую композицию функций, свободных переменных и констант.

Среди множества индуктивно-порожденных суперпозиций могут присутствовать избыточные суперпозиции [252, 253, 259]. В них есть подвыражения, которые могут быть удалены или заменены на более простые; при этом отображение из пространства свободных переменных в пространство зависимых переменных остается неизменным. Проблема решается методами упрощения суперпозиций, например, методом упрощения выражений по правилам [85, 86] и методом замены поддеревьев на эквивалентные поддеревья меньшей сложности [211]. Суперпозиции представляются в виде направленного ациклического графа с объединением общих поддеревьев. Такое представление позволяет существенно расширить класс допустимых преобразований суперпозиций.

Ниже предлагается алгоритм, основанный на упрощении суперпозиций по правилам. При этом суперпозицией называется некоторая композиция элементарных функций, а правила описывают, как функции связаны между собой. В качестве примера таких правил можно указать $\log \circ \exp \equiv \mathrm{id}$ или $t^n \times t^m \equiv t^{n+m}$.

В работах [239, 253, 59, 257, 258] предлагается представлять суперпозиции в виде соответствующего им дерева, над которым и оперировали предлагавшиеся алгоритмы. В работе Г. И. Рудого [338] суперпозиция представляется не в виде дерева, а в виде направленного ациклического графа, где различные функции могут принимать в качестве аргумента одно и то же подвыражение. Примером может являться суперпозиция $\cos^2 t + \sin^2 t$, где $t$ — некое сложное подвыражение. Таким образом, искомая задача сводится к задаче нахождения общих подвыражений в исходном дереве суперпозиции и задания правил упрощения на множестве подобных ациклических графов.

В работах А. Н. Колмогорова [163] сформулирована следующая теорема, приведенная здесь в локальных обозначениях: «Каждая непрерывная функция $n$ переменных, заданная на единичном кубе $n$-мерного пространства, представима в виде

$$f(\mathbf{x}) = \sum_{i=1}^{2n+1} h_i \left( \sum_{j=1}^{n} g_{ij}(x_j) \right), \quad \text{где} \quad \mathbf{x} = [x_1, \ldots, x_n]^\intercal,$$

где функции $h_g(\xi)$ непрерывны, а функции $g_{ij}$ также непрерывны и зависят от выбора функции $f$». Отсюда следует, что класс порождаемых регрессионных моделей должен быть ограничен двухслойными нейронными сетями. Однако при решении прикладных задач, особенно

задач математического моделирования, структура модели должна быть интепретируема экспертом в контексте моделируемого явления. Поэтому далее при постановке задач мы будем считать, что множество порождающих функций и вид индуктивно-порождаемых моделей заданы экспертно.

## 2.1. Допустимые суперпозиции

Рассмотрим некоторые алгоритмы, порождающий все возможные суперпозиции заданной сложности за конечное число шагов. Для описания алгоритмов, задающих классы моделей $\mathfrak{F}$, введем необходимые обозначения.

### 2.1.1. Порождающие функции и их суперпозиции

Рассмотрим две функции $g : \mathbb{X} \to \mathbb{Y}$ и $h : \mathbb{Y}' \supseteq \mathbb{Y} \to \mathbb{Z}$ и пусть $\mathbb{Y}' \cup \mathbb{Y} \neq \emptyset$. Их композицией называется функция $f = g \circ h : \mathbb{X} \to \mathbb{Z}$, определенная равенством

$$(h \circ g)(\mathbf{x}) = h(g)(\mathbf{x}),\ \mathbf{x} \in \mathbb{X}.$$

Пусть задано множество $G = \{g_i\}$ функций. Для каждой функции $g_i$ задана область определения $\mathbb{X}_i = \mathrm{dom}(g_i)$ и область значения $\mathbb{Y}_i = \mathrm{cod}(g_i)$. Пусть множество значений $\mathbb{Y}_i$ функции $g_i$ содержится в области определения $\mathbb{X}_{i+1}$ функции $g_{i+1}$, то есть

$$g_i : \mathbb{X}_i \to \mathbb{Y}_i \subseteq \mathbb{X}_{i+1}, \quad i = 1, 2, \ldots, K-1, \tag{70}$$

то функция

$$f = g_K \circ g_{K-1} \circ \cdots \circ g_1, \quad K \geq 2, \tag{71}$$

определяемая равенством

$$f(\mathbf{x}) = (g_K \circ g_{K-1} \circ \cdots \circ g_1)(\mathbf{x}) = g_K\big(g_{K-1}\big(\ldots(g_1)\big)\big)(\mathbf{x}), \quad \mathbf{x} \in \mathbb{X}, \tag{72}$$

называется *сложной функцией* или *суперпозицией функций* $g_1, g_2, \ldots, g_K$.

**Определение 9.** *Суперпозиция $f$ функций $\{g_1, \ldots, g_K\}$ — функция, представленная как композиция нескольких функций, определяемая выражениями (71–72) при выполнении условия* (70).

Функции $g = g(\mathbf{b}, \xi)$ с параметрами $\mathbf{b}$ и аргументом $\xi$, принадлежащие множеству $G$, далее будут называться *порождающими функциями*.

**Определение 10.** *Допустимой суперпозицией $f$ называется такая суперпозиция, в которой*

$$cod(g_{i(k+1)}) \subseteq dom(g_{i(k)}) \quad для\ всех \quad k = 1, \ldots, K-1.$$

Для обобщения этого определения случай функций нескольких аргументов будем считать, что функции $g_1, \ldots, g_K$, входящие в суперпозицию, являются вектор-функциями от векторных величин $\boldsymbol{\xi}$. При этом и области определения $\mathbb{X}_i$, и области значений $\mathbb{Y}_i$ этих вектор-функций являются подмножествами декартова произведения пространств соответствующих аргументов.

Пусть задано множество порождающих функций $G = \{g_1, \ldots, g_l | g = g(\mathbf{b}, \cdot)\}$, то есть, заданы

1) сама функция $g : (\mathbf{b}, \xi) \mapsto \xi'$,

2) число ее параметров $\mathbf{b}$ (возможен пустой набор),

3) число аргументов (арность) $v(g)$ функции $g$ (возможен пустой набор) и порядок следования аргументов,

4) домен $\mathrm{dom}(g)$ и кодомен $\mathrm{cod}(g)$.

Требуется построить функцию $f$ как суперпозицию порождающих функций из заданного множества $G$. Модель $f(\mathbf{w}, \mathbf{x})$ рассматривается как суперпозиция

$$f(\mathbf{w}, \mathbf{x}) = (g_{i(1)} \circ \cdots \circ g_{i(K)})(\mathbf{x}), \quad \text{где} \quad \mathbf{w} = [\mathbf{b}_{i(1)}^{\mathsf{T}}, \ldots, \mathbf{b}_{i(K)}^{\mathsf{T}}]^{\mathsf{T}},$$

в которой вектор $\mathbf{w}$ состоит из присоединенных векторов-параметров $\mathbf{b}$ функций $g$, водящих в суперпозицию $f$.

Для порождения моделей требуется задать:

1) множество непорождаемых переменных $\{\xi\}$ с заданным $\mathrm{dom}(\xi)$,

2) множество порождающих функций $G = \{g_u, \mathrm{id}\}$, $g : x \mapsto x'$,

3) правило Gen порождения допустимых суперпозиций $\mathfrak{G} \supset G$, где суперпозиция $g_j = g_u \circ g_v \in \mathfrak{G}$, построена с учетом ограничений

   на число аргументов $v(g_u)$,

   на область определения $\mathrm{cod}(g_u)$,

   на структурную сложность суперпозиции $C(g_j) \leqslant C_{\max}$,

   на число и типы входных и выходных переменных,

4) правило Rem упрощения суперпозиций: $g_k \notin \mathfrak{G}$, если

$$\mathrm{Rem} : g_k = g_u \circ g_v \mapsto g_j \in \mathfrak{G}.$$

Результатом порождения допустимых суперпозиций является набор $\mathfrak{F}$ моделей-претендентов $f$, из которого производится выбор.

Поставим в во взаимно-однозначное соответствие каждой суперпозиции $f$ дерево $\Gamma_f$, которое строится следующим образом:

1) в вершинах $V_i$ дерева $\Gamma_f$ находятся соответствующие порождающие функции $g_s, s = s(i)$;

2) число дочерних вершин у некоторой вершины $V_i$ равно арности соответствующей функции $g_s$;

3) порядок вершин, дочерних для $V_i$, соотвествует порядку аргументов соответствующей функции $g_{s(i)}$;

4) в листьях дерева $\Gamma_f$ находятся свободные переменные $x_i$.

Вычисление значения выражения $f = f(\mathbf{w}, \mathbf{x})$ в некоторой точке $\mathbf{x}$ с заданным вектором параметров $\mathbf{w} = \{w_1, w_2, \ldots, w_\eta\}$ эквивалентно подстановке соответствующих значений свободных переменных $\mathbf{x}$ и параметров $\mathbf{w}$ функцию $f$, соответствующую дереву $\Gamma_f$. Каждое поддерево $\Gamma_f^i$ дерева $\Gamma_f$, соответствующее вершине $V_i$, также соответствует некоторой суперпозиции, являющейся составляющей исходной суперпозиции $f$.

### 2.1.2. Условия допустимости суперпозиций

Методы индуктивного построения регрессионных моделей используют в качестве моделей-претендентов различные суперпозиции свободных переменных. Рассмотрим условия при которых суперпозиции являются допустимыми. Функции $g_v \in G$ проиндексированы числами $v \in \mathcal{V} = \{1, ..., V\}$. Задано отображение $\iota : \mathcal{V}^R \to \mathcal{A}$. Элементы $A_\iota \in \mathcal{A}$ — всевозможные сочетания с повторениями из $V$ по $K$, где $K = 1, \ldots, R$. Мощность множества $\mathcal{A}$ равна

$$|\mathcal{A}| = \sum_{K=1}^{R} \bar{C}_K^V = \sum_{K=1}^{R} \frac{(K+V-1)!}{(K-1)!V!},$$

где $\bar{C}$ — число сочетаний с повторениями.

Элементы набора $A_\iota = \{a_\iota(k)\}$ проиндексированы числами $k = 1, \ldots, K_\iota$. Так как $a \in \mathcal{V}$, элементы $a_\iota(k)$ однозначно соответствуют функциям $g_v$ из $G$. Каждому набору $A_\iota$ поставим в соответствие набор матриц инцидентности $\{\rho_i(A_\iota)\}, i \in \mathbb{N}$. Индекс $i$ матрицы $\rho$ задает уникальную суперпозицию $f_i$ функций $g$ из $G$; обозначим $\rho_i = \rho_i(A_\iota)$. Число элементов этой суперпозиции равно $K_\iota$. Матрица инцидентности

$$\rho_i : \{1, \ldots, K_\iota\} \times \{1, \ldots, K_\iota\} \to \{0, 1\}$$

задает орграф и суперпозицию функций $f_i$ нескольких аргументов. Суперпозиция $f_i$ называется *допустимой*, если выполнены следующие условия.

1. Орграф $\rho_i$ является ациклическим.

2. Орграф является односвязным без изолированных вершин, то есть справедливо равенство

$$\sum_{k=1}^{K_\iota} \sum_{l=1}^{K_\iota} \rho_i(l,k) = \sum_{k=1}^{K_\iota} s(a_\iota(k)),$$

   где $s = s(v)$ — число аргументов функции $g_v$. Число единиц в орграфе $\rho_i$ равно суммарному числу аргументов в суперпозиции $f_i$.

3. Число аргументов каждого элемента суперпозиции должно совпадать с числом аргументов соответствующей порождающей функции

$$\sum_{l=1}^{K_\iota} \rho_i(l,k) = s(a_\iota(k)), \quad \text{для всех} \quad k = 1, \ldots, K_\iota.$$

   Число вершин орграфа, смежных вершине с номером $k$, есть число $s(a_\iota(k))$ аргументов функции $g_v$ при $v = a_\iota(k)$.

### 2.1.3. Порождение произвольных суперпозиций

Опишем алгоритм, порождающий произвольную суперпозицию конечной глубины за конечное число шагов. *Глубина суперпозиции* $f$ — максимальная глубина дерева $\Gamma_f$.

Пусть дано множество примитивных функций $G = \{g_1, \ldots, g_l\}$ и множество свободных переменных $X = \{x_1, \ldots, x_n\}$. Для удобства будем исходить из предположения, что множество $G$ состоит только из унарных и бинарных функций, и разделим его соответствующим образом на два подмножества: $G = G_b \cup G_u \mid G_b = \{g_{b_1}, \ldots, g_{b_k}\}, G_u = \{g_{u_1}, \ldots, g_{u_l}\}$, где $G_b$ — множество всех бинарных функций, а $G_u$ — множество всех унарных функций из $G$. Потребуем также наличия id в $G_b$.

Алгоритм итеративного порождения суперпозиций.

1. Перед первым шагом зададим начальные значения множества $\mathfrak{F}_0$ и вспомогательного индексного множества $\mathcal{I}$, служащего для запоминания, на какой итерации впервые встречена каждая суперпозиция:
$$\mathfrak{F}_0 = X,$$
$$\mathcal{I} = \{(x, 0) \mid x \in X\}.$$

2. Для множества $\mathfrak{F}_i$ построим вспомогательное множество $U_i$, состоящее из суперпозиций, полученных в результате применения функций $g_u \in G_u$ к элементам $\mathfrak{F}_i$:
$$U_i = \{g_u \circ f \mid g_u \in G_u, f \in \mathfrak{F}_i\}.$$

3. Аналогичным образом построим вспомогательное множество $B_i$ для бинарных функций $g_b \in G_b$:
$$B_i = \{g_b \circ (f, h) \mid g_b \in G_b, f, h \in \mathfrak{F}_i\}.$$

4. Обозначим $\mathfrak{F}_{i+1} = \mathfrak{F}_i \cup U_i \cup B_i$.

5. Для каждой суперпозиции $f$ из $\mathfrak{F}_{i+1}$ добавим пару $(f, i+1)$ в множество $\mathcal{I}_f$, если суперпозиция $f$ еще там не присутствует.

6. Перейдем к следующей итерации, п. 2.

Тогда $\mathfrak{F} = \cup_{i=0}^{\infty} \mathfrak{F}_i$ — множество всех возможных суперпозиций конечной длины, которые можно построить из данного множества примитивных функций.

Вспомогательное множество $\mathcal{I}$ позволяет запоминать, на какой итерации впервые встречается каждая суперпозиция. Это необходимо, так как каждая суперпозиция, впервые порожденная на $i$-ой итерации, будет порождена также и на любой итерации после $i$. Одной из возможностей избежать необходимости в этом множестве является построение $\mathfrak{F}_{i+1}$ как $\mathfrak{F}_{i+1} = U_i \cup B_i$ (без $\mathfrak{F}_i$), а множества $U_i$ и $B_i$ строить следующим образом:

$$U_i = \{g_u \circ f \mid g_u \in G_u, f \in \cup_{j=0}^{i} \mathfrak{F}_j\},$$

$$B_i = \{g_b \circ (f, h) \mid g_b \in G_b, f, h \in \cup_{j=0}^{i} \mathfrak{F}_j\}.$$

Алгоритм очевидным образом обобщается на случай, когда множество $G$ содержит функции произвольной (но конечной) арности. Действительно, для такого обобщения достаточно строить аналогичным образом вспомогательные множества для этих функций, а именно: для множества функций $G_n$ арности $n$ построим вспомогательное множество $H_i^n$ вида:

$$H_i^n = \{g \circ (f_1, f_2, \ldots, f_n) \mid g \in G_n, f_j \in \mathfrak{F}_i\}.$$

В этих обозначениях $U_i \equiv H_i^1$, а $B_i \equiv H_i^2$.

Тогда множество $\mathfrak{F}_{i+1} = \mathfrak{F}_i \cup_{n=0}^{n_{max}} H_i^n$, где $n_{max}$ — максимальное значение арности функций из $G$.

**Теорема 3.** *Вышеописанный алгоритм действительно породит любую конечную суперпозицию за конечное число шагов.*

Действительно, найдем номер итерации, на котором будет порождена некоторая конечная суперпозиция $f$. Для этого, пронумеруем вершины графа $\Gamma_f$ по следующим правилам:

1) если это вершина со свободной переменной, то она имеет номер 0;

2) если вершина $V$ соответствует унарной функции, то она имеет номер $i+1$, где $i$ — номер дочерней для этой функции вершины;

3) если вершина $V$ соответствует бинарной функции, то она имеет номер $i+1$, где $i = \max(l, r)$, а $l$ и $r$ — номера, соответственно, первой и второй дочерней вершины.

Номер вершины, соответствующей корню графа, будет номером итерации, на которой получена суперпозиция $f$. Иначе, для любой суперпозиции мы можем указать конкретный номер итерации, на котором она будет получена.

В предложенных ранее методах построения суперпозиций [285] требовалась отдельная проверка отсутствия частично-рекурсивных суперпозиций вида $f(x, y) = g(f(x, y), x, y)$ в ходе итеративного порождения. В вышеописанном алгоритме такие суперпозиции не могут возникнуть по построению.

### 2.1.4. Суперпозиции с дополнительными параметрами

При задании некоторых классов моделей (например, нейронных сетей) состоящих из порождающих функций, не включающих параметры, предлагается включать параметры $\mathbf{w}$ непосредственно в модель $f(\mathbf{w}, \mathbf{x})$ в процедуре индуктивного порождения. Пусть порождающие функции не имеют параметров. Модифицируем алгоритм порождения произвольных суперпозиций следующим образом:

$$U_i = g_u \circ (\alpha f + \beta),$$

$$B_i = g_b \circ (\alpha f + \beta, \psi h + \phi).$$

Здесь параметры $\alpha, \beta$ зависят только от комбинации $g_u, f$ (или $g_b, f, h$ для $\alpha, \beta, \psi, \phi$), индексы параметров опущены. Каждая суперпозиция из предшествующих итераций входит в следующие, будучи умноженной на некоторой коэффициент и с добавленной константой. При таком

включении параметров $\mathbf{w} = [\alpha, \beta, \ldots, \psi, \phi]^\intercal$ мощность полученного множества суперпозиций и свойства порождающего алгоритма остаются неизменными.

Этот алгоритм, как и предыдущий, может быть обобщен на случай порождающих функций произвольного конечного числа аргументов.

### 2.1.5. Порождение обобщенно-линейных моделей

Линейные модели

$$f(\mathbf{w}, \mathbf{x}) = \mathbf{w}^\intercal \mathbf{x} = \sum_{j=1}^{n} w_j x_j = \sum_{j \in \mathcal{A}} w_j x_j, \quad \text{где} \quad \mathcal{A} \subseteq \mathcal{J} = \{1, \ldots, n\}$$

порождаются путем перебора всех наборов слагаемых $w_j x_j$ при ограничении на их число. Так как при нулевом значении параметра $w_j = 0$ соответствующее слагаемое не учитывается в модели, то алгоритм порождения моделей можно представить в следующем виде. Пусть модель

$$f(\mathbf{w}, \mathbf{x}) = \alpha_1 w_1 x_1 + \alpha_2 w_2 x_2 + \cdots + \alpha_n w_n x_n.$$

Структурный параметр $\alpha_j$ не является частью модели и может принимать значение из множества $\{0, 1\}$. Для получения класса линейных моделей $\mathfrak{F}_{\text{lin}}$, имеющих не более $n$ слагаемых необходимо перебрать все значения вектора $\boldsymbol{\alpha}$:

| $\alpha_1$ | $\alpha_2$ | $\ldots$ | $\alpha_n$ |
|---|---|---|---|
| 1 | 0 | $\ldots$ | 0 |
| 0 | 1 | $\ldots$ | 0 |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| 1 | 1 | $\ldots$ | 1 |

.

Всего в классе $2^n - 1$ модель. Таким образом класс моделей $\mathfrak{F}_{\text{lin}} = \{f_{\mathcal{A}} | f_{\mathcal{A}} = \mathbf{w}_{\mathcal{A}}^\intercal \mathbf{x}_A j\}$ задается всеми наборами индексов $\mathcal{A} \subseteq \mathcal{J}$. В векторных обозначениях $\mathbf{f} = X_{\mathcal{A}} \mathbf{w}_{\mathcal{A}}$.

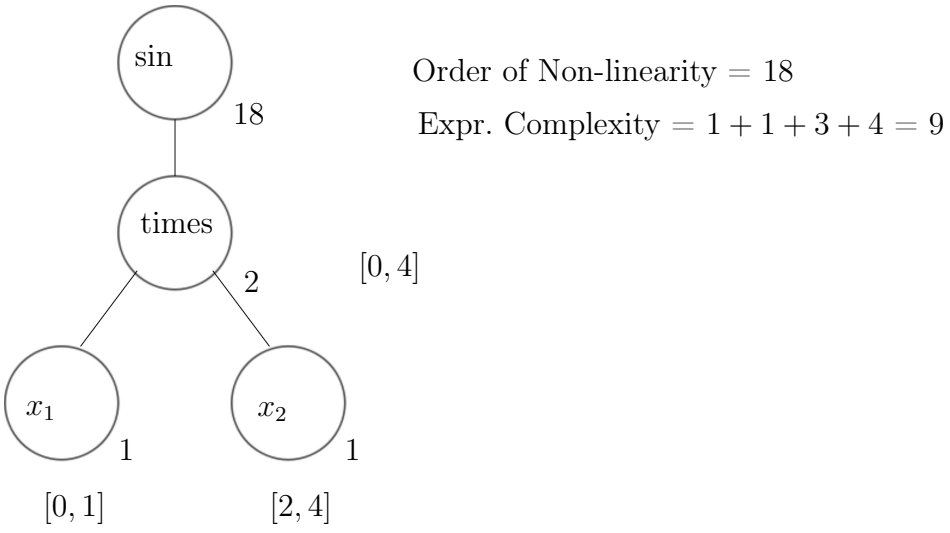



Рис. 16. Вычисление порядка нелинейности для модели, содержащей две свободных переменных.

### 2.1.6. Структурная сложность суперпозиций

Понятние структурной сложности вводится с целью построения процедур индуктивного порождения суперпозиций. Эта сложность не включает никаких сведений о выборке. Например, структурной сложность можно определить как количество вершин в дереве или высота дерева, соответствующего суперпозиции. Ниже в качестве структурной рассматривается сумму вершин всех поддеревьев заданного дерева (см. рис. 18). Такая мера сложности делает более предпочтительными те деревья у которых при том же количестве вершин число ветвей больше. В [157] доказана теорема, упрощающая вычисление структурной сложности: для структурной сложности $\phi$ верна формула $\phi = \eta + s + \pi$, где $\eta$ — длина обхода дерева начиная с самой левой листовой вершины, $\pi$ — длина обхода дерева, полученного из исходного выкидыванием листовых вершин, $s$ — число вершин дерева.

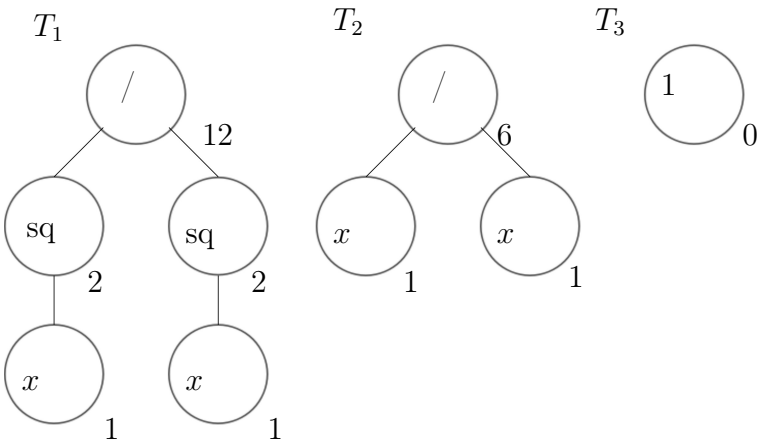



Рис. 17. Индуктивное вычисление порядка нелинейности.

### 2.1.7. Число суперпозиций ограниченной сложности

Оценим число суперпозиций, получаемых после каждой итерации алгоритма индуктивного порождения, описанного в разделе 2.1.3.. Рассмотрим $n$ независимых переменных. Мощность множества $G$ представим как мощности его подмножеств функций соответствующей арности: $|G_1| = l_1, |G_2| = l_2, \ldots, |G_p| = l_p$. На нулевой итерации порождается $P_0 = n$ суперпозиций. На первой итерации порождается

$$P_1 = l_1 n + l_2 n^2 + \cdots + l_n n^p = \sum_{i=1}^{p} l_i P_0^i$$

суперпозиций. Общее число суперпозиций после первой итерации —

$$\hat{P}_1 = P_1 + P_0 = \sum_{i=1}^{p} l_i P_0^i + P_0.$$

Суперпозиции, порожденные на $k$-ой итерации, будут также порождены и на любой следующей после $k$ й. Поэтому обзее число суперпозиций после второй итерации будет равно

$$\hat{P}_2 = \sum_{i=1}^{p} l_i \hat{P}_1^i.$$

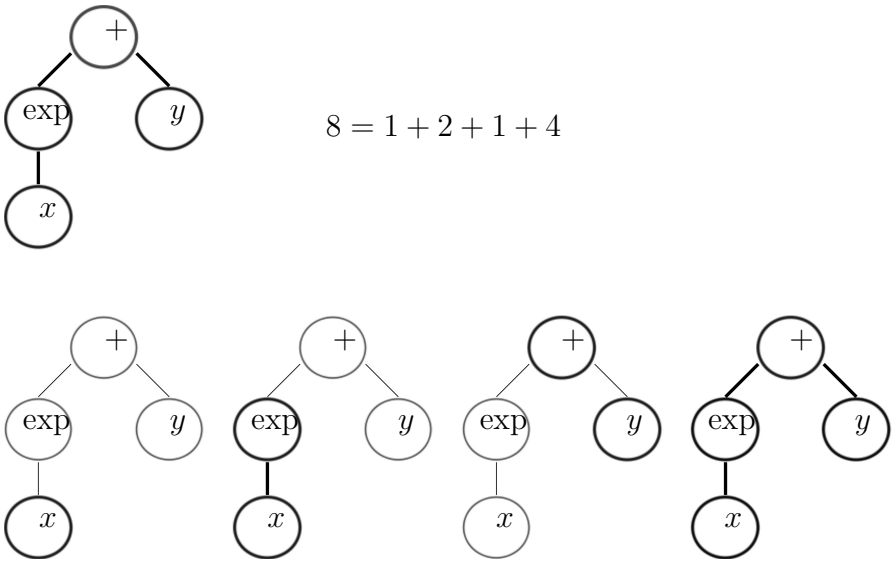



Рис. 18. Вычисление сложности суперпозиции.

После $k$-ой итерации будет порождено

$$\hat{P}_k = \sum_{j=1}^{p} l_i \hat{P}_{k-1}^i$$

суперпозиций. Оценим порядок роста количества функций, порожденных после $k$-ой итерации.

**Теорема 4** (Сологуб)**.** *Пусть в множестве порождающих функций $G$ содержится $l_p$ функций арности $p > 1$ и ни одной функции арности $p+k \mid k > 0$, и имеется $n > 1$ независимых переменных. Тогда справедлива следующая оценка количества суперпозиций, порожденных алгоритмом $\mathfrak{A}$ после $k$-ой итерации:*

$$|\mathfrak{F}_k| = \mathcal{O}(l_p^{\sum_{i=0}^{k-1} p^i} n^{p^k}).$$

*Доказательство.* Оценим сначала порядок роста для случая, когда есть лишь одна $m$-арная функция и $n$ свободных переменных. После первой итерации алгоритма будет порождено $n^m + n$ суперпозиций. После второй — $(n^m + n)^m + n^m + n$, что можно оценить как $(n^m)^m = n^{m^2}$. После $k$-ой итерации количество суперпозиций можно оценить как $n^{m^k}$. Для оценки скорости роста количества порожденных суперпозиций можно учитывать только функции с наибольшей арностью.

Рассмотрим случай, когда имеется не одна функция арности $m$, а $l_m$ таких функций. Тогда на первой итерации порождается $l_m n^m + n$ суперпозиций, на второй —

$$l_m(l_m n^m + n)^m + l_m n^m + n \approx l_m^{m+1} n^{m^2},$$

на третьей, с учетом этого приближения —

$$l_m(l_m^{m+1} n^{m^2})^m = l_m l_m^{m(m+1)} n^{m^3} = l_m^{m^2+m+1} n^{m^3}.$$

Скорость роста количества порожденных суперпозиций оценим как

$$|\mathfrak{F}_k| = \mathcal{O}(l_m^{\sum_{i=0}^{k-1} m^i} n^{m^k}).$$

Таким образом, получаем указанную оценку в случае, когда в множестве $G$ содержится $l_p$ функций арности $p$ и ни одной функции арности $p+k \mid k>0$. □

## 2.2. Порождение суперпозиций

### 2.2.1. Стохастическое порождение суперпозиций

Пусть дана регрессионная выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i) \mid i \in \{1,\ldots,N\}, \mathbf{x}_i \in \mathbb{X} \subset \mathbb{R}^n, y_i \in \mathbb{Y} \subset \mathbb{R}\},$$

где $N$ — объем регресионной выборки (число объектов), $\mathbf{x}_i$ — вектор значений независимых переменных $i$-ого объекта, $y_i$ — значение зависимой переменной у $i$-ого объекта, $\mathbb{X}$ — множество значений независимых переменных, лежащее в $\mathbb{R}^n$, $\mathbb{Y}$ — множество значений зависимой переменной.

Требуется выбрать параметрическую функцию $f : \Omega \times \mathbb{X} \to \mathbb{R}$ из порождаемого множества $\mathfrak{F} = \{f_r\}$, где $\Omega$ — пространство параметров, доставляющую минимум некоторому функционалу ошибки, определяемому ниже.

То есть, для множества всех суперпозиций

$$\mathfrak{F} = \{f_r \mid f_r : (\mathbf{w}, \mathbf{x}) \mapsto y \in \mathbb{Y}, r \in \mathbb{N}\},$$

требуется найти такой индекс $\hat{r}$, что функция $f_r$ среди всех $f \in \mathfrak{F}$ доставляет минимум функционалу качества $S$ при данной регрессионной выборке $\mathfrak{D}$:

$$\hat{r} = \arg\min_{r\in\mathbb{N}} S(f_r \mid \hat{\boldsymbol{\omega}}_{\boldsymbol{r}}, \mathfrak{D}), \tag{73}$$

где $\hat{\boldsymbol{\omega}}_r$ — оптимальный вектор параметров функции $f_r$ для каждой $f \in \mathfrak{F}$ при данной регрессионной выборке $\mathfrak{D}$:

$$\hat{\boldsymbol{\omega}}_{\boldsymbol{r}} = \arg\min_{\mathbf{w}\in\Omega} S(\mathbf{w} \mid f_r, \mathfrak{D}). \tag{74}$$

В качестве функции ошибки $S$ используется сумма квадратов регрессионных остатков:

$$S(\mathbf{w}, f, \mathfrak{D}) = \sum_{i=1}^{N} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2, \text{ при } (\mathbf{x}_i, y_i) \in \mathfrak{D}. \tag{75}$$

Несмотря на то, что построенный ранее итеративный алгоритм $\mathfrak{A}$ порождения суперпозиций позволяет получить за конечное число шагов произвольную суперпозицию, для практических применений он непригоден в связи с чрезмерной вычислительной сложностью, как и любой алгоритм, реализующий полный перебор. Вместо него предлагается использовать стохастические алгоритмы и ряд эвристик, позволяющих на практике получать за приемлемое время результаты, удовлетворяющие заранее заданным условиям. В данном разделе описывается практически реализуемый вариант алгоритма $\mathfrak{A}$, который и использован в вычислительном эксперименте.

Сначала опишем вспомогательный алгоритм случайного порождения суперпозиции:

Алгоритм случайного порождения суперпозиции $\mathcal{RF}$. Заданы

1) набор пороговых значений $0 < \xi_1 < \xi_2 < \xi_3 < 1$.

2) максимальная глубина порождаемой суперпозиции $Td$.

Алгоритм работает следующим образом. Генерируется случайное число $\xi$ на интервале $(0; 1)$, и рассматриваются следующие случаи:

1) $\xi \leq \xi_1$: результатом алгоритма является некоторая случайно выбранная свободная переменная.

2) $\xi_1 < \xi \leq \xi_2$: результатом алгоритма является числовой параметр.

3) $\xi_2 < \xi \leq \xi_3$: результатом алгоритма является некоторая случайно выбранная унарная функция, для определения аргумента которой данный алгоритм рекурсивно запускается еще раз.

4) $\xi_3 < \xi$: результатом алгоритма является некоторая случайно выбранная бинарная функция, аргументы которой порождаются аналогичным образом.

Опишем теперь итеративный алгоритм стохастического порождения суперпозиций. Заданы

1) множество порождающих функций $G$, состоящее только из унарных и бинарных функций;

2) регрессионная выборка $\mathfrak{D}$;

3) параметры $N_{\max}, I_{\max}, \gamma_{\text{mut}}, \gamma_{\text{cross}}$ — максимальное число одновременно рассматриваемых суперпозиций; максимальное число итераций алгоритма; доля суперпозиций, подверженных случайной замене узлов их деревьев; доля суперпозиций, для которых выполняется случайный обмен поддеревьями;

4) прочие параметры, используемые в алгоритме 2.2.1.

Выполнение алгоритма.

1. Инициализируется упорядоченный набор $\mathcal{X}_f$ суперпозиций. А именно, порождается $N_{\max}$ суперпозиций алгоритмом 2.2.1..

2. Оптимизируются параметры $\mathbf{w}$ суперпозиций из $\mathcal{X}_f$ алгоритмом Левенберга-Марквардта.

3. Вычисляется значение $Q_f$ для каждой еще не оцененной суперпозиции $f$ из $\mathcal{X}_f$: для нее рассчитывается значение функции ошибки $S_f$ согласно (75) на выборке $\mathfrak{D}$, и ставится в соответствие значение $Q_f$. Для суперпозиций, при вычислении $Q_f$ которых была хотя бы раз получена ошибка вычислений из-за несовпадения областей определений и значений, принимается $Q_f = -\infty$.

4. Набор суперпозиций $\mathcal{X}_f$ сортируется согласно их приспособленности.

5. Наименее приспособленные суперпозиции удаляются из массива $\mathcal{X}_f$ до тех пор, пока его размер не станет равен $N_{\max}$.

6. Отбирается некоторая часть $\gamma_{\text{mut}}$ наименее приспособленных суперпозиций из $\mathcal{X}_f$. У этой части происходит случайная замена одной функции или свободной переменной на другую: генерируются две случайные величины, одна из которых служит для выбора вершины дерева $\Gamma_f$, которую предстоит изменить, а другая — для выбора нового элемента для этой вершины. Замена такова, чтобы сохранилась структура суперпозиции, а именно: в случае замены функции сохраняется арность, а свободная переменная заменяется только на другую свободную переменную. При этом исходные суперпозиции сохраняются в массиве $\mathcal{X}_f$.

7. Повторяются шаги $3-4$.

8. Производится случайный обмен поддеревьями у $\gamma_{\text{cross}}$ наиболее приспособленных суперпозиций. Вершины, соответствующие этим поддеревьям, выбираются случайным образом. При этом исходные суперпозиции сохраняются в массиве $\mathcal{X}_f$.

9. Повторяются шаги $2-4$.

10. Проверяются условия останова: если либо число итераций больше $I_{\max}$, либо в массиве $\mathcal{X}_f$ есть хотя бы одна суперпозиция с приспособленностью больше, чем $\hat{Q}$, то алгоритм останавливается, и результатом является наиболее приспособленная суперпозиция, иначе осуществляется переход к шагу 2.

Заметим, что выборка $\mathfrak{D}$ не делится на обучающую и контрольную — контроль качества оставляется различным стандартным методикам типа скользящего контроля.

### 2.2.2. Стохастическая процедура порождения модели

**Стохастическая процедура порождения модели.** Процедура из итеративно повторяемых шагов. Из множества моделей-претендентов отбирается заданное число лучших моделей. С помощью операций скрещивания и модификации происходит порождение новых моделей. Процедура повторяется, пока не выполнится условие останова.

Используем переменную выбора признака — вектор $\mathbf{c} = (c_1, \ldots, c_n)$. Алгоритм содержит следующие параметры для отбора моделей: $F$ — число лучших моделей в популяции, $F_1$ — число моделей для скрещивания, $P_2$ — вероятность выбора модели для мутации. Начальный набор моделей выбирается случайным образом. Итеративно выполняются следующие операции.

1. Отбор: согласно критерию (36) при $\mathcal{X} \stackrel{\text{def}}{=} \mathcal{C}$ выбирается $F$ лучших моделей.

2. Случайным образом выбираются $F_1$ моделей для скрещивания и модификации.

3. Скрещивание: операция, при которой из двух моделей порождается две новые. Выбранные модели случайным образом разбиваются на пары. В каждой паре переменные выбора $\mathbf{c}^q = (c_1^q, \ldots, c_n^q)$ и $\mathbf{c}^p = (c_1^p, \ldots, c_n^p)$ разбиваются точкой скрещивания, выбираемой случайно из множества $\{1, \ldots, n\}$, на две части. Происходит обмен элементов векторов $\mathbf{c}^p$ и $\mathbf{c}^p$:

$$\begin{cases} (c_1^q, \ldots, c_k^q, c_{k+1}^q, \ldots, c_n^q) \\ (c_1^p, \ldots, c_k^p, c_{k+1}^p, \ldots, c_n^p) \end{cases} \mapsto$$

$$\mapsto \left\{ \begin{array}{l} (c_1^q, \ldots, c_k^q, c_{k+1}^p, \ldots, c_n^p) \\ (c_1^p, \ldots, c_k^p, c_{k+1}^q, \ldots, c_n^q) \end{array} \right. .$$

4. Каждая модель из полученного множества с вероятностью $P_2$ подвергается модификации: случайным образом из множества $\{1, \ldots, n\}$ выбирается индекс переменной выбора $j$. В результате этой операции значение компоненты $c_j$ меняется на противоположное (если был выбран элемент $c_j = 0$, то после операции он меняет свое значение на 1 и наоборот).

После операций 3 и 4 новые модели настраиваются исходя из условия минимизации критерия (36) при $\mathcal{X} \stackrel{\text{def}}{=} \mathcal{L}$. Вышеперечисленные шаги выполняются заданное число раз.

### 2.2.3. Порождающие функции и классы моделей

В качестве примера множества порождающих функций приведем набор, показанный в таблице 7. Здесь при задании функций приняты следующие обозначения:

$\#\,\mathrm{arg}$ — число аргументов функции,

$\#\mathrm{vec}$ — число элементов вектор-функции,

comm — аргументы коммутируют,

$\#\mathbf{b}$ – число элементов вектора параметров,

$\mathbb{R}_+$ — множество неотрицательных действительных чисел,

$\mathbb{U}$ — неупорядоченное конечное множество, «номинальная» шкала,

$\mathbb{N}^*$ — линейно-упорядоченное конечное множество, «порядковая» шкала,

$\mathbb{B}$ — множество $\{0, 1\}$.

### 2.2.4. Порождаемые модели

Ниже приведены модели, которые используются при регрессионном анализе измеряемых данных. Параметры моделей обозначены латинскими и греческими буквами: $\{a, b, c, ..., \chi, \psi, \omega\}$, $x, y$ — свободная и зависимая переменные. Присоединение параметров-скаляров для их представления в виде вектора $\mathbf{w}$ выполняется в том порядке, в котором они появляются, если представить формулу регрессионной модели в виде строки.

#### Линейные модели

1. Полином $y = \sum_{i=1}^{n} a_i x^{i-1}$ и его частный случай прямая $y = ax + b$. Следует отмечать, что полиномы высоких степеней крайне неустойчивы и могут неадекватно описывать измеряемые данные [291].

2. Гипербола $y = k/x$, а также прочие нелинейные функции с линейно-входящими параметрами: тригонометрические функции $\sin(x), \arcsin(x)$, гиперболический синус $\mathrm{sh}(x)$, корневые $\sqrt{x}$ и обратно-корневые $x^{-\frac{1}{2}}$ функции. Эти функции используются в финансовом анализе и других приложениях.

**Нелинейные модели**

1. Экспонента $y = e^b x$, экспонента с линейным коэффициентом $y = ae^b x$. Распространена двухкомпонентная экспоненциальная модель $y = ae^b x + ce^d x$. Модель может быть использована, в частности, если коэффициент изменения величины свободной переменной пропорционален ее начальной величине.

2. Монотонные модели:

   1) очень быстрый рост $y = \exp(a + bx)$, параметр $b > 0$,

   2) быстрый степенной рост $y = \exp(a_b \ln(x))$, параметр $b > 1$,

   3) медленный рост $y = \exp(a_b \ln(x))$, параметр $0 < b < 1$,

   4) очень медленный рост $y = \ + b \ln(x)$, параметр $b > 0$,

   5) медленная стабилизация $y = \ + b/x$, параметр $b \neq 0$, свободная переменная $x \neq 0$,

   6) быстрая стабилизация $y = a + b \exp(-x)$, параметр $b \neq 0$,

   7) логистическая кривая $y = 1/(a + b \exp(-x))$, параметр $b > 0$.

3. Ряд Фурье $y = a_0 + \sum_{i=1}^{n} \big(a_i \cos(i\omega x) + b_i \sin(i\omega x)\big)$. Используется для описания периодических сигналов.

4. Сумма гауссианов $y = \sum_{i=1}^{n} a_i \exp(-\frac{(x-b_i)^2}{c_i})$. Используется для аппроксимации пиков. Коэффициент $a_i$ является амплитудой, $b_i$ — смещением, коэффициент $c_i$ задает ширину пика. Всего в сумме может быть до $n$ пиков.

5. Моном $y = x^b$, моном с линейным коэффициентом $y = ax^b$. Используется при моделировании размерности физических или химических величин. Например, количество некоторого реагирующего в химической реакции вещества считается пропорциональным концентрации этого вещества, возведенного в некоторую степень.

6. Рациональный полином $y = \frac{\sum_{i=0}^{n} a_i x^i}{x^m + \sum_{i=0}^{m-1} b_i x^i}$. Принято считать коэффициент перед $x^m$ единицей. Например, если $m = n$, такое соглашение позволит получить уникальные числитель и знаменатель.

7. Сумма синусов $y = \sum_{i=1}^{n} a_i \sin(b_i x + c_i)$. Здесь $a_i$ — амплитуда, $b_i$ — частота, $c_i$ — фаза некоторого периодического процесса.

8. Двухпараметрическое распределение Вейбулла $y = abx^{b-1} \exp(-ax^b)$. Параметр $a$ является масштабирующим, а параметр $b$ определяет форму кривой. Трехпараметрическое распределение Вейбулла $y = abx^{b-1} \exp(-a(x-c)^b)$ с параметром смещения $c$.

9. Логистическая функция $(1 + e^{-n})^{-1}$ используются в нейронных сетях, например в MLP в качестве функций активации.

10. Тангенциальная сигмоида $y = 2(1 + e^{-2n})^{-1} - 1$ также используются в качестве функций активации.

Этот список не является исчерпывающим. Выбираемая регрессионная модель зависит прежде всего от экспертных предположений относительно моделируемого явления.

Таблица 7. Набор порождающих функций для задач логистической регрессии.

| Название | dom | # arg | cod | # vec | comm | #**b** |
|---|---|---|---|---|---|---|
| Nominal to binary | $\mathbb{U}$ | 1 | $\mathbb{B}$ | 1–4 | – | – |
| Ordinal to binary | $\mathbb{N}^*$ | 1 | $\mathbb{B}$ | 1–4 | – | – |
| Linear to linear segments | $\mathbb{R}$ | 1 | $[0,1]$ | 1–4 | – | 1–4 |
| Linear segments to binary | $\mathbb{R}$ | 1 | $\mathbb{B}$ | 1–4 | – | 1–4 |
| Get one column of n-matrix | $\mathbb{B}$ | 1–4 | $\mathbb{B}$ | 1 | – | 1 |
| Conjunction | $\mathbb{B}$ | 2–6 | $\mathbb{B}$ | 1 | Да | –n |
| Dijsunction | $\mathbb{B}$ | 2–6 | $\mathbb{B}$ | 1 | Да | – |
| Negate binary | $\mathbb{B}$ | 1 | $\mathbb{B}$ | 1 | – | – |
| Logarithm | $\mathbb{R}_+$ | 1 | $\mathbb{R}$ | 1 | – | – |
| Hyperbolic tangent sigmiod | $\mathbb{R}$ | 1 | $(-1,1)$ | 1 | – | – |
| Logistic sigmoid | $\mathbb{R}$ | 1 | $(0,1)$ | 1 | – | – |
| Sum | $\mathbb{R}$ | 2–3 | $\mathbb{R}$ | 1 | Да | – |
| Divfference | $\mathbb{R}$ | 2 | $\mathbb{R}$ | 1 | Нет | – |
| $\text{Multiplication}_1$ | $\mathbb{R}$ | 2–3 | $\mathbb{R}$ | 1 | Да | – |
| $\text{Multiplication}_2$ | $\mathbb{B}$ | 2–3 | $\mathbb{B}$ | 1 | Да | – |
| Division | $\mathbb{R} \setminus \{0\}$ | 2 | $\mathbb{R}$ | 1 | Нет | – |
| Inverse | $\mathbb{R} \setminus \{0\}$ | 1 | $\mathbb{R}$ | 1 | – | – |
| Polynomial transformation | $\mathbb{R}$ | 1 | $\mathbb{R}$ | 1 | – | $>0$ |
| Radial basis function | $\mathbb{R}$ | 1 | $\mathbb{R}$ | 1 | – | $>0$ |
| Rational $x\sqrt{x}$ | $\mathbb{R}_+$ | 1 | $\mathbb{R}$ | 1 | – | – |

## 2.3. Упрощение суперпозиций

### 2.3.1. Порождение допустимых суперпозиций

Алгоритм, описанный в 2.1.3. позволяет получить все возможные суперпозиции, однако, не все они будут пригодны в практических приложениях: например, выражение $\ln x$ имеет смысл только при $x > 0$, а $\frac{x}{0}$ не имеет смысла вообще никогда. Выражения типа $\frac{x}{\sin x}$ имеют смысл только при $x \neq \pi k$. Используем веденное ранее понятие множества *допустимых* суперпозиций.

Одним из способов построения только допустимых суперпозиций является модификация предложенного алгоритма таким образом, чтобы проверять вложенность области определения и области значения соответствующих функций в ходе построения суперпозиций.

### 2.3.2. Изоморфные суперпозиции

Пусть дана суперпозиция $f$, состоящая из произвольного набора $N$-арных функций, свободных переменных и констант. Пусть также дан набор правил-аксиом, указывающих на существующие соотношения между элементарными функциями. Требуется построить суперпозицию, изоморфную исходной и обладающую наименьшей сложностью, согласно этим правилам.

Здесь под изоморфизмом двух суперпозиций понимается такое эквивалентное преобразование, что обе суперпозиции дают одинаковые результаты при одних и тех же значениях свободных переменных.

Определим понятие сложности суперпозиции $C(f)$:

**Определение 11.** *Сложность суперпозиции $f$, обозначаемая $C(f)$ — число элементарных функций, констант и свободных переменных, каждые из которых считаются столько раз, сколько встречаются в суперпозиции.*

Например, сложность суперпозиции $x + y + y$ равна 5.

Введем также множество $\mathfrak{F}$ всех возможных суперпозиций, составленных из элементарных функций $g \in G$.

Исходная задача формулируется следующим образом. Для данной суперпозиции $f$ требуется найти суперпозицию $\hat{f}$, имеющую минимальную сложность среди всех суперпозиций, изоморфных $f$:

$$\hat{f} = \arg \min_{\varphi \in \mathfrak{F}_f \subset \mathfrak{F}} C(\varphi),$$

где $\mathfrak{F}_f \subset \mathfrak{F}$ — множество всех возможных суперпозиций, изоморфных $f$. Множество $\mathfrak{F}_f$ строится путем последовательного применения заданных правил, описывающих возможные соотношения между элементарными функциями, и являющихся правилами преобразования суперпозиций.

### 2.3.3. Преобразование суперпозиций по правилам

В работе [338] предложен следующий метод преобразования суперпозиций. Условимся считать, что каждой суперпозиции $f$ сопоставлено дерево $\Gamma_f$, строящееся следующим образом:

1) вершине $V_i$ дерева $\Gamma_f$ соответствует элементарная функция $g_s, s = s(i)$,

2) число дочерних вершин $V_j$ у вершины $V_i$ равно арности соответствующей функции $g_s$,

3) порядок вершин, смежных вершине $V_i$, соответствует порядку аргументов соответствующей функции $g_{s(i)}$,

4) листьям дерева $\Gamma_f$ соответствуют свободные переменные $x_i$.

Вычисление значения суперпозиции $f$ в некоторой точке $\mathbf{x} = \{x_i\}$ эквивалентно подстановке соответствующих значений свободных переменных $x_i$ в дерево $\Gamma_f$.

**Определение 12.** *Два дерева $\Gamma_1$ и $\Gamma_2$ равны тогда и только тогда, когда при любой упорядоченной нумерации вершин содержимое вершин с одинаковым номером совпадает.*

**Определение 13.** *Дерево $\Gamma$ является поддеревом дерева $\Gamma'$, если $\Gamma'$ содержит поддерево, равное $\Gamma$.*

Отметим важное свойство таких деревьев: каждое поддерево $\Gamma_f^i$ дерева $\Gamma_f$, соответствующее вершине $V_i$, также соответствует некоторой суперпозиции, являющейся составляющей исходной суперпозиции $f$. Будем обозначать такую суперпозицию, соответствующую вершине $V_i$, как $f_{V_i}$.

**Определение 14.** *Общая компонента дерева — поддерево $\Delta_{f'}$ дерева $\Gamma_f$ суперпозиции $f$, встречающееся более чем один раз.*

Здесь $f'$ — подвыражение в суперпозиции $f$, соответствующее дереву $\Delta_{f'}$.

Так как суперпозиция $f$ и ее дерево $\Gamma_f$ эквивалентны, общей компонентой дерева также можно называть суперпозицию, эквивалентную поддереву, являющемуся общей компонентой в смысле определения 14. Например, для суперпозиции $(x+2)+(x+2)+x$ общие компоненты: $x$, 2, $x+2$.

**Определение 15.** *Наибольшая общая компонента дерева — такая общая компонента $\hat{\Delta}_{f'}$, что не существует другой общей компоненты, включающей данную.*

Для вышеупомянутой суперпозиции $(x+2)+(x+2)+x$ наибольшей общей компонентой является поддерево, соответствующее суперпозиции $x+2$.

Наибольших общих компонент может быть несколько. Например, для суперпозиции $(x+2)+(x+2)+2\sin(x+y)\cos(x+y)$ наибольшими общими компонентами будут подвыражения $x+2$ и $x+y$. Обозначим множество всех наибольших общих компонент графа $\Gamma_f$ суперпозиции $f$ как $\tilde{\Delta}_f$.

Введем понятие унифицированного графа суперпозиции $f$:

**Определение 16.** *Унифицированный граф $\hat{\Gamma}_f$ суперпозиции $f$ — направленный ациклический граф, полученный из дерева $\Gamma_f$ следующим итеративным алгоритмом:*

1. *Граф $\tilde{\Gamma}_f$ на первом шаге равен $\Gamma_f$.*
2. *Для графа $\tilde{\Gamma}_f$ находятся все наибольшие общие компоненты $\hat{\Delta}_{f'}$.*
3. *Если таких компонент не найдено, то полученный граф $\tilde{\Gamma}_f$ объявляется искомым графом $\hat{\Gamma}_f$, и алгоритм завершается.*
4. *Иначе выбирается компонента $\hat{\Delta}_{f'}$ с наибольшей сложностью $C(f')$. Если несколько наибольших общих компонент имеют максимальную сложность, то выбирается первая из них согласно некоторому фиксированному порядку на множестве суперпозиций.*
5. *Выбранная компонента остается в единственном экземпляре: удаляются все вершины $V_i$ такие, что $f_{V_i} = f'$, кроме одной вершины $V_i'$, и ребра, входящие в удаленные вершины, изменяются таким образом, чтобы они входили в $V_i'$.*

То есть, в графе $\hat{\Gamma}_f$ каждая наибольшая общая компонента $\hat{\Delta}_{f'}$ присутствует не более одного раза.

Таким образом, $\hat{\Gamma}_f$ отличается от $\Gamma_f$ тем, что в одну вершину могут входить несколько ребер, причем в том и только в том случае, если эта вершина соответствует поддереву, являющемуся некоторой наибольшей общей компонентой дерева $\Gamma_f$.

**Теорема 5.** *Граф $\hat{\Gamma}_f$ по данному дереву $\Gamma_f$ строится единственным образом.*

*Доказательство.* Утверждение напрямую следует из метода построения $\hat{\Gamma}_f$, если учесть, что не важно, какая именно вершина не была удалена на шаге 5, так как после указанного преобразования графа все такие вершины равнозначны. □

**Теорема 6.** *Дерево $\Gamma_f$ по данному графу $\hat{\Gamma}_f$ восстанавливается единственным образом.*

*Доказательство.* Для того, чтобы построить дерево $\Gamma_f$ по данному графу $\hat{\Gamma}_f$, достаточно найти в $\hat{\Gamma}_f$ все вершины $V_i$, имеющие более одной родительской вершины $V_j$. Каждой такой вершине $V_i$ соответствует некоторая суперпозиция $f_{V_i}$ и ее подграф $\Gamma^i_f$. Добавим в граф $\Gamma_f$ такое минимальное число подграфов, равных $\Gamma^i_f$, чтобы каждая вершина из $V_j$ указывала в свой собственный подграф, равный $\Gamma^i_f$, то есть, чтобы множества вершин дочерних подграфов не пересекались. Так как $\hat{\Gamma}_f$ построен путем выделения и объединения общих компонент дерева $\Gamma_f$, то полученный в результате указанной процедуры граф будет являться деревом, причем, эквивалентным $\Gamma_f$. □

Из этих двух утверждений непосредственно следует следующая теорема:

**Теорема 7.** *Между $\hat{\Gamma}_f$ и $\Gamma_f$ существует взаимно однозначное преобразование.*

Эта теорема позволяет говорить об эквивалентности $\Gamma_f$ и $\hat{\Gamma}_f$ и о том, что взаимно однозначные преобразования для $\hat{\Gamma}_f$ являются также и взаимно однозначными преобразованиями для суперпозиции $f$.

Работа алгоритма преобразования суперпозиций по правилам состоит из двух этапов.

1. Находятся наибольшие равные в смысле определения 12 поддеревья, и дерево $\Gamma_f$ суперпозиции $f$ преобразуется в соответствующий унифицированный граф $\hat{\Gamma}_f$.

2. К полученному графу $\hat{\Gamma}_f$ применяются правила преобразования, уменьшающие его сложность, до тех пор, пока правила возможно применять.

Заметим, что, например, правило, описывающее вынесение общего множителя за скобки, заменяющее $ax+bx$ на $(a+b)x$, не применимо к суперпозиции типа $nx+x$ в описанном выше виде. Шаблон Rst у такого правила представляет умножение константы на переменную, в то время как в суперпозиции $nx+x$ второй аргумент — константа.

Чтобы избежать подобной ситуации, предлагается указывать каждое подобное соотношение не в виде правил переписывания графа, а в виде отношений эквивалентности, что позволяет сократить количество экспертно заданных правил.

## 2.4. Структурное обучение при порождении суперпозиций

Предложен метод прогнозирования структуры суперпозиции регрессионной модели, описывающей предъявленную выборку оптимальным образом. Алгоритмы выбора моделей имеют значительную вычислительную сложность в связи с необходимостью перебора большого числа моделей. Основываясь на собранных прецедентах выбора моделей, адекватно описывающих выборки, предлагается построить алгоритм прогноза структуры таких моделей.

### 2.4.1. Постановка задачи структурного обучения

Задан набор $\mathfrak{D} = \{(\mathbf{D}_k, f_k)\}$, состоящий из регрессионных выборок $\mathbf{D}$. Каждая пара $\mathbf{D}_k = (\mathbf{X}_k, \mathbf{y}_k)$ состоит из $(m \times n)$-матрицы $\mathbf{X}$ и $(m \times 1)$-вектора $\mathbf{y}$. Для каждой регрессионной выборки $\mathbf{D}_k$ известна модель $f_k$, оптимально приближающая данную выборку. Задано множество $\mathcal{G}$ порождающих функций. Для каждой функции $g : \mathbb{R} \times \ldots \times \mathbb{R} \to \mathbb{R}$ из набора $\mathcal{G}$ определены её арность $v = v(g)$, области определения и значений: $\mathrm{dom}(g), \mathrm{cod}(g)$. Известно множество $\mathfrak{F}$ суперпозиций порождающих функций, при этом заданы правила индуктивного порождения функции $f \in \mathfrak{F}$:

$$\mathfrak{F} = \{f_s \mid \mathbf{f}_s : (\hat{\mathbf{w}}_k, \mathbf{X}) \mapsto \mathbf{y}, s \in \mathbb{N}\}.$$

Каждой выборке $\mathbf{D}$ требуется поставить в соответствие оптимальную модель $f : \mathbb{W} \times \mathbb{X} \to \mathbb{R}$ из порождаемого множества моделей $\mathfrak{F} = \{f_s\}$, где $\mathbb{W}$ — пространство параметров, доставляющую минимум заданной функции ошибки, определяемой ниже.

Другими словами, для множества моделей $\mathfrak{F}$ требуется найти такой индекс $\hat{s}$, что функция $f_{\hat{s}}$ среди всех $f \in \mathfrak{F}$ доставляет минимум функции ошибки $S$ при фиксированной регрессионной выборке $\mathbf{D}$:

$$\hat{s} = \arg\min_{s \in \mathbb{N}} S(f_s \mid \hat{\mathbf{w}}_k, \mathbf{D}_k), \tag{76}$$

где $\hat{\mathbf{w}}_k$ — оптимальный вектор параметров модели $f_s$ для каждой $f \in \mathfrak{F}$ при данной регрессионной выборке $\mathbf{D}$:

$$\hat{\mathbf{w}}_k = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\hat{\mathbf{w}} \mid f_s, \mathbf{D}_k). \tag{77}$$

В качестве функции ошибки $S$ используется сумма квадратов регрессионных остатков 75.

### 2.4.2. Способ задания структуры регрессионной модели

Каждой суперпозиции $f$ ставится в соответствие дерево $\Gamma_f$ вида (рис.19), строящееся по следующим правилам:

1) корнем дерева является специальный символ “*”, имеющий одну дочернюю вершину,

2) в остальных вершинах $V_i$ дерева $\Gamma_f$ находятся соответствующие порождающие функции $g_{r(i)}$ из набора $\mathcal{G}$,

3) число дочерних вершин $V_j$ у некоторой вершины $V_i$ равно арности соответствующей функции $g_r$: $v = v(g_r)$,

4) область определения порождающей функции дочерней вершины $V_j$ содержит область значений функции родительской вершины $V_i$: $\mathrm{dom}(g_{r(i)}) \supset \mathrm{cod}(g_{r(j)})$,

5) в листьях дерева $\Gamma_f$ находятся свободные переменные $x_i$.

Каждому дереву $\Gamma_f$ ставится в соответствие бинарная матрица $\mathbf{Z}$ (табл.8) размера $(1 + l) \times (l + n)$, где $l$ – число элементарных функций набора $\mathcal{G}$, $n$ — число свободных переменных $\mathbf{x}_i$. Элементы матрицы $\mathbf{Z}$ отвечают за наличие ребра между двумя вершинами в

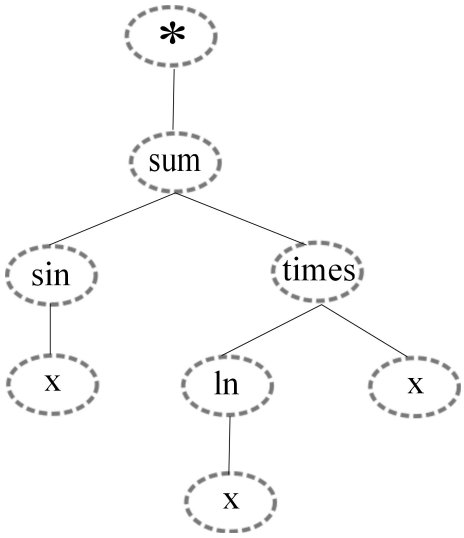



Рис. 19. Пример дерева суперпозиции функций.

| | sum | times | ln | sin | $x$ |
|---|---|---|---|---|---|
| $*$ | 1 | 0 | 0 | 0 | 0 |
| sum | 0 | 1 | 1 | 0 | 0 |
| times | 0 | 0 | 0 | 1 | 1 |
| ln | 0 | 0 | 0 | 0 | 1 |
| sin | 0 | 0 | 0 | 0 | 1 |

| | sum | times | ln | sin | $x$ |
|---|---|---|---|---|---|
| $*$ | 0.7 | 0.1 | 0.1 | 0.1 | 0.2 |
| sum | 0.2 | 0.7 | 0.8 | 0.1 | 0.2 |
| times | 0.1 | 0.3 | 0 | 0.8 | 0.8 |
| ln | 0.2 | 0.1 | 0.3 | 0.1 | 0.9 |
| sin | 0.1 | 0.2 | 0.1 | 0 | 0.8 |

Таблица 8. Пример матрицы связей и матрицы вероятностей связей для суперпозиции функций.

дереве. При этом строки матрицы отвечают только за те вершины, которые могут быть родительскими: вершина дерева “ * ” и порождающие функции $g_s$. Столбцы матрицы отвечают за потенциальные дочерние вершины: порождающие функции $g_r$ и свободные переменные $x_i$. Таким образом, матрица $\mathbf{Z}$ состоит из квадратного блока, отвечающего за набор порождающих функций, добавленной сверху строки, отвечающей за вершину дерева “ * ”, и $n$ добавленных справа столбцов, отвечающих за свободные переменные $x_i$. На матрицу $\mathbf{Z}$ по построению накладываются следующие ограничения:

1) в каждой строке $i$ содержится либо количество единиц, равное арности $v = v(g_{r(i)})$ элементарной функции $g_{r(i)}$, отвечающей за $i$-ый столбец матрицы, либо ноль;

2) в каждом столбце, отвечающем за порождающую функцию, может содержаться только одна единица;

3) заполнение строк и столбцов проходит сверху–вниз и слева–направо, т.е. для записи очередного ребра в матрицу выбирается самый левый и верхний из “свободных” столбцов и строк, отвечающий тем же родительским и дочерним элементам.

Обозначим для удобства множество матриц, удовлетворяющих данным условиям как $\mathcal{M}$.

### 2.4.3. Оценка вероятности переходов в дереве суперпозиции

Поскольку по матрице из множества $\mathcal{M}$ можно однозначно восстановить суперпозицию функции, задача прогнозирования суперпозиции $f$ сводится к поиску матрицы $\mathbf{Z}_f$ из множества $\mathcal{M}$, максимизирующей вероятность переходов в дереве суперпозиции:

$$\mathbf{Z}_f = \arg\max_{\mathbf{Z}\in\mathcal{M}} \sum_{i,j} P_{ij} \times Z_{ij}, \tag{78}$$

где матрица вероятностей переходов $\mathbf{P}$ определяется с помощью векторной логистической регрессии с функцией ошибки, соответствующей гипотезе порождения данных биномиальным распределением.

### 2.4.4. Решение задачи структурного обучения

Пусть с помощью векторной логистической регрессии найдена матрица вероятностей переходов $\mathbf{P}_f$ вида табл. 8. Ставится задача отыскания матрицы $\mathbf{Z}_f$ из допустимого множества матриц $\mathcal{M}$, удовлетворяющей условию 78. Для этого разобьем матрицу $\mathbf{P}_f$ на два блока. Блок $P'_{(1+l)\times l}$:

$$P'_{ij} = p(g_i \to g_j)$$

отвечает за вероятности переходов между порождающими функциями. Блок $P''_{(1+l)\times n}$:

$$P''_{ik} = p(g_i \to x_k)$$

содержит значения вероятностей перехода от порождающих функций к независимым переменным. Введем понятия открытой вершины. Назовем вершину $i$ — *открытой*, если она

относится к порождающей функции, и при этом существует вершина, являющаяся для вершины $i$ родительской, но у нее нет дочерних вершин:

$$(i \le l)\&(\exists j : (j,i) = 1)\&(\nexists k : (i,k) = 1).$$

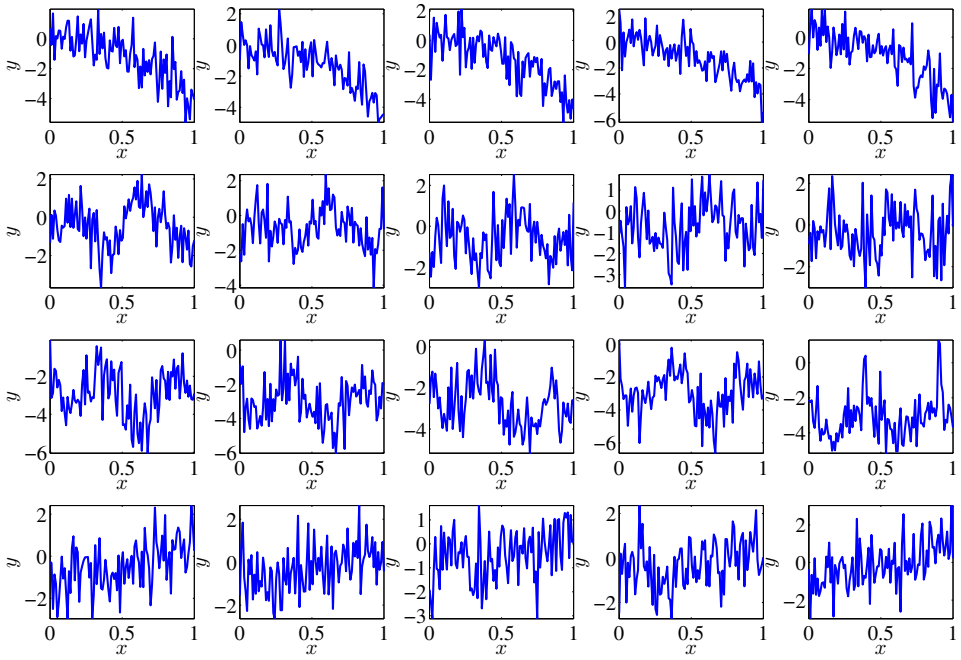


Рис. 20. Исходные временные ряды, порожденные различными моделями.

Также зададим значение $K$ максимально допустимой сложности суперпозиции. Опишем процедуру построения оптимального дерева $\hat{\Gamma}_f$.

1. На нулевом шаге процедуры объявляем вершину дерева открытой: $i = 1$.

2. Пока количество единиц в матрице не превышает $K$, повторяем:

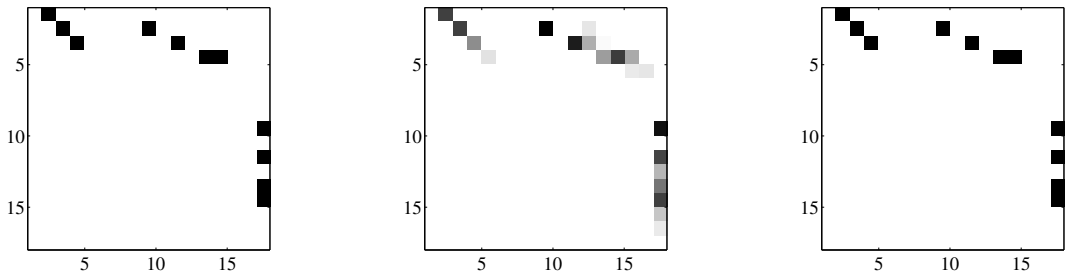
   1) выбираем максимальные вероятности переходов $c_j = \max\limits_{j=1,\dots,l} P_{ij}$ для всех открытых вершин $i$;

   2) достраиваем матрицу из условия максимизации вероятности перехода: $j^* = \arg\max\limits_j c_j, \quad (i,j^*) = 1$;

   3) добавляем $j^*$ к списку открытых вершин, если $(i,j^*) \in P'$;

3. если количество единиц превышает $K$, ставим в соответствие всем открытым вершинам независимые переменные: $k^* = \arg\max\limits_k P''_{ik}, \quad (i,k^*) = 1$ для всех $i$-открытых.

Процедура может быть прервана, если множество открытых вершин пусто, но сложность суперпозиции еще не превысила заданную максимальную сложность $K$ — в таком случае построенная оптимальная суперпозиция будет иметь меньшую сложность.




(a) Исходная матрица переходов, $\mathbf{Z}_f$ (b) Полученная матрица вероятности переходов, $\mathbf{P}_f$ (c) Полученная матрица переходов, $\hat{\mathbf{Z}}_f$

Рис. 21. Матрицы переходов в графе суперпозиции.

### 2.4.5. Процедура прогнозирования структуры модели

Алгоритм протестирован на выборке синтетических данных, полученных следующим образом. Экспертно задан набор порождающих функций $\mathcal{G}$, для каждой из которых известны арность функции $v = v(g)$, области определения и значений: $\mathrm{dom}(g), \mathrm{cod}(g)$. По набору $\mathcal{G}$ построено конечное множество суперпозиций $\mathfrak{F}$ — библиотека функций. Экспертно заданы значения независимых переменных $\mathbf{X}$ и вектор параметров модели $\mathbf{w}_s$. Значения зависимых переменных заданы как

$$\mathbf{y}_s = f_s(\mathbf{w}_s, \mathbf{X}) + \tau_f,$$

где $\tau_f$ — шумовая добавка, являющаяся случайной величиной из нормального распределения. На рис.20 изображен вид исходных моделей. В каждой строке $i$ содержатся графики модели $f_i$, зашумленной независимо друг от друга 5 раз одним и тем же распределением. Таким образом, сгенерировано множество регрессионных выборок $\mathfrak{D} = \{(\mathbf{D}_s, f_s)\}$, где $\mathbf{D}_s = (\mathbf{X}, \mathbf{y})$.

Для обучения алгоритма векторной логистической регрессии использована нейронная сеть с двумя скрытыми слоями, имеющая на выходном слое сигмоидную функцию активации. На вход такой нейросети подается регрессионная выборка $\mathbf{D}_s = (\mathbf{X}, \mathbf{y})$, выходом алгоритма является матрица вероятностей $\mathbf{P}$. Далее при помощи указанной выше процедуры построения оптимального дерева $\hat{\Gamma}_f$, прогнозировалась искомая структура суперпозиции модели. Пример работы алгоритма изображен на рис.(21a, 21b, 21c). Левая матрица соответствует исходной суперпозиции $f_s$, средняя матрица — построенной матрице вероятностей переходов $\mathbf{P}_f$, по которой, используя предложенную процедуру, вычисляется оптимальная прогнозируемая суперпозиция $\hat{f}_s$.

Для тестирования качества алгоритма использован метод LOO(Leave-One-Out), по которому множество регрессионных выборок разбивается таким образом, что в обучении алгоритма использованы все выборки, за исключением одной: $\mathfrak{D} \setminus \{\mathbf{D}_k\}$. Контроль проведен на одной выборке $\mathbf{D}_k$, для которой по полученному алгоритму вычислено оптимальное дерево суперпозиции $\hat{\Gamma}_k$, построена модель $\hat{f}_k$, вычислены ее оптимальные параметры $\hat{\mathbf{w}}_k$ и вычислено значение ошибки

$$S(\hat{\mathbf{w}}_k, \hat{f}_s, f_s) = \| \mathbf{y} - f(\mathbf{w}_k, \mathbf{X}) \|_2.$$

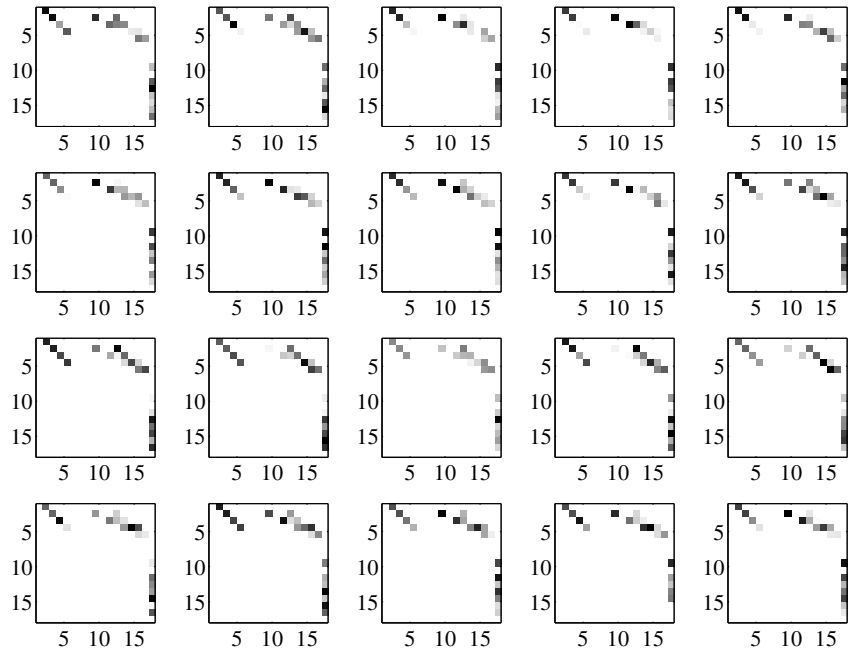



Рис. 22. Полученные матрицы вероятностей переходов в графе суперпозиции.

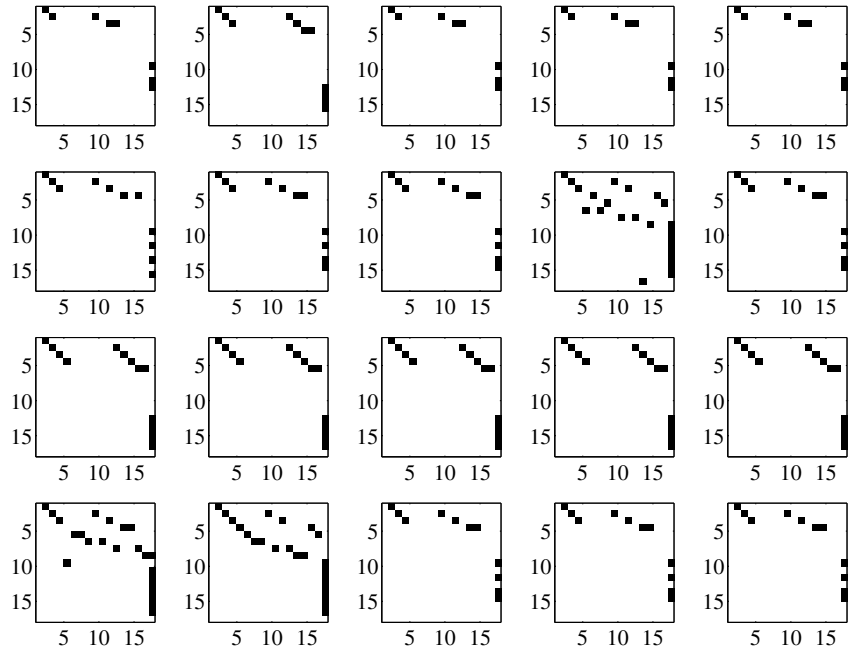
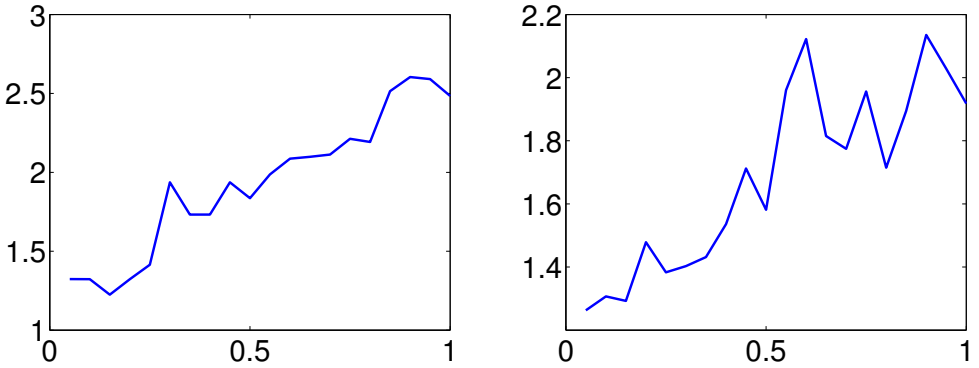



Рис. 23. Полученные матрицы переходов в графе суперпозиции.

Результаты прогнозирования по методу LOO представлены на рис.22, отражающем полученные вероятности переходов $\mathbf{P}$, и рис.23, отражающем оптимальные прогнозируемые матрицы переходов $\hat{\mathbf{Z}}$.




(a) Зависимость ошибки $S$ от дисперсии шума $\sigma$

(b) Зависимость ошибки $S$ от возмущения параметров модели $\delta\mathbf{w}$

Рис. 24. Зависимость ошибки от возмущения шума и параметров модели.

Из рис.23 видно, что не во всех строках получились одинаковые матрицы переходов, что отражает некую неустойчивость предложенного метода относительно вводимого шума $\tau$. Качество работы алгоритма в зависимости от размера $\sigma$ дисперсии нормальной случайной величины $\tau$ изображено на рис.24a. Видно, что при увеличении шума качество прогнозирования падает, растет ошибка $S$. Немонотонность графика объясняется малыми для обучения размерами выборки. На рис.24b изображен график зависимости ошибки прогнозирования от возмущения $\delta\mathbf{w}$ вектора параметров $\mathbf{w}_k$ модели $f_k$ контрольной выборки

$$\mathbf{D}_k = (\mathbf{X}_k, \mathbf{y}_k), \mathbf{y}_k = f(\mathbf{w}_k, \mathbf{X}_k) + \tau_f.$$

В данном случае также можно отметить увеличение ошибки $S$ от размера возмущения $\delta\mathbf{w}$, но при этом качество прогнозирования остается достаточно хорошим при небольших значениях возмущения.

Таким образом поставлена решена задача прогнозирования суперпозиции модели исходных данных. Предложен алгоритм поиска оптимальной структуры модели. Качество предлагаемого метода проверено на выборке синтетических данных.

# 3. Сравнение элементов моделей

Процедура выбора моделей может быть организована двумя способами: прямое сравнение сложности моделей из некоторого ранее порожденного множества и сравнение сложности последовательно порождаемых моделей. Первый способ предполагает, что сложность каждой модели уже известна [256]. При этом процедура прямого сравнения требует большего количества шагов, так как необходимо оценить сложность для каждой модели из множества допустимых моделей. Второй способ — последовательное сравнение моделей — позволяет избежать вычисления оценки для всех моделей. При этом модель должна быть представима в виде композиции отдельных ее элементов. Выбор модели при этом выполняется следующим образом. Вводится критерий сравнения элементов, после чего по результатам сравнения один или несколько элементов добавляются в модель или удаляются из модели. Предлагается алгоритм выбора признаков или элементов, оптимизирующий структуру модели.

Этот алгоритм различается для обобщенно-линейных и нелинейных моделей. В первом случае речь идет о *выборе признаков*, во втором — о *выборе элементов суперпозиции*. Мотивацией работы является тот факт, что решение практических задач восстановления регрессионной зависимости требует рассмотрения большого числа порождаемых признаков или элементов. Требуется предложить такой алгоритм выбора признаков, который за «небольшое» число шагов выбрал бы набор признаков, задающий модель оптимальной сложности.

Процедура построения регрессионных моделей состоит из двух шагов. На первом шаге на основе свободных переменных, результатов измерений, порождается набор признаков. На втором шаге выбираются признаки. При выборе признаков выполняется оценка параметров модели и вычисляется ее сложность.

Развитие методов выбора признаков в регрессионном анализе имеет насыщенную историю. Широкое распространение получил шаговый метод, впервые предложенный в 1960 г. М. А. Эфроимсоном [83]. Он состоит из процедур поочередного добавления и удаления признаков. На каждом шаге признаки проверяются на возможность добавления признака в модель или возможность удаления из модели. Выбираются признаки, которые вносят наибольший вклад в зависимую переменную. Выбор выполняется процедурой, состоящей из серии $F$-тестов. Для выбора оптимального набора признаков используется критерий Акаике, Байесовский критерий [15, 135] или критерий Маллоуза [197].

В 1963 г. А. И. Тихонов ввел понятие регуляризации — дополнительного ограничения на задачу [310]. В работах [348, 349] введено понятие класса регуляризуемых некорректно поставленных задач и предложен общий метод решения таких задач, названный методом регуляризации. Работы А. И. Тихонова были опубликованы на западе только в 1977 г. В 1970 г. А. Хоэрл и Р. Кеннард предложили метод гребневой регрессии, в котором использовалась регуляризация [137]. Было введено дополнительное регуляризующее слагаемое в минимизируемую функцию, что дало улучшение устойчивости решения [50, 360].

В первом издании книги Н. Дрейпера 1966 г. [80] приведен ступенчатый алгоритм выбора признаков. На каждой итерации выбирается признак, имеющий наибольшую проекцию на вектор зависимых переменных, после чего делается небольшой шаг в направлении

решения [130]. Среди полученных на каждом шаге моделей находится оптимальная, то есть выбираются признаки.

В 1971 г. А. Г. Ивахненко начал разрабатывать семейство методов группового учета аргументов [344]. Согласно его подходу, на каждом шаге происходит выбор моделей и построение на их основе более сложных моделей [192, 308, 307]. Метод позволяет сократить перебор и выбрать признаки.

В работе Д. Холланда 1975 г. [139] рассматривается общий подход к построению адаптивных систем. Любая адаптивная задача может быть описана в терминах теории эволюции. Задача, сформулированная таким образом, может быть решена с помощью генетического алгоритма. В основе подхода лежит теорема схем, из которой следует, что при определенных условиях алгоритм дает экспоненциально быструю сходимость решения к локальному оптимуму.

В работах С. Шена 1980 г. и 1991 г. [66, 65] рассмотрен алгоритм последовательного добавления признаков с ортогонализацией. Отбор признаков происходит автоматически при выборе оптимальной модели [38, 122].

Для упрощения структуры модели также используется метод оптимального прореживания, согласно которому элементы модели, оказывающие малое влияние на ошибку аппроксимации, можно исключить из модели без значительного ухудшения качества аппроксимации [343, 353, 176]. Метод предложен в 1990 г. Я. ЛеКюном и развит Б. Хассиби. Он основан на анализе первых производных в ходе обучения градиентными методами [114, 270, 129, 323, 114].

Еще один метод регуляризации, Лассо, был предложен Р. Тибширани в 1996 г. [267]. В нем вводится ограничение на $L_1$-норму вектора параметров модели, что приводит к обнулению части параметров модели и улучшению устойчивости решения.

В 2002 г. Б. Эфрон, Т. Хасти, И. Джонстон и Р. Тибширани предложили метод наименьших углов (Least Angle Regression) [82]. Алгоритм заключается в последовательном добавлении признаков. На каждом шаге признак выбирается таким образом, чтобы вектор регрессионных остатков был равноуголен уже добавленным в модель признакам [173]. Обобщение метода на многоиндексную матрицу плана можно найти в [212].

Перечислим наиболее часто используемые методы выбора признаков:

1) полный перебор моделей [308];

2) генетический алгоритм [110];

3) метод группового учета аргументов [192, 308, 359];

4) шаговая регрессия [80, 83, 233];

5) гребневая регрессия [80];

6) алгоритм Лассо [267];

7) ступенчатая регрессия [80];

8) последовательное добавление признаков с ортогонализацией [66, 65];

9) метод наименьших углов [82];

10) оптимальное прореживание в линейной регрессии [343, 353].

## 3.1. Методы эмпирического выбора признаков

### 3.1.1. Регуляризующие методы

**Гребневая регрессия.** Метод заключается во введении дополнительного регуляризующего слагаемого в минимизируемую функцию ошибок. Этот метод не является методом выбора признаков, так как не указывает, какие признаки следует исключить из модели. Плохая обусловленность матрицы $\mathbf{X}^\intercal\mathbf{X}$ приводит к неустойчивости решения $\hat{\mathbf{w}}$ уравнения линейной регрессии (38). Регуляризация позволяет уменьшить число обусловленности матрицы $\mathbf{X}^\intercal\mathbf{X}$ и получить более устойчивое решение.

При регуляризации параметры модели находятся из условия минимизации функции

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}), \tag{79}$$

$$S(\mathbf{w}) = \big(\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \tau\|\mathbf{w}\|^2\big).$$

Решением задачи минимизации является вектор $\hat{\mathbf{w}} = (\mathbf{X}^\intercal\mathbf{X} + \tau\mathbf{I}_n)^{-1}\mathbf{X}^\intercal\mathbf{y}$.

Увеличение параметра $\tau$ приводит к уменьшению нормы вектора параметров модели и повышению эффективной размерности пространства признаков [131]. Действительно, рассмотрим сингулярное разложение $\mathbf{X}^\intercal\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$. Числом обусловленности матрицы называется отношение максимального сингулярного числа к минимальному:

$$\varkappa = \frac{\lambda_1}{\lambda_n}.$$

Рассмотрим число обусловленности $\varkappa$ регуляризованной матрицы $\mathbf{X}^\intercal\mathbf{X}$ нормального уравнения:

$$\varkappa\big(\mathbf{X}^\intercal\mathbf{X} + \tau\mathbf{I}\big) = \frac{\lambda_1 + \tau}{\lambda_n + \tau}$$

где $\lambda_i$ — сингулярные числа матрицы $\mathbf{X}^\intercal\mathbf{X}$. Чем больше $\tau$, тем устойчивее решение задачи. С увеличением коэффициента регуляризации $\tau$ уменьшается число обусловленности матрицы $\mathbf{X}^\intercal\mathbf{X}$. Возможен другой способ регуляризации

$$\mathbf{X}^\intercal\mathbf{X} \mapsto (1-\tau)\mathbf{X}^\intercal\mathbf{X} + \tau\mathrm{diag}(\mathbf{X}^\intercal\mathbf{X})\mathbf{I}_n.$$

Функция ошибки $S(\mathbf{w})$ — квадратична относительно параметров $\mathbf{w}$, поэтому поверхность $S = \mathrm{const}$ является эллипсоидом. Как видно из уравнения (79), коэффициент регуляризации, отличный от нуля, задает радиус сферы в этом пространстве. Точка касания эллипсоида и сферы является решением уравнения (79) при фиксированном $\tau$. При этом касание эллипсоида в нулевой точке исключено. То есть, обнуление параметров $\mathbf{w}$ не происходит. Метод гребневой регресии улучшает устойчивость параметров регрессионной модели, но не приводит к обращению в ноль ни одного из них.

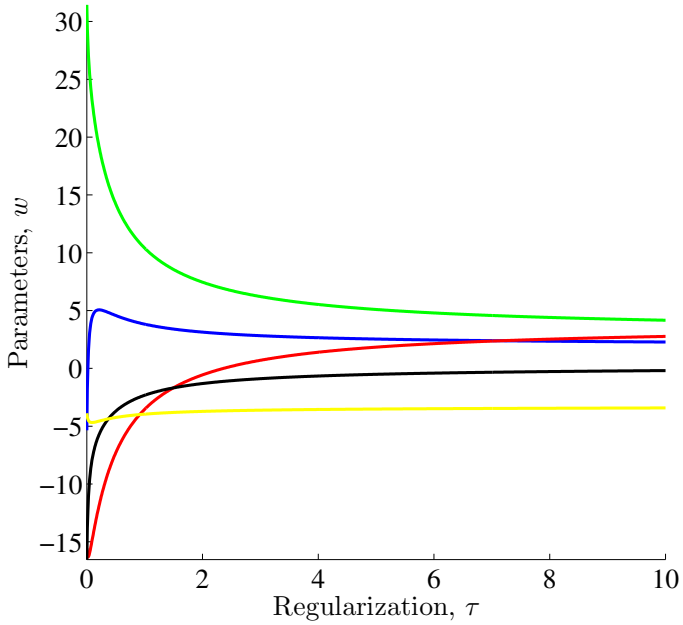



Рис. 25. Зависимость значений параметров при гребневой регрессии от коэффициента регуляризации $\tau$.

**Метод Лассо.** Метод нахождения оценки параметров линейной модели при ограничении на сумму их абсолютных значений. В отличие от гребневой регрессии, в Лассо некоторые параметры становятся равными нулю, а значит, выполняется отбор признаков. Рассматривается сумма модулей параметров модели, $T(\mathbf{w}) = \|\mathbf{w}\|_1$.

Регрессионные параметры выбираются из условия минимизации функции ошибки $S(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - y\|_2^2$ при ограничении

$$T(\mathbf{w}) \leq \tau, \tag{80}$$

где $\tau$ — параметр регуляризации. Для решения используется метод квадратичного программирования с ограничением в виде линейного неравенства. При больших $\tau$ решение, получаемое методом квадратичного программирования, совпадает с решением, полученным методом наименьших квадратов. Чем меньше $\tau$, тем большее число параметров $w_j, j \in \mathcal{J}$ принимает нулевое значение, см. рис.26. Соответствующие признаки исключаются из регрессионной модели.

Задача может решаться методом наименьших квадратов (38) с $2^n$ ограничениями-неравенствами соответствующими $2^n$ возможным наборам знаков параметров. Элемент $\text{sign}(w_j)$ набора параметров принадлежит множеству $\{-1, +1\}$. Найдем решение при фиксированном $\tau \geq 0$. Введем $\boldsymbol{\delta}_i$, $i \in \{1, 2, \ldots, 2^n\}$ — $n$-мерные векторы вида $[\pm 1, \pm 1, \ldots, \pm 1]^\intercal$. Тогда условия (80) эквивалентны системе линейных неравенств

$$\boldsymbol{\delta}_i^\intercal \mathbf{w} \leq \tau \quad \text{для } i = 1, \ldots, 2^n.$$

Для заданного вектора $\mathbf{w}$ пусть

$$E = \{i : \boldsymbol{\delta}_i^\intercal \mathbf{w} \leqslant \tau\}.$$

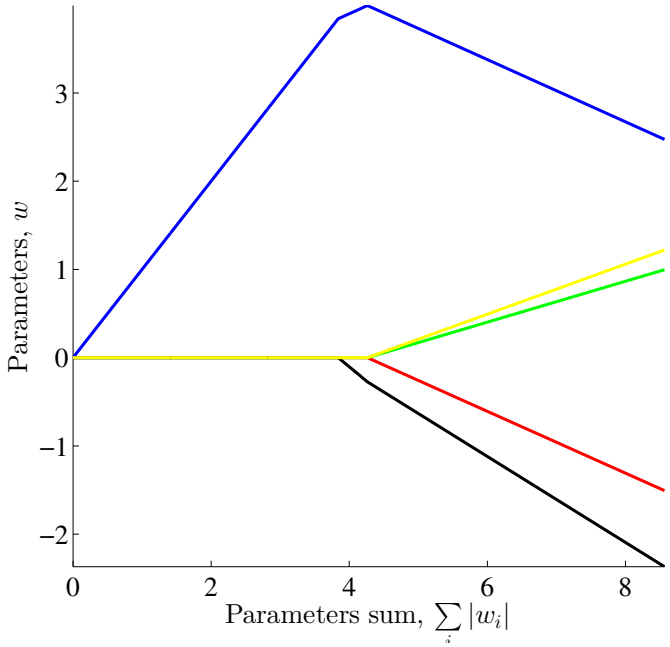



Рис. 26. Оценки параметров, полученные с помощью метода Лассо, в зависимости от нормы $\|\mathbf{w}\|_1$.

Введем матрицу $G_E$, строками которой являются $\boldsymbol{\delta}_i$, $i \in E$, и $\mathbf{1}$ — вектор-строка из единиц длиной, равной числу столбцов в $G_E$. Начальное приближение для алгоритма: $E = \{i_0\}$, где $\boldsymbol{\delta}_{i_0} = \text{sign}(\hat{\mathbf{w}})$, $\hat{\mathbf{w}}$ — оценка вектора параметров методом наименьших квадратов без ограничений в виде системы линейных неравенств. Пока $\|\hat{\mathbf{w}}\|_1 > \tau$

1) найти $\hat{\mathbf{w}}$, минимизирующий функцию ошибки $S(\mathbf{w})$ при $G_E\mathbf{w} \leq \mathbf{1}\tau$,

2) добавить такое $i$ в набор $E$, что $\boldsymbol{\delta}_i = \text{sign}(\hat{\mathbf{w}})$.

Эта процедура сходится за конечное число шагов, так как на каждом шаге добавляется по одному вектору $\boldsymbol{\delta}_i$ и число добавляемых векторов конечно. Вектор $\hat{\mathbf{w}}$ получаемый на последнем шаге, является решением задачи.

Рассмотрим альтернативный метод решения. Каждый параметр $w_j$ задачи оптимизации записывается в виде $wj = w_j^+ - w_j^-$, где $w_j^+$ и $w_j^-$ неотрицательны,

$$w_j^+ = \begin{cases} w_j, w_j \geq 0, \\ 0, w_j < 0, \end{cases} \qquad w_j^- = \begin{cases} 0, w_j > 0, \\ w_j, w_j \leq 0. \end{cases}$$

Тогда ограничения в виде системы линейных неравенств принимают вид

$$\begin{cases} w_j^+ \geq 0, \\ w_j^- \geq 0, \\ \sum\limits_{j \in \mathcal{J}} (w_j^+ + w_j^-) \leq \tau. \end{cases}$$

Таким образом, оптимизационная задача с $n$ переменными и $2^n$ ограничениями преобразована в задачу с $2n$ переменными и $(2n + 1)$ ограничениями.

Поверхность $S(\mathbf{w}) = \text{const}$ в пространстве параметров модели является эллипсоидом. Система линейных неравенств, записанная выше, задает многомерный октаэдр в этом пространстве, а параметр $\tau$ — его размер. Точка касания эллипсоида и октаэдра является решением нормального уравнения при фиксированном $\tau$.

### 3.1.2. Корреляционные методы

**Ступенчатая регрессия.** Алгоритм состоит в последовательном добавлении признаков, наиболее коррелирущих с вектором регрессионных остатков. Начальный набор признаков пуст, $\mathcal{A} = \emptyset$; вектор остатков $\boldsymbol{\varepsilon}_0 = \mathbf{y}$. Рассмотрим $k$-й шаг алгоритма. Сначала находится признак с номером $j_k$, корреляция которого с вектором остатков максимальна:

$$j_k = \arg\max_{j \in \mathcal{J} \setminus \mathcal{A}} \boldsymbol{\varepsilon}_k^{\mathsf{T}} \boldsymbol{\chi}_j.$$

Затем оценивается параметр $w_{j_k}$ для найденного признака $j_k$:

$$w_{j_k} = \frac{\boldsymbol{\varepsilon}_k^T \boldsymbol{\chi}_j}{\|\boldsymbol{\chi}_j\|^2}.$$

Признак с номером $j_k$ включается в набор $\mathcal{A}$ и исключается из дальнейшего рассмотрения. Обновляется вектор регрессионных остатков

$$\boldsymbol{\varepsilon}_{k+1} = \boldsymbol{\varepsilon}_k + \tau \text{sign}\big(\boldsymbol{\varepsilon}_k^T \boldsymbol{\chi}_{j_k}\big) \boldsymbol{\chi}_{j_k},$$

где $\tau$ — достаточно маленькое число. Выбор большого значения $\tau$, например,

$$\tau = \frac{|\boldsymbol{\varepsilon}_k^T \boldsymbol{\chi}_{j_k}|}{\|\boldsymbol{\chi}_{j_k}\|^2},$$

приводит к алгоритму последовательного добавления признаков, имеющих наибольшую корреляцию с вектором регрессионных остатков. В этом случае обновление вектора регрессионных остатков задано как $\boldsymbol{\varepsilon}_{k+1} = \boldsymbol{\varepsilon}_k - w_{j_k} \boldsymbol{\chi}_{j_k}$.

**Добавление признаков с ортогонализацией.** Метод последовательного добавления признаков с ортогонализацией основан на ортогонализации признаков-столбцов $\boldsymbol{\chi}$ матрицы $X$. Ортогонализация делает возможным вычисление индивидуального вклада каждого признака в вектор значений зависимой переменной. Матрица $X$ может быть ортогонализована с помощью процедур Грамма-Шмидта или Хаусхолдера.

Запишем ортогональное разложение матрицы $\mathbf{X} = \mathbf{QR}$, где $\mathbf{Q}$ — матрица, столбцы которой являются ортогональным базисом, а $\mathbf{R}$ — верхняя треугольная матрица. Тогда

$$\mathbf{y} = \mathbf{Xw} = \mathbf{Qv},$$

где $\mathbf{v} = \mathbf{Rw}$. Пусть на $k$-м шаге получен вектор регрессионных остатков $\boldsymbol{\varepsilon}_k$. Обозначим $\mathcal{A}_k$ — набор индексов признаков, а $\bar{\mathcal{A}}_k$ — остальные признаки, $\mathcal{J} = \mathcal{A}_k \sqcup \bar{\mathcal{A}}_k$. Начальное значение вектора регрессионных остатков $\boldsymbol{\varepsilon}_0 = \mathbf{y}$, см. рис. 27 а). Начальный набор признаков $\mathcal{A}_0$ пуст. Рассмотрим $k$-й шаг алгоритма.

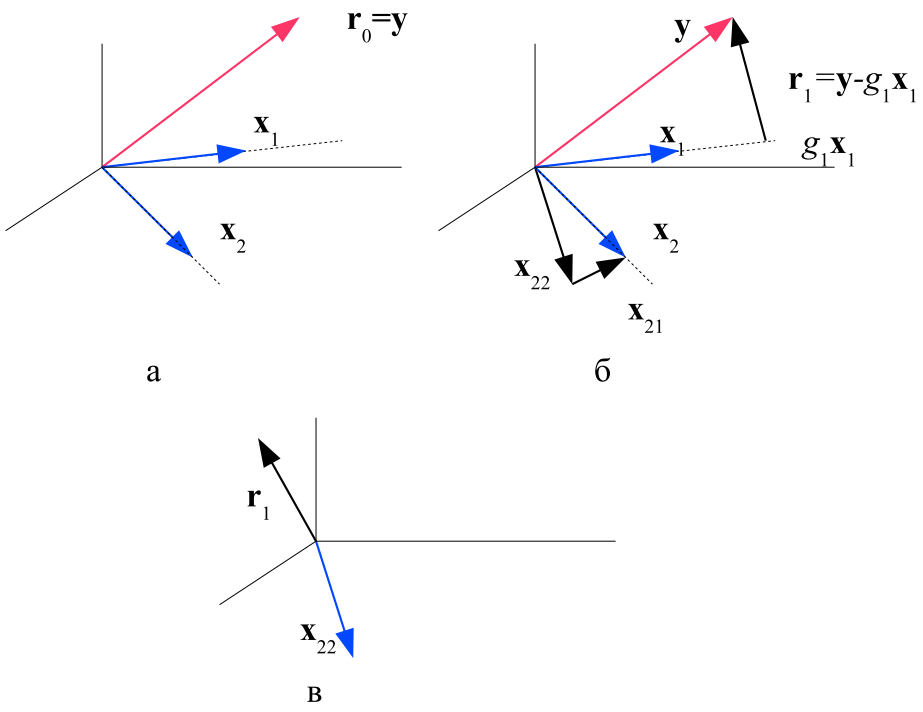



Рис. 27. Шаги алгоритма последовательного добавления признаков с ортогонализацией.

1. Находим признак $j_k$, который составляет наименьший угол с вектором остатков $\boldsymbol{\varepsilon}_k$:

$$j_k = \max_{j \in \bar{\mathcal{A}}_{k-1}} \frac{\boldsymbol{\chi}_j^{\mathsf{T}} \boldsymbol{\varepsilon}_k}{\|\boldsymbol{\chi}_j\| \|\boldsymbol{\varepsilon}_k\|}.$$

   Признак $j_k$ добавляется в набор $\mathcal{A}_k$ и удаляется из $\bar{\mathcal{A}}_k$.

2. Находим $\boldsymbol{\varepsilon}_k^{\mathrm{Pr}}$ — проекцию вектора остатков $\boldsymbol{\varepsilon}_k$ на $\boldsymbol{\chi}_{j_k}$, см. рис. 27, б):

$$\boldsymbol{\varepsilon}_k^{\mathrm{Pr}} = \frac{\boldsymbol{\chi}_{j_k}^{\mathsf{T}} \boldsymbol{\varepsilon}_k}{\|\boldsymbol{\chi}_{j_k}\|^2} \boldsymbol{\chi}_{j_k}.$$

3. Находим параметр, соответствующий добавленному признаку:

$$w_{j_k} = \frac{\|\boldsymbol{\varepsilon}_k^{\mathrm{Pr}}\|}{\|\boldsymbol{\chi}_{j_k}\|}. \tag{81}$$

4. Обновляется вектор остатков, см. рис. 27, б):

$$\boldsymbol{\varepsilon}_{k+1} = \boldsymbol{\varepsilon}_k - \boldsymbol{\varepsilon}_k^{\mathrm{Pr}}.$$

5. Признаки $\boldsymbol{\chi}_j, j \in \bar{\mathcal{A}}_k$, не входящие в набор $\mathcal{A}_k$, проецируются на подпространство, ортогональное пространству признаков из $\mathcal{A}_k$, как показано на рис. 27, в). Выполняется ортогонализация векторов признаков:

$$\mathbf{X}_{\bar{\mathcal{A}}_k} = \mathbf{X}_{\bar{\mathcal{A}}_{k-1}} - \boldsymbol{\chi}_{j_k} \frac{\boldsymbol{\chi}_{j_k}^{\mathsf{T}} \mathbf{X}_{\bar{\mathcal{A}}_{k-1}}}{\|\boldsymbol{\chi}_{j_k}\|^2}.$$

Вектор параметров $\mathbf{w}$ находится из (81):

$$\mathbf{w} = \mathbf{R}^{-1}\mathbf{v}.$$

Алгоритм последовательно добавляет признаки и, за счет ортогонализации матрицы плана $\mathbf{X}$, позволяет отбирать наименее коррелированные признаки. Всего требуется $n = |\mathcal{J}|$ шагов.

**Метод наименьших углов.** На каждом шаге алгоритма выбирается признак $\boldsymbol{\chi}_j$, имеющий наибольшую корреляцию с биссектрисой между вектором $\mathbf{y}$ и ранее добавленными признаками с индексами из множества $\mathcal{A}_{k-1}$. На $k$-м шаге только $k$ элементов вектора $\mathbf{w}$ отличны от нуля. Алгоритм последовательно вычисляет приближение

$$\mathbf{f} = \mathbf{X}\mathbf{w}$$

зависимой переменной. Для приближения используется вектор корреляций столбцов матрицы $\mathbf{X}$ с вектором остатков $\mathbf{y} - \mathbf{f}$:

$$\mathbf{c}(\mathbf{f}) = \mathbf{X}^{\intercal}(\mathbf{y} - \mathbf{f}).$$

На $k$-м шаге новое значение приближения вектора зависимой переменной $\mathbf{y}$ вычисляется как

$$\mathbf{f}_k = \mathbf{f}_{k-1} + \gamma_k \mathbf{u}_k.$$

Здесь $\mathbf{u}_k$ — вектор единичной длины, вычисляемый следующим образом. Подмножество $\mathcal{A} \subseteq \{1, \ldots, n\} = \mathcal{J}$ задает матрицу

$$\mathbf{X}_{\mathcal{A}} = [s_{j_1}\boldsymbol{\chi}_{j_1}, \ldots, s_{j_{|\mathcal{A}|}}\boldsymbol{\chi}_{j_{|\mathcal{A}|}}], j \in \mathcal{A},$$

где множитель $s \in \{+1, -1\}$ и $|\mathcal{A}|$ — количество элементов множества $\mathcal{A}$. Обозначим ковариационную матрицу

$$\boldsymbol{\Sigma}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^{\intercal}\mathbf{X}_{\mathcal{A}}$$

и

$$\delta_{\mathcal{A}} = \frac{1}{\sqrt{\mathbf{1}_{\mathcal{A}}^{\intercal}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}}}},$$

где $\mathbf{1}_{\mathcal{A}}$ — вектор, состоящий из $|\mathcal{A}|$ единиц.

Вычислим единичный вектор

$$\mathbf{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}},$$

где

$$\mathbf{w}_{\mathcal{A}} = \delta_{\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}}.$$

Вектор $\mathbf{u}_{\mathcal{A}}$ образует со столбцами матрицы $\mathbf{X}_{\mathcal{A}}$ одинаковые углы, меньшие $\frac{\pi}{2}$, поскольку справедливы равенства

$$\mathbf{X}_{\mathcal{A}}^{\intercal}\mathbf{u}_{\mathcal{A}} = \delta_{\mathcal{A}}\mathbf{1}_{\mathcal{A}} \quad \text{и} \quad \|\mathbf{u}_{\mathcal{A}}\| = 1.$$

Действительно,

$$\mathbf{X}_{\mathcal{A}}^{\intercal}\mathbf{u}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^{\intercal}\mathbf{X}_{\mathcal{A}}\mathbf{w}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}^{\intercal}\mathbf{X}_{\mathcal{A}}\delta_{\mathcal{A}}\Sigma_{\mathcal{A}}^{-1}\mathbf{1}_{\mathcal{A}} = \delta_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}^{\intercal}\mathbf{X}_{\mathcal{A}}\big(\mathbf{X}_{\mathcal{A}}^{\intercal}\mathbf{X}_{\mathcal{A}}\big)^{-1}\mathbf{1}_{\mathcal{A}} = \delta_{\mathcal{A}}\mathbf{1}_{\mathcal{A}}$$

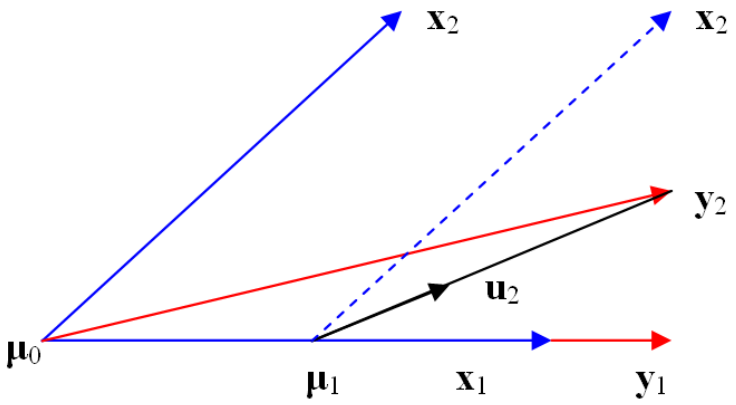



Рис. 28. Метод наименьших углов для случая $n = 2$.

и норма вектора

$$\|\mathbf{u}_{\mathcal{A}}\|^2 = \mathbf{u}_{\mathcal{A}}^T\mathbf{u}_{\mathcal{A}} = \mathbf{w}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}}^T\mathbf{u}_{\mathcal{A}} = \delta_{\mathcal{A}}^2\mathbf{1}_{\mathcal{A}}^T\big(\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\big)^{\intercal}\mathbf{1}_{\mathcal{A}} = 1.$$

Выполнение алгоритма. Назначим первое приближение вектора значений зависимой переменной $\mathbf{f} = \mathbf{0}$. Вычислим текущую оценку $\mathbf{f}_{\mathcal{A}}$ и вектор корреляций

$$\mathbf{c} = \mathbf{X}^{\intercal}(\mathbf{y} - \mathbf{f}_{\mathcal{A}}).$$

Найдем текущий набор индексов $\mathcal{A}$, соответствующих признакам с наибольшими абсолютными значениями корреляций

$$c^{\max} = \max_{j\in\mathcal{J}}|c_j| \quad \text{и} \quad \mathcal{A} = \{j : |c_j| = c^{\max}\}.$$

Пусть

$$s_j = \operatorname{sign}(c_j) \quad \text{для} \quad j \in \mathcal{A}.$$

Построим матрицу $\mathbf{X}_{\mathcal{A}}$, вычислим $\delta_{\mathcal{A}}$. Вычислим вектор $\mathbf{u}_{\mathcal{A}}$ и вектор скалярных произведений

$$\mathbf{a} = \mathbf{X}^{\intercal}\mathbf{u}_{\mathcal{A}}.$$

Пересчитаем значение вектора $\mathbf{f}_{\mathcal{A}}$:

$$\mathbf{f}_{\mathcal{A}} = \mathbf{f}_{\mathcal{A}} + \hat{\gamma}\mathbf{u}_{\mathcal{A}}, \tag{82}$$

где

$$\hat{\gamma} = \min_{j\in\mathcal{J}\setminus\mathcal{A}}{}^{+}\left\{\frac{c^{\max} - c_j}{\delta_{\mathcal{A}} - a_j}, \frac{c^{\max} + c_j}{\delta_{\mathcal{A}} + a_j}\right\}. \tag{83}$$

Минимум, обозначенный здесь как $\min_+$ здесь берется по положительным значениям аргументов для каждого $j$.

Добавим в множество $\mathcal{A}$ индекс $\hat{j}$, где $\hat{j}$ доставляет минимум соответствующему значению $\hat{\gamma}$ из выражения (83),

$$\hat{j} = \arg(\hat{\gamma}(j)).$$

Алгоритм повторяется $n$ раз.

Так как столбцы матрицы $\mathbf{X}$ предполагаются линейно независимыми, то матрица $\mathbf{\Sigma}$ не вырождена. В случае если матрица $\mathbf{\Sigma}$ является плохо обусловленной, для получения псевдообратной матрицы можно использовать сингулярное разложение.

Разъясним смысл коэффициента

$$\delta_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^{\mathsf{T}} \mathbf{\Sigma}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-\frac{1}{2}}.$$

Пусть $\mathcal{S}_{\mathcal{A}}$ — гиперплоскость

$$\mathcal{S}_{\mathcal{A}} = \left\{ \mathbf{v} = \sum_{j \in \mathcal{A}} s_j \boldsymbol{\chi}_j \rho_j : \sum_{j \in \mathcal{A}} \rho_j = 1 \right\}, \tag{84}$$

где $\rho_j$ может быть отрицательным. Вектор из геометрического множества точек $\mathcal{S}_{\mathcal{A}}$, имеющий наименьшую длину, равен

$$\mathbf{v}_{\mathcal{A}} = \delta_{\mathcal{A}} \mathbf{u}_{\mathcal{A}} = \delta_{\mathcal{A}} \mathbf{X}_{\mathcal{A}} \mathbf{w}_{\mathcal{A}},$$

где

$$\mathbf{w}_{\mathcal{A}} = \delta_{\mathcal{A}} \mathbf{\Sigma}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}$$

и

$$\|\mathbf{v}_{\mathcal{A}}\| = \delta_{\mathcal{A}}.$$

Действительно, квадрат нормы любого вектора из $\mathcal{S}_{\mathcal{A}}$ равен

$$\|\mathbf{X}_{\mathcal{A}} \boldsymbol{\rho}\|^2 = \boldsymbol{\rho}^{\mathsf{T}} \mathbf{\Sigma}_{\mathcal{A}} \boldsymbol{\rho}.$$

Выпишем лагранжиан с ограничением на сумму элементов $\rho_j$ вектора $\boldsymbol{\rho}$:

$$\boldsymbol{\rho}^{\mathsf{T}} \mathbf{\Sigma}_{\mathcal{A}} \boldsymbol{\rho} - \lambda\big(\mathbf{1}_{\mathcal{A}}^{\mathsf{T}} \boldsymbol{\rho} - 1\big).$$

Минимизируя по $\boldsymbol{\rho}$, получаем

$$\boldsymbol{\rho} = \lambda \mathbf{\Sigma}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}}, \quad \boldsymbol{\rho} = \delta_{\mathcal{A}}^2 \mathbf{\Sigma}_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}} = \delta_{\mathcal{A}} \mathbf{w}_{\mathcal{A}}$$

и

$$\mathbf{v}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} \boldsymbol{\rho} \in \mathcal{S}_{\mathcal{A}}.$$

Норма вектора $\mathbf{v}_{\mathcal{A}}$

$$\|\mathbf{v}_{\mathcal{A}}\|^2 = \boldsymbol{\rho}^{\mathsf{T}} \mathbf{\Sigma}_{\mathcal{A}}^{-1} \boldsymbol{\rho} = \delta_{\mathcal{A}}^4 \mathbf{1}_{\mathcal{A}}^{\mathsf{T}} (\mathbf{\Sigma}^{-1})_{\mathcal{A}}^{\mathsf{T}} \mathbf{1}_{\mathcal{A}} = \delta_{\mathcal{A}}^2.$$

Величина $\hat{\gamma}$ в (83) интерпретируется следующим образом. Запишем оценку вектора зависимой переменных $\mathbf{f}$ как функцию от $\hat{\gamma}$

$$\mathbf{f}(\hat{\gamma}) = \mathbf{f}_{\mathcal{A}} + \hat{\gamma} \mathbf{u}_{\mathcal{A}}$$

при условии $\hat{\gamma} > 0$. Корреляция вектора регрессионных остатков с добавляемым $j$-м признаком равна

$$c_j(\hat{\gamma}) = \boldsymbol{\chi}_j^{\mathsf{T}} \big(\mathbf{y} - \mathbf{f}(\hat{\gamma})\big) = c_j - \hat{\gamma} a_j.$$

Для $j \in \mathcal{A}$ получаем

$$|c_j(\hat{\gamma})| = c^{\max} - \hat{\gamma} \delta_{\mathcal{A}}.$$

Это означает, что все рассматриваемые на данном шаге максимальные абсолютные корреляции уменьшаются на одну и ту же величину. Из предыдущих двух соотношений видно, что если $j \in \mathcal{A}^c$, то корреляция $c_j(\gamma)$ принимает наибольшее значение при

$$\hat{\gamma} = \frac{c^{\max} - c_j}{\delta_{\mathcal{A}} - a_j}.$$

Аналогично корреляция $-c_j(\hat{\gamma})$ принимает наибольшее значение при

$$\hat{\gamma} = \frac{c^{\max} + c_j}{\delta_{\mathcal{A}} + a_j}.$$

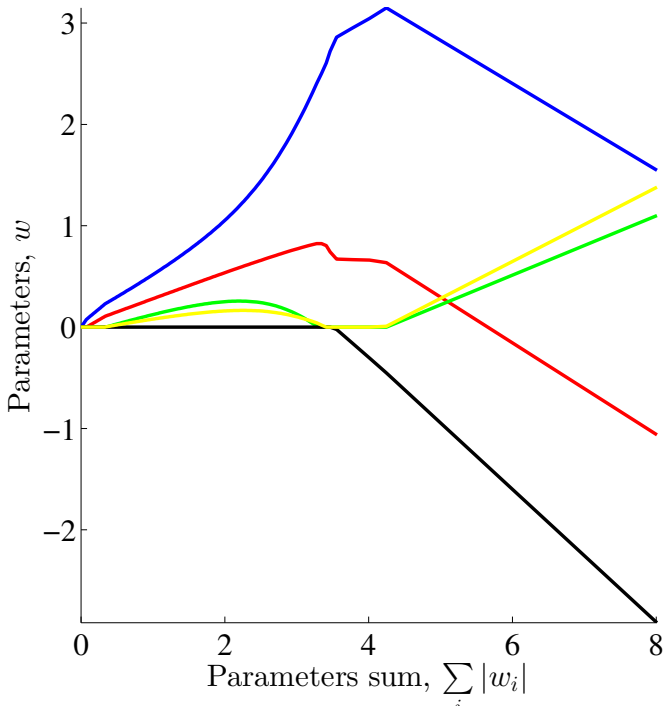



Рис. 29. Оценки параметров для метода наименьших углов в зависимости от их нормы $\|\mathbf{w}\|_1$.

Таким образом, $\hat{\gamma}$ в выражении (83) — минимальная положительная величина, при которой новый индекс $j$ может быть добавлен в набор $\mathcal{A}$.

С помощью модификаций алгоритма наименьших углов можно получить решения Лассо и ступенчатой регрессии. Основным достоинством алгоритма является то, что он выполняется за число шагов, равное числу свободных переменных.

Для иллюстрации основного недостатка алгоритма рассмотрим следующий пример. Пусть матрица $\mathbf{X}$ состоит из столбцов значений трех признаков. Первый признак $\boldsymbol{\chi}_1$ значительно коррелирует с вектором зависимых переменных $\mathbf{y}$, который является линейной комбинацией остальных двух признаков $\mathbf{x}_2$ и $\mathbf{x}_3$. Например,

$$\mathbf{X} = [\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

Алгоритм на первом шаге выберет первый признак, так как он сильнее коррелирует с вектором зависимых переменных, и затем присоединит остальные признаки. Ошибка модели, полученной с помощью этого алгоритма, будет отлична от нуля в то время, когда существует модель, доставляющая нулевую ошибку, и векторы-признаки, входящие в модель, ортогональны. Для разрешения этого недостатка ниже предложен алгоритм, позволяющий удалять мультиколлинеарные признаки и добавлять признаки, уменьшающие значение функции ошибки.

### 3.1.3. Прореживающие методы

Прореживающие методы являются обобщением методов последовательного удаления признаков. Они последовательно исключают параметры моделей согласно принятым критериям. Благодаря этому, такие методы можно использовать как для удаления признаков обобщенно-линейных моделей, так и для удаления элементов нелинейных моделей. При этом каждому элементу нелинейной модели должен быть поставлен в соответствие параметр.

**Оптимальное прореживание.** Оптимальное прореживание — метод упрощения структуры регрессионной модели. Основная идея прореживания заключается в том, что те элементы модели, которые оказывают малое влияние на функцию ошибки $S(\mathbf{w})$, можно исключить из модели без значительного ухудшения качества аппроксимации.

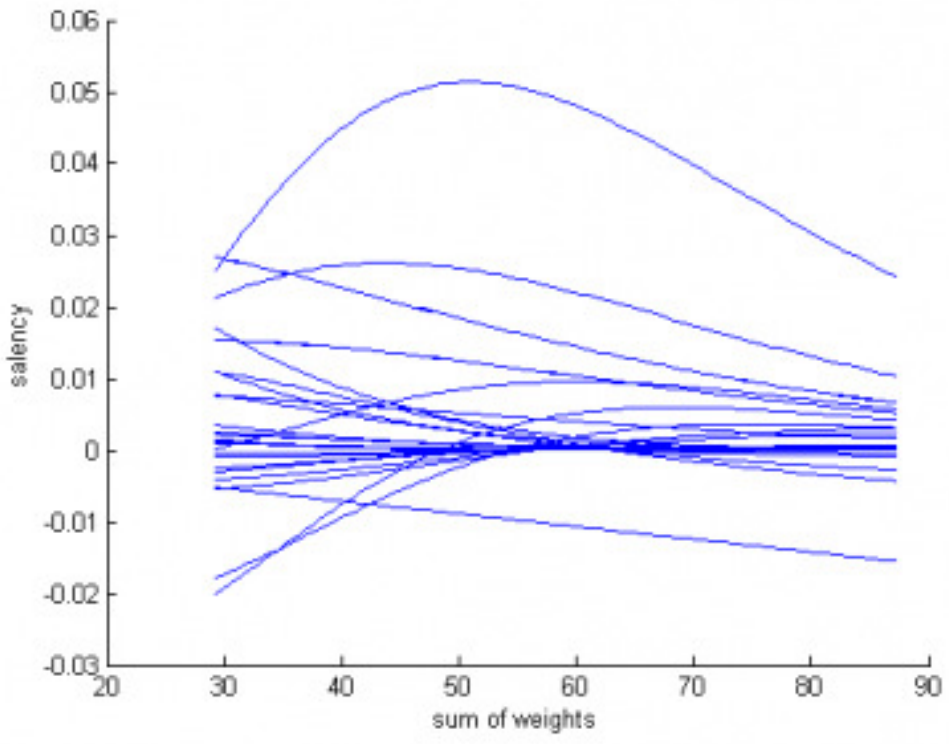



Рис. 30. Функция выпуклости параметров модели.

Рассмотрим регрессионную модель общего вида $\mathbf{f}(\mathbf{w}, \mathbf{X})$ и функцию ошибки $S(\mathbf{w}) = \|\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\|^2$. Найдем локальную аппроксимацию функции $S(\mathbf{w})$ в окрестности произвольной

точки $\mathbf{w}$ с помощью разложения в ряд Тейлора:

$$S(\mathbf{w}+\Delta\mathbf{w}) = S(\mathbf{w}) + \mathbf{g}^{\intercal}(\mathbf{w})\Delta\mathbf{w} + \frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}\Delta\mathbf{w} + O(\|\Delta\mathbf{w}\|^3),$$

где $\Delta\mathbf{w}$ — приращение вектора параметров $\mathbf{w}$, $\mathbf{g}$ — градиент,

$$\mathbf{g} = \frac{\partial S(\mathbf{w})}{\partial \mathbf{w}},$$

и $\mathbf{H} = \mathbf{H}(\mathbf{w})$ — матрица вторых производных, или матрица Гессе,

$$H = \frac{\partial^2 S(\mathbf{w})}{\partial \mathbf{w}^2}.$$

Предполагается, что функция $S(\mathbf{w})$ достигает своего минимума при значении параметров $\mathbf{w} = \hat{\mathbf{w}}$. Таким образом, предыдущее выражение можно упростить и представить в виде

$$\Delta S(\hat{\mathbf{w}}) = S(\hat{\mathbf{w}}+\Delta\mathbf{w}) - S(\hat{\mathbf{w}}) \approx \frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}(\hat{\mathbf{w}})\Delta\mathbf{w}.$$

Пусть исключение элемента функции регрессии $\mathbf{f}(\hat{\mathbf{w}}, \mathbf{X})$ есть исключение одного параметра, например, $w_j$. Индекс $j \in \mathcal{J}$ в данном случае считаем номером элемента вектора параметров $\mathbf{w} = [w_1, \ldots, w_j, \ldots, w_{|\mathcal{J}|}]^{\intercal}$. Число элементов вектора $\mathbf{w}$ может быть не равно числу элементов вектора $\mathbf{x}$ — строки матрицы плана $\mathbf{X}$.

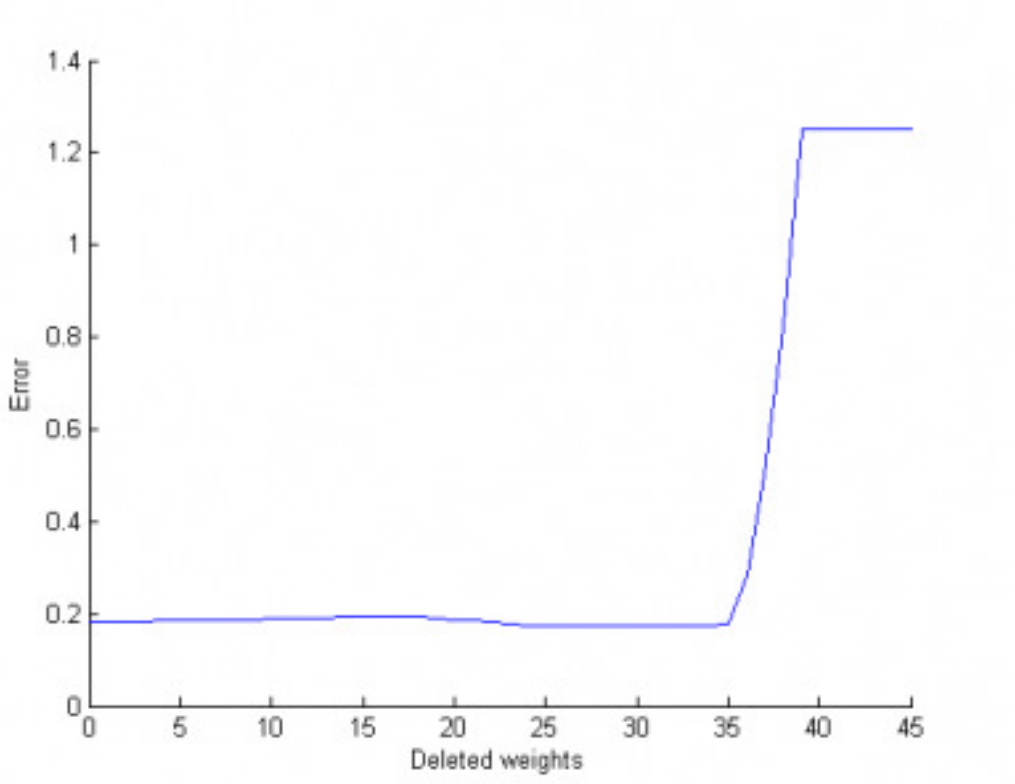
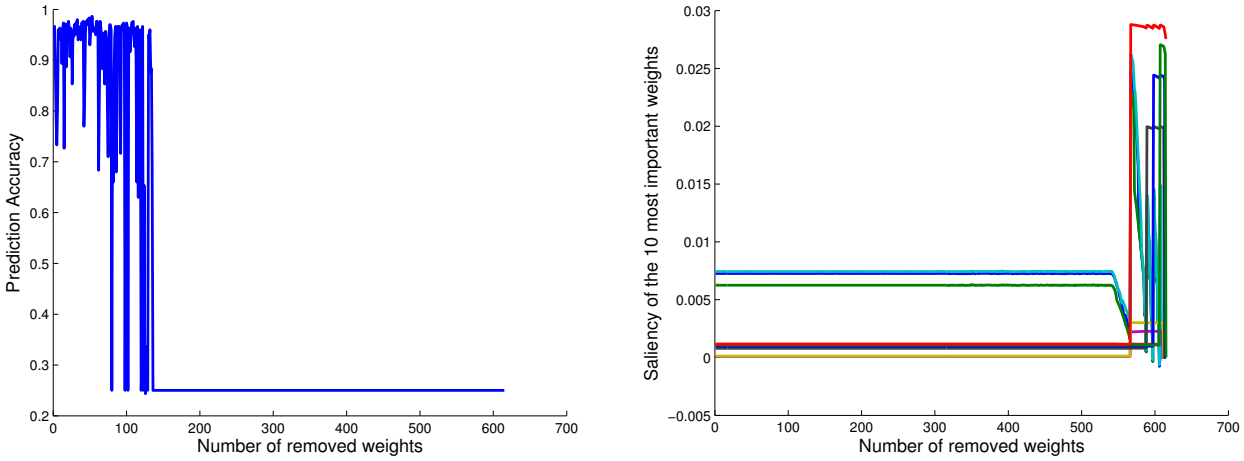



Рис. 31. Зависимость функции ошибки от числа параметров модели.

На рис. 32 показан метод Optimal Brain Damage для удаления параметров в двухслойной нейронной сети.

Пусть исключение элемента эквивалентно выражению $\Delta w_j + w_j = 0$, иначе

$$\mathbf{e}_j^{\intercal}\Delta\mathbf{w} + w_j = 0,$$




(a) Зависимость точности прогноза от количества удаляемых параметров

(b) Зависимость функции выпуклости от количества удаляемых параметров

Рис. 32. Метод OBD для двухслойной нейронной сети.

где $\mathbf{e}_j$ — вектор, $j$-й элемент которого равен единице, все остальные элементы равны нулю. Это самое сильное ограничение, не позволяющее применять данный метод для регрессионных моделей произвольного вида.

Для нахождения элемента, который нужно исключить, требуется минимизировать квадратичную форму $\Delta\mathbf{w}^\intercal\mathbf{H}\Delta\mathbf{w}$ относительно $\Delta\mathbf{w}$ при ограничениях $\mathbf{e}_j^\intercal\Delta\mathbf{w} + w_j = 0$, для всех значений $j \in \mathcal{J}$. Индекс $\hat{j}$, который доставляет минимум квадратичной форме, задает номер исключаемого элемента:

$$\hat{j} = \underset{j\in\mathcal{J},\ \Delta\mathbf{w}\in\mathbb{W}}{\arg\min}\ (\Delta\mathbf{w}^\intercal\mathbf{H}\Delta\mathbf{w}) \quad \text{при} \quad \mathbf{e}_j^\intercal\Delta\mathbf{w} + w_j = 0.$$

Задача условной минимизации решается с помощью введения множителя Лагранжа $\lambda$:

$$S(\Delta\mathbf{w}) = \Delta\mathbf{w}^\intercal\mathbf{H}\Delta\mathbf{w} - \lambda(\mathbf{e}_j^\intercal\Delta\mathbf{w} + w_j). \tag{85}$$

Дифференцируя лагранжиан (85) по приращению параметров $\Delta\mathbf{w}$ и приравнивая его градиент к нулю, получаем для каждого индекса $j$ параметра $w_j$

$$\Delta\mathbf{w} = -\frac{w_j}{[\mathbf{H}^{-1}]_{jj}}\mathbf{H}^{-1}\mathbf{e}_j.$$

Этому значению вектора приращений параметров соответствует минимальное значение лагранжиана

$$L_j = \frac{w_j^2}{2[\mathbf{H}^{-1}]_{jj}}.$$

Полученное выражение называется мерой выпуклости функции ошибки $S(\Delta\mathbf{w})$ при изменении параметра $w_j$.

Функция $L_j$ зависит от квадрата параметра $w_j$. Это говорит о том, что параметр с малым значением $w_j$ должен быть удален из модели. Однако если $j$-й диагональный элемент $[\mathbf{H}^{-1}]_{jj}$ матрицы, обратной матрицы Гессе, достаточно мал, это означает, что данный параметр оказывает существенное влияние на функцию ошибки.

Для упрощения структуры регрессионной модели выполняются следующие шаги.

1. Оцениваем параметры модели $\hat{\mathbf{w}} = \arg\min\limits_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|f, \mathfrak{D})$.

2. Для вектора $\hat{\mathbf{w}} + \Delta\mathbf{w}$ решаем оптимизационную задачу, находим для каждого индекса $j$ минимальное значение Лагранжиана $L_j$.

3. Выбираем среди $L_j$ минимальное, исключаем элемент модели, соответствующий $j$-му параметру.

4. Добавляем к вектору параметров $\hat{\mathbf{w}}$, вектор приращений $\Delta\mathbf{w}_j$, соответствующий исключаемому параметру с индексом $j$, либо переходим к первому шагу.

Шаги 1–4 процедуры повторяются до тех пор, пока значение функции ошибки $S(\mathbf{w})$ не превзойдет заданное или до достижения минимума одного из критериев сложности модели.

### 3.1.4. Шаговые методы

Шаговыми методами называются методы, заключающиеся в последовательном удалении или добавлении признаков линейных или обобщенно-линейных моделей, согласно определенному критерию.

**Шаговая регрессия и критерии формирования набора признаков.** На шаге с номером $k$ сравнивается текущая модель задаваемая набором признаков $\mathcal{A}_k$ и все модели, построенные на этом наборе путем удаления или добавления одного признака. При выборе модели следующего, $k+1$-го шага, используется критерий Фишера или $F$-критерий, который сравнивает значения функций среднеквадратичных ошибок двух моделей:

$$F = \frac{S_1 - S_2}{S_2}\frac{m - n_2}{n_1 - n_2}.$$

Индекс 2 соответствует текущей линейной модели, а индекс 1 соответствует новой линейной модели, которая является модификацией первой модели; $n_1, n_2$ — соответствующие числа параметров моделей, $m$ — объем регрессионной выборки. Если значение критерия больше заданного, то вторая модель считается лучше первой. Отметим, что согласно изложенному подходу, знаменатель правого сомножителя $n_1 - n_2$ равен $+1$ или $-1$ в зависимости от шага.

Обозначим $S_{\mathcal{A}}$ значение функции ошибки, которое модель, заданная набором индексов признаков $\mathcal{A}$, имеет после оценки своих параметров. Пусть на $k$-м шаге набор признаков задан множеством $\mathcal{A}$. На первом шаге начальным набором является пустой набор $\mathcal{A} = \emptyset$. На $k$-м шаге к текущему набору $\mathcal{A}$ присоединяется признак $\hat{j} \in \mathcal{J} \setminus \mathcal{A}$, который доставляет максимум $F$-критерию:

$$\hat{j} = \arg\max_{j\in\mathcal{J}\setminus\mathcal{A}} F = \arg\max_{j\in\mathcal{J}\setminus\mathcal{A}} \left(\frac{S_{\mathcal{A}\cup\{j\}} - S_{\mathcal{A}}}{S_{\mathcal{A}}}(m - |\mathcal{A}|)\right). \tag{86}$$

При последовательном удалении признаков начальный набор состоит из всех признаков $\mathcal{A} = \mathcal{J}$. На каждом шаге происходит удаление признака $j \in \mathcal{A}$ так, чтобы значение $F$-критерия было минимально:

$$\hat{j} = \arg\min_{j\in\mathcal{A}} F = \arg\min_{j\in\mathcal{A}} \left(\frac{S_{\mathcal{A}\setminus\{j\}} - S_{\mathcal{A}}}{S_{\mathcal{A}}}(|\mathcal{A}| - m)\right). \tag{87}$$

В ходе процедуры происходит смена шагов (86) и (87). Задаются некоторые фиксированные пороговые значения $F_{\text{Add}}$ и $F_{\text{Del}}$. Индексы признаков добавляются в набор $\mathcal{A}$ до тех пор, пока значение $F$-критерия на некотором шаге не станет меньше $F_{\text{Add}}$. Затем из набора $\mathcal{A}$ удаляются индексы до тех пор, пока значение $F$-критерия для которых не превзойдет $F_{\text{Del}}$. Останов процедуры производится при достижении минимума, заданного критерием Маллоуза $C_p$:

$$C_p = \frac{S_{\mathcal{A}}}{S_{\mathcal{J}}} + 2n - m,$$

где $S_{\mathcal{J}}$ — среднеквадратичная ошибка, вычисленная для модели, настроенной с помощью метода наименьших квадратов на всем множестве признаков $\mathcal{J}$, $n = |\mathcal{J}|$ — число признаков. Критерий штрафует модели с большим числом признаков. Минимизация критерия позволяет найти множество значимых признаков, иначе — оптимальную модель.

Основное преимущество шаговой регрессии — она применима в случае, когда число признаков, из которых надо выбрать оптимальный набор, велико.

## 3.2. Сходимость при последовательном добавлении признаков

Рассматривается задача последовательного добавления признаков в регрессионную модель. Решается вопрос остановки процедуры добавления. Используемые при этом внешние критерии или критерии сложности модели позволяют определить сложность модели, выражаемую, например, числом включаемых в модель признаков. Однако, с их помощью невозможно определить, насколько отличается текущая порождаемая модель от оптимальной [168]. Для этого вводится понятие расстояния между последовательно порождаемыми моделями.

Ранее в работе рассматривались функции структурных расстояний между порождаемыми моделями. В настоящем разделе рассмотрим расстояния, основанные на сравнении векторов регрессионных остатков моделей. В задачах выбора последовательно порождаемых моделей требуется определить расстояние до неизвестной модели оптимальной сложности. В случае, когда число элементов регрессионной выборки стремится к бесконечности, используется сходимость по вероятности матрицы ковариации.

### 3.2.1. Расстояние между последовательно порождаемыми моделями

Рассмотрим линейные модели вида

$$f_1 : \mathbf{y} = \mathbf{X}_1\mathbf{w}_1 + \boldsymbol{\varepsilon}_1, \tag{88}$$

$$f_2 : \mathbf{y} = \mathbf{X}_1\mathbf{w}_1 + \mathbf{X}_2\mathbf{w}_2 + \boldsymbol{\varepsilon}_2. \tag{89}$$

Матрица плана $\mathbf{X}_1, \mathbf{X}_2$ имеют соответственно размер $m \times n_1$ и $m \times n_2$ элементов. Предполагается, что многомерная случайная величина $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ распределена нормально с нулевым матожиданием.

Также предполагается существование некоторой «истинной» модели

$$f_3 : \mathbf{y} = \underset{m\times n_1}{\mathbf{X}_1}\ \mathbf{w}_1 + \underset{m\times n_2}{\mathbf{X}_2}\ \mathbf{w}_2 + \underset{m\times n_3}{\mathbf{X}_3}\ \mathbf{w}_3 + \boldsymbol{\varepsilon}, \tag{90}$$

в которой матрица $X_3$ имеет $m \times n_3$ элементов. Заметим, что для моделей $f_1, f_2$ случайные величины $\boldsymbol{\varepsilon}_1$ и $\boldsymbol{\varepsilon}_2$ определены через случайную величину $\boldsymbol{\varepsilon}$ истинной модели:

$$\boldsymbol{\varepsilon}_1 = \mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon} \quad \text{и} \quad \boldsymbol{\varepsilon}_2 = \mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon}.$$

Сравним дисперсии регрессионных остатков истинной модели $f_3$ и моделей-претендентов $f_1, f_2$. Идемпотентная симметричная матрица проекции случайной величины $\mathbf{y}$ на пространство столбцов матрицы $\mathbf{X}$ равна $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$. Обозначим $\mathbf{R} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$ матрицу, проецирующую произвольный вектор $\mathbf{y}$ на дополнение к пространству столбцов $\mathbf{X}$. Здесь матрица $\mathbf{X} = \mathbf{X}_1 \vdots \mathbf{X}_2$ получена путем соединения матриц $\mathbf{X}_1$ и $\mathbf{X}_2$ по столбцам.

За оценку дисперсии регрессионных остатков $\boldsymbol{\varepsilon}_2$ модели $f_2$ примем величину

$$\hat{\sigma}^2(\boldsymbol{\varepsilon}_2) = \frac{1}{m}\mathbf{y}^\intercal\mathbf{R}\mathbf{y}, \tag{91}$$

поскольку

$$\hat{\boldsymbol{\varepsilon}}_2 = \mathbf{R}\mathbf{y}.$$

В силу истинности модели $f_3$, подставим (90) в (91) и получим:

$$m\sigma^2(\boldsymbol{\varepsilon}_2) = \mathbf{w}_3^\intercal\mathbf{X}_3^\intercal\mathbf{R}\mathbf{X}_3\mathbf{w}_3 + 2\mathbf{w}_3^\intercal\mathbf{X}_3^\intercal\mathbf{R}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^\intercal\mathbf{R}\boldsymbol{\varepsilon}, \tag{92}$$

что доказывается путем использования свойств проекционной матрицы $\mathbf{R}$:

$$\mathbf{R}^2 = \mathbf{R}, \quad \mathbf{R}^\intercal = \mathbf{R} \quad \text{и} \quad \mathbf{R}\mathbf{X} = \mathbf{0}.$$

Рассмотрим оценку дисперсии регрессионных остатков $\hat{\sigma}^2(\boldsymbol{\varepsilon}_3)$ третьей, истинной модели (90). В [160] показано, что при достаточном числе $m$ элементов выборки неравенство

$$\hat{\sigma}^2(\boldsymbol{\varepsilon}_3) < \hat{\sigma}^2(\boldsymbol{\varepsilon}_2)$$

справедливо с вероятностью, близкой к единице; евклидова норма $\|2\mathbf{w}_3^\intercal\mathbf{X}_3^\intercal\mathbf{R}\boldsymbol{\varepsilon}\|$ второго слагаемого (92) стремится к нулю при увеличении объема выборки $m$.

### 3.2.2. Расстояние между функциями регрессии

Оценим наиболее правдоподобные параметры $\mathbf{w}$ моделей $f$ и обозначим вектор регрессионных остатков функции регрессии, которая соответствует модели $f_2$ как

$$\hat{\boldsymbol{\varepsilon}}_2 = \mathbf{y} - \mathbf{X}_1\hat{\mathbf{w}}_1 - \mathbf{X}_2\hat{\mathbf{w}}_2 = \mathbf{y} - \mathbf{X}\hat{\mathbf{w}}.$$

Введем функцию расстояния $\rho(f_k, f_l) = \|\cdot\|^2$ — евклидову норму разности векторов регрессионных остатков двух моделей и обозначим расстояния между моделями $D_1 = \rho(f_1, f_3)$, $D_2 = \rho(f_2, f_3)$ и $D = \rho(f_1, f_2)$. При этом считаем, что объем выборки $m$ фиксирован, конечен, и поэтому в дальнейшем не учитывается. Таким образом,

$$D_1 = \|\hat{\boldsymbol{\varepsilon}}_1 - \boldsymbol{\varepsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}} - (\mathbf{y} - \mathbf{X}\mathbf{w} - \mathbf{X}_3\mathbf{w}_3)\|^2. \tag{93}$$

Считая модель (90) истинной, запишем

$$\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{P}\mathbf{X}_3\mathbf{w}_3 + \mathbf{P}\boldsymbol{\varepsilon}, \tag{94}$$

где

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal$$

и, как было сказано ранее, $\mathbf{P} = \mathbf{I} - \mathbf{R}$ — идемпотентная симметричная проекционная матрица, получаемая при проектировании вектора $\mathbf{y}$ на пространство столбцов матрицы $\mathbf{X}$. Эта матрица имеет свойство

$$\mathbf{RP} = (\mathbf{I} - \mathbf{P})\mathbf{P} = \mathbf{0}.$$

Подставляя (94) в (93), получаем

$$D_1 = \|\mathbf{RX}_3\mathbf{w}_3 - \mathbf{P}\hat{\boldsymbol{\varepsilon}}\|^2 = (\mathbf{RX}_3\mathbf{w}_3 - \mathbf{P}\boldsymbol{\varepsilon})^\intercal(\mathbf{RX}_3\mathbf{w}_3 - \mathbf{P}\boldsymbol{\varepsilon}) = \mathbf{w}_3^\intercal\mathbf{X}_3^\intercal\mathbf{RX}_3\mathbf{w}_3 + \boldsymbol{\varepsilon}^\intercal\mathbf{P}\boldsymbol{\varepsilon}.$$

Первая производная $D_1$ относительно вектора $\mathbf{X}_3\mathbf{w}_3$ равна

$$\frac{\partial D_1}{\partial \mathbf{X}_3\mathbf{w}_3} = 2\mathbf{RX}_3\mathbf{w}_3. \tag{95}$$

Идемпотентная и симметричная матрица является неотрицательно определенной. Так как предполагается, что никакой вектор-столбец матрицы $\mathbf{X}_3$ не может быть представлен в виде линейной комбинации некоторого набора столбцов матрицы $\mathbf{X}_1\vdots\mathbf{X}_2$, экстремум функции расстояния $D_1$ может быть получен при $\mathbf{X}_3\mathbf{w}_3 = \mathbf{0}$. Вторая производная функции расстояния $D_1$ по $\mathbf{X}_3\mathbf{w}_3$ —

$$\frac{\partial^2 D_1}{\partial \mathbf{X}_3\mathbf{w}_3 \partial(\mathbf{X}_3\mathbf{w}_3)^\intercal} = 2\mathbf{R}$$

также является неотрицательно определенной, и следовательно, экстремум, полученный выражением (95), является минимумом.

Расстояние между моделями $f_2$ и $f_3$ определено аналогично (93) как

$$D_2 = \|\hat{\boldsymbol{\varepsilon}}_2 - \boldsymbol{\varepsilon}\|^2, \tag{96}$$

где $\hat{\boldsymbol{\varepsilon}}_2 = \mathbf{y} - \mathbf{X}_1\hat{\mathbf{w}}_1$. Правая часть предыдущего равенства может быть переписана в виде

$$D_2 = \| - \mathbf{X}_1\hat{\mathbf{w}}_1 + \mathbf{X}_1\mathbf{w}_1 + \mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3\|^2.$$

При

$$\mathbf{X}_1\hat{\mathbf{w}} = \mathbf{X}_1(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal\mathbf{y} = \mathbf{X}_1\mathbf{w}_1 + \mathbf{P}_1\mathbf{X}_2\mathbf{w}_2 + \mathbf{P}_1\mathbf{X}_2\mathbf{w}_3 + \mathbf{P}_1\boldsymbol{\varepsilon},$$

где

$$\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal = \mathbf{I} - \mathbf{R}_1$$

подставлено в предыдущее выражение, получаем (96) в виде

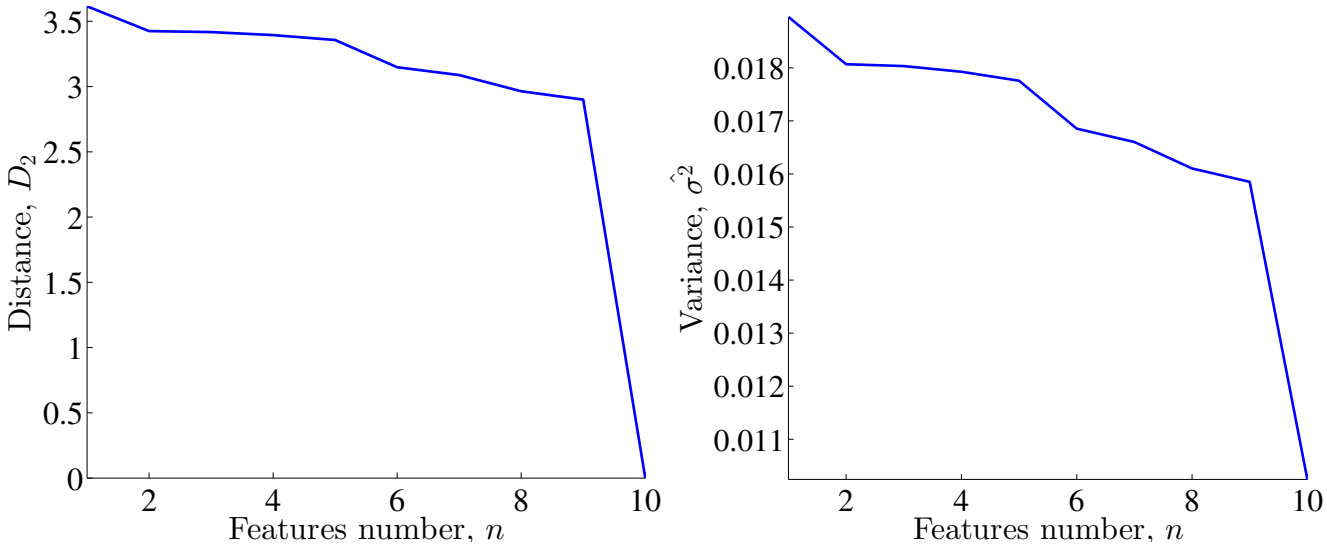
$$D_2 = \|\mathbf{R}_1(\mathbf{X}_1\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3) - \mathbf{P}_1\boldsymbol{\varepsilon}\|^2.$$

Используем свойство $\mathbf{R}_1\mathbf{P}_1 = \mathbf{R}_1(\mathbf{I} - \mathbf{R}_1) = \mathbf{R}_1 - \mathbf{R}_1^2 = \mathbf{0}$ (матрицы, обратной к $\mathbf{R}_1$ не существует) получаем

$$D_2 = (\mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3)^\intercal\mathbf{R}_1(\mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3) + \boldsymbol{\varepsilon}^\intercal\mathbf{P}_1\boldsymbol{\varepsilon}.$$

Первая производная функции расстояния $D_2$ по вектору $\mathbf{X}_3\mathbf{w}_3$ равна

$$\frac{\partial D_2}{\partial \mathbf{X}_3\mathbf{w}_3} = \mathbf{R}_1(\mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3) + (\mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3)^\intercal\mathbf{R}_1 = 2\mathbf{R}_1(\mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3).$$




(a) (b)

Рис. 33. Сходимость при последовательном добавлении признаков.

При нахождении экстремума функции $D_2$ предполагаем, что матрицы $\mathbf{X}_2$ и $\mathbf{X}_3$ невырождены. Так же как и ранее, предполагается, что каждый из столбцов этих матриц не является линейной комбинацией столбцов прочих матриц или присоединений этих матриц. В таком случае экстремум функции $D_2$ достигается при

$$\mathbf{X}_3\mathbf{w}_3 = -\mathbf{X}_2\mathbf{w}_2.$$

Этот экстремум является минимумом, так как вторая производная функции $D_2$ равна

$$\frac{\partial^2 D_2}{\partial \mathbf{X}_3\mathbf{w}_3 \partial (\mathbf{X}_3\mathbf{w}_3)^\intercal} = 2\mathbf{R}_1$$

и неотрицательно определена, то есть для произвольного вектора $\mathbf{y} \in \mathbb{R}^m$ выполняется неравенство

$$\mathbf{x}^\intercal \mathbf{R}_1 \mathbf{x} \geqslant 0. \tag{97}$$

Расстояние между двумя функциями регрессии, которые соответствуют моделям $f_1$ и $f_2$ определено аналогично (93) и (96) как

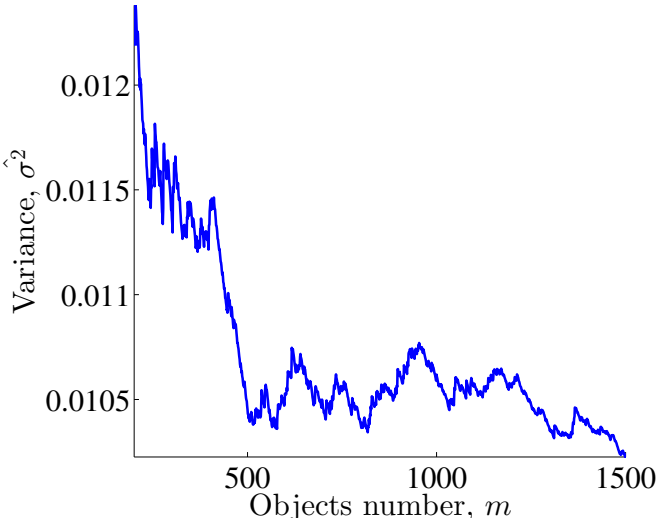
$$D = \|\hat{\boldsymbol{\varepsilon}}_2 - \hat{\boldsymbol{\varepsilon}}_1\|^2 = \hat{\boldsymbol{\varepsilon}}_2^\intercal \hat{\boldsymbol{\varepsilon}}_2 - 2\hat{\boldsymbol{\varepsilon}}_2^\intercal \hat{\boldsymbol{\varepsilon}}_1 + \hat{\boldsymbol{\varepsilon}}_1^\intercal \hat{\boldsymbol{\varepsilon}}_1.$$

### 3.2.3. Критерии сходимости при выборе моделей

При неизвестной истинной модели $f_3$, для выбора между моделями $f_1$ и $f_2$ может быть использован следующий критерий. Рассмотрим гипотезу выбора модели $f_1$ при условии неравенства

$$\hat{\boldsymbol{\varepsilon}}_2^\intercal \hat{\boldsymbol{\varepsilon}}_2 > q\hat{\boldsymbol{\varepsilon}}_1^\intercal \hat{\boldsymbol{\varepsilon}}_1 \quad \text{при} \quad q > 1, \tag{98}$$

в котором $q = q(m, k_1, k_2)$ — функция от объема выборки $m$ и числа признаков, которые включают первая и вторая модели. Увеличение значения $q$ снижает вероятность того, что




(a)

Рис. 34. Зависимость дисперсии регрессионных остатков от числа объектов.

модель $f_1$ будет принята. Прибавим одновременно к левой и правой части одни и те же члены, взятые из (3.2.2.):

$$\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 - 2\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1 + \hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1 > q\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1 - 2\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1 + \hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1.$$

Так как проекционные матрицы $\mathbf{P}_1\mathbf{P} = \mathbf{P}_1$ и $\mathbf{R}_1\mathbf{R} = \mathbf{R}$, то $\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1 = \hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1$. При этом расстояние между функциями $f_1$ и $f_2$, см. (3.2.2.), может быть записано как

$$D = \hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 - \hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1.$$

Из этого следует, что критерий выбора модели может быть задан следующим образом: гипотеза $f_1$ принимается, если расстояние $D$ превосходит некоторый критический порог $Q$, например

$$D > Q, \quad \text{где} Q = (q-1)\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1.$$

Учитывая предположение об истинности модели (90), выразим функции $D$ и $Q$, используя векторы $\mathbf{X}_3\mathbf{w}_3$ и $\boldsymbol{\varepsilon}$:

$$D = \hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 - \hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1 = \mathbf{y}^{\mathsf{T}}\mathbf{R}_1\mathbf{y} - \mathbf{y}^{\mathsf{T}}\mathbf{R}\mathbf{y} = (\mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon})(\mathbf{R}_1 - \mathbf{R})(\mathbf{X}_2\mathbf{w}_2 + \mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon})$$

и

$$Q = (q-1)\big((\mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon})^{\mathsf{T}}\mathbf{R}(\mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon})\big).$$

Функции $D$ и $Q$ являются неотрицательными симметричными квадратичными функциями от вектора $\mathbf{X}_3\mathbf{w}_3$, так как $\mathbf{R} - \mathbf{R}_1 = \mathbf{R}\mathbf{P}_1$, и матрица $\mathbf{R}$ неотрицательно определена. Первые производные функций $D$ и $Q$ по $\mathbf{X}_3\mathbf{w}_3$ имеют вид

$$\frac{\partial D}{\partial \mathbf{X}_3\mathbf{w}_3} = 2\mathbf{R}_1\mathbf{P}(\mathbf{X}_2\mathbf{w}_3 + \mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon})$$

и

$$\frac{\partial Q}{\partial \mathbf{X}_3\mathbf{w}_3} = 2(q-1)\mathbf{R}(\mathbf{X}_3\mathbf{w}_3 + \boldsymbol{\varepsilon}).$$

Снова предполагая, что матрицы $\mathbf{X}_2$ и $\mathbf{X}_3$ невырожденны, минимальное значение, ноль, для функции расстояния $D$ и критерия $Q$ будет достигнуто при

$$\mathbf{X}_3\mathbf{w}_3 = -\mathbf{X}_2\mathbf{w}_2 - \boldsymbol{\varepsilon} \quad \text{и} \quad \mathbf{X}_3\mathbf{w}_3 = -\boldsymbol{\varepsilon},$$

соответственно. Из вышеизложенного следует, что выбор модели может определяться значением вектора $\mathbf{X}_3\mathbf{w}_3$. Однако на выбор также влияют вектор $\mathbf{X}_3\mathbf{w}_2$, проекционная матрица $\mathbf{R}_1$ и вектор регрессионных остатков $\boldsymbol{\varepsilon}$.

**Определение множителя $q$.** Рассмотрим тест Фишера вида

$$F = \frac{k_2^{-1}(\hat{\boldsymbol{\varepsilon}}_2^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_2 - \hat{\boldsymbol{\varepsilon}}_1^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_1)}{(m-k_1-k_2)^{-1}(\hat{\boldsymbol{\varepsilon}}_1^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_1)}.$$

Нуль-гипотеза — принятие модели$f_1$ может быть отвергнута при $F > c$, где $c$ — некоторый заданный процентиль (как правило, 95-й) $F$-распределения со степенями свободы $k_2$ и $m - k_1 - k_2$. Перепишем последнее выражение как

$$(m-k_1-k_2)\hat{\boldsymbol{\varepsilon}}_2^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_2 > (m-k_1-ck_2)(\hat{\boldsymbol{\varepsilon}}_1^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_1)$$

при этом множитель $q$ определяется выражением

$$\frac{m-k_1-k_2+ck_2}{m-k_1-k_2} = 1 + \frac{ck_2}{m-k_1-k_1}, \tag{99}$$

значение которого больше единицы при $k_2 \neq 0$. Заметим, что если $F$-тест может быть представлен в виде (98), то другие нижеперечисленные критерии могут быть представлены посредством $F$-тестов с критическими значениями, отличными от $c$.

Рассмотрим другой критерий выбора моделей — скорректированный коэффициент детерминации

$$R^2_{\text{adj}} = 1 - \frac{m-1}{m-k}\frac{\hat{\boldsymbol{\varepsilon}}^\mathsf{T}\hat{\boldsymbol{\varepsilon}}}{\mathbf{y}^\mathsf{T}\mathbf{y}}.$$

Используя этот критерий при выборе первой из двух моделей (88) и (89) получим

$$1 - \frac{m-1}{m-k_1-k_2}\frac{\hat{\boldsymbol{\varepsilon}}_1^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_1}{\mathbf{y}^\mathsf{T}\mathbf{y}} > 1 - \frac{m-1}{m-k_1}\frac{\hat{\boldsymbol{\varepsilon}}_2^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_2}{\mathbf{y}^\mathsf{T}\mathbf{y}},$$

откуда следует

$$\frac{m-1}{m-k_1}\hat{\boldsymbol{\varepsilon}}_2^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_2 > \frac{m-1}{m-k_1-k_2}\hat{\boldsymbol{\varepsilon}}_1^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_1.$$

Множитель $q$ для этого критерия определен как

$$\frac{m-k}{m-k_1-k_2} = 1 + \frac{k_2}{m-k_1-k_2}. \tag{100}$$

Так как введенное ранее пороговое значение $> 1$, то значение $q$, полученное из формулы (99) больше значения $q$ из формулы (100).

Критерий предсказательной способности, введенный в [19] предполагает принятие гипотезы $f_1$ при

$$\frac{m+k_1}{m-k_1}\hat{\boldsymbol{\varepsilon}}_2^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_2 > \frac{m+k_1=k_2}{m-k_1-k_2}\hat{\boldsymbol{\varepsilon}}_1^\mathsf{T}\hat{\boldsymbol{\varepsilon}}_1.$$

При этом множитель $q$ задан отношением

$$\frac{(m+k_1+k_2)(m-k_1)}{(m-k_1-k_2)(m+k_1)} = 1 + \frac{2mk_2}{(m-k_1-k_2)(m+k_1)}.$$

Согласно информационному критерию Акаике [14] гипотезе $f_1$ отдается предпочтение при

$$\exp\left(\frac{2k_1}{m}\right)\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 > \exp\left(\frac{2(k_1+k_2)}{m}\right)\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1.$$

При этом множителя $q$ задан как

$$\exp\left(\frac{2k_2}{m}\right).$$

Заметим, что в данном случае значение множителя не зависит от числа признаков первой модели $k_1$. При $k_2 \neq 0$ значение определенного в последнем выражении множителя $q > 1$.

В работе [247] введен следующий критерий предпочтения гипотезы $f_1$

$$\left(1+\frac{2k_1}{m}\right)\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 > \exp\left(1+\frac{2(k_1+k_2)}{m}\right)\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1,$$

множитель $q$ для этого критерия определен выражением

$$\frac{m+2(k_1+k_2)}{m+2k_1}1 + \frac{2k_1}{m+2k_1}.$$

Для оценки необходимого числа признаков, в работе [241] был предложен критерий, штрафующий модель за чрезмерное количество параметров. При этом модель $f_2$ отвергается как гипотеза только тогда, когда

$$m^{\frac{k_1}{m}}\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 > m^{\frac{k_1+k_2}{m}}\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1.$$

значение множителя при этом определено выражением

$$m^{\frac{k_2}{m}},$$

который, как и критерий Акаике, не зависит от значения $k_1$. Если разложить предыдущее выражение в ряд Тейлора,

$$1 + \frac{k_2 \log m}{m} + \frac{1}{2}\frac{k_2^2 \log m}{m^2} + \ldots,$$

то можно увидеть, что значение $q > 1$.

Критерий обобщенного скользящего контроля был предложен в работе [73]. Согласно ему, модель $f_1$ более предпочтительна, если

$$\left(\frac{m-k_1}{m-k_1}\right)^2\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 > \left(\frac{m-k_1}{m-k_1-k_2}\right)^2\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1.$$

В этом случае множитель $q$ имеет вид

$$\left(\frac{m-k_1}{m-k_1}\right)^2\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 = 1 + \frac{2k_2}{m-k_1-k_2} + \frac{k_2^2}{m-k_1-k_2}.$$

Согласно критерию Райса [234]. Модель $f_1$ более предпочтительна, если

$$\frac{1}{1-(2k_1m^{-1})}\hat{\boldsymbol{\varepsilon}}_2^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_2 > \frac{1}{1-\left((2k_1+2k_2)m^{-1}\right)}\hat{\boldsymbol{\varepsilon}}_1^{\mathsf{T}}\hat{\boldsymbol{\varepsilon}}_1,$$

при это множитель $q$ задан выражением

$$\frac{m}{m-k_1-k_2}\frac{m-2k_1}{m} = 1 + \frac{2k_2}{m-2(k_1+k_2)}.$$

Таким образом, известные критерии качества могут быть выражены с помощью введенного расстояния между моделями с использованием множителя $q$.

## 3.3. Выбор признаков при последовательном порождении моделей

В данной работе предлагается компромиссный вариант алгоритма выбора регрессионных моделей. Его целью является получение наиболее адекватной модели, включающей наименее мультикоррелирующие признаки. Он заключается в последовательном порождении наиболее правдоподобных моделей и основан на работах по символьной регрессии и байесовскому подходу к оценке параметров моделей [273, 47, 48, 285]. При этом значения значения правдоподобия различных моделей сравниваются. В ходе порождения модели модифицируются таким образом, что при добавлении признаков увеличивается правдоподобие модели, а при удалении признаков уменьшается число мультикоррелирующих признаков. При удалении признаков статистическая зависимость между параметрами модели не увеличивается.

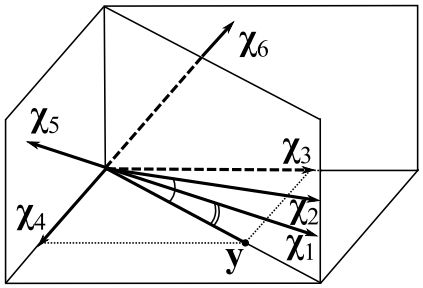



Рис. 35. Проблема фильтрации шумовых и мультикоррелирующих признаков.

Рис. 35 иллюстрирует основные проблемы, возникающие при выборе признаков. Вектор $\mathbf{y}$ зависимых переменных лежит в пространстве, натянутом на векторы-столбцы $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n$ матрицы плана $X$. Пусть эти векторы расположены так, что вектор $\mathbf{y}$ образует наименьший угол к вектору $\boldsymbol{\chi}_1$, а вектор $\boldsymbol{\chi}_1$ образует малый угол с вектором $\boldsymbol{\chi}_2$. При этом вектор $\mathbf{y}$ лежит в одной плоскости с векторами $\boldsymbol{\chi}_3$ и $\boldsymbol{\chi}_4$, причем эти два вектора ортогональны. Векторы $\boldsymbol{\chi}_5, \boldsymbol{\chi}_6$ «почти» ортогональны вектору $\mathbf{y}$. Иначе говоря, векторы $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ считаются мультикоррелирующими, а векторы $\boldsymbol{\chi}_5, \boldsymbol{\chi}_6$ считаются шумовыми.

Предположим, что модель линейна и имеет не более двух параметров. Тогда алгоритм последовательного добавления признаков, а также ряд других алгоритмов выберут векторы $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$, получив таким образом не самое точное приближение и, в терминах возможного значительного изменения параметров при незначительном изменении данных, не самую устойчивую модель. Очевидно, что оптимальным решением для этой иллюстрации является выбор признаков $\boldsymbol{\chi}_3, \boldsymbol{\chi}_4$, так как они дают точное приближение и устойчивую модель. Ни один из вышеперечисленных алгоритмов не дает такого решения. Во-первых, потому, что он не содержит критерии, позволяющие выявить наличие мультикоррелирующих признаков, и, как следствие, устойчивость модели к небольшим изменениям выборки.

Предположим, что к некоторой модели линейной регрессии

$$\mathsf{E}(\mathbf{y}|\mathbf{X}_1) = \mathbf{X}_1\mathbf{w}_1, \quad \mathsf{D}(\mathbf{y}) = \sigma^2\mathbf{I}_m, \quad \mathsf{D}(\mathbf{w}_1) = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1},$$

добавляются новые признаки $\{\boldsymbol{\chi}_{n_1+1}, \ldots, \boldsymbol{\chi}_{n_1+n_2}\}$. При этом модель принимает вид

$$\mathsf{E}(\mathbf{y}|\mathbf{X}) = \underset{m\times n_1}{\mathbf{X}_1}\ \mathbf{w}_1 + \underset{m\times n_2}{\mathbf{X}_2}\ \mathbf{w}_2 = [\mathbf{X}_1, \mathbf{X}_1]\begin{bmatrix}\mathbf{w}_1\\ \mathbf{w}_2\end{bmatrix} = \underset{m\times(n_1+n_2)}{\mathbf{X}_3}\ \mathbf{w}_3.$$

Пусть матрица $\mathbf{X}_3$, полученная в результате соединения матриц $\mathbf{X}_1, \mathbf{X}_2$, имеет ранг $n_1 + n_2$, равный числу ее столбцов. Оценку $\hat{\mathbf{w}}_3$ вектора параметров $\mathbf{w}_3$ можно получить двумя путями: непосредственно вычислить ее методом наименьших квадратов или воспользоваться теоремой о дополнительных регрессорах [340].

Пусть

$$\begin{aligned}
\underset{m\times m}{\mathbf{P}} &= \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal,\\
\underset{m\times m}{\mathbf{R}} &= \mathbf{I}_m - \mathbf{X}_3(\mathbf{X}_3^\intercal\mathbf{X}_3)^{-1}\mathbf{X}_3^\intercal,\\
\underset{n_1\times n_2}{\mathbf{L}} &= (\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal\mathbf{X}_2,\\
\underset{n_2\times n_2}{\mathbf{M}} &= (\mathbf{X}_2^\intercal\mathbf{R}\mathbf{X}_2)^{-1}
\end{aligned}$$

и

$$\underset{(n_1+n_2)\times 1}{\hat{\mathbf{w}}_3} = \begin{bmatrix} \hat{\mathbf{w}}_1 \\ \hat{\mathbf{w}}_2 \end{bmatrix};$$

тогда:

1) $\hat{\mathbf{w}}_1 = (\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}\mathbf{X}_1^\intercal(\mathbf{y} - \mathbf{X}_2\hat{\mathbf{w}}_2) = \hat{\mathbf{w}}_1 - \mathbf{L}\hat{\mathbf{w}}_2$,

2) $\hat{\mathbf{w}}_2 = (\mathbf{X}_2^\intercal\mathbf{P}\mathbf{X}_2)^{-1}\mathbf{X}_2\mathbf{P}\mathbf{y}$,

3) $\mathbf{y}^\intercal\mathbf{R}\mathbf{y} = (\mathbf{y} - \mathbf{X}_2\hat{\mathbf{w}}_2)^\intercal\mathbf{P}(\mathbf{y} - \mathbf{X}_2\hat{\mathbf{w}}_2)$,

4) $\mathbf{y}^\intercal\mathbf{R}\mathbf{y} = \mathbf{y}^\intercal\mathbf{P}\mathbf{y} - \hat{\mathbf{w}}_2^\intercal\mathbf{X}_2^\intercal\mathbf{P}\mathbf{y}$,

5) $\mathsf{D}(\hat{\mathbf{w}}_3) = \begin{pmatrix} (\mathbf{X}_1^\intercal\mathbf{X}_1)^\intercal + \mathbf{L}\mathbf{M}\mathbf{L}^\intercal, & -\mathbf{L}\mathbf{M} \\ -\mathbf{M}\mathbf{L}^\intercal, & \mathbf{M} \end{pmatrix}$.

Таким образом получена оценка $\hat{\mathbf{w}}_3$ без обращения соединенной матрицы. Получение оценки путем пересчета части системы линейных уравнений снижает время работы алгоритма выбора признаков, особенно в случае, когда число $n_1$ выбранных столбцов велико, а число $n_2$ столбцов, которые рассматриваются в качестве претендентов на добавление, мало.

### 3.3.1. Процедура последовательного выбора признаков

В процедуре шаговой регрессии не учитывается то, что добавление или удаление признака может существенно изменить значения параметров $\mathbf{w}$ остальных признаков $[\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_r]$, заданных набором $\mathcal{A}$. Ранее вошедший в набор признак может перестать быть значимым после добавления других признаков. Поэтому предполагая, что дисперсии признаков меняются с изменением набора $\mathcal{A}$ незначительно, предложим модификацию шаговой процедуры, использующую альтернативные критерии добавления или выбора признаков и выбора оптимальной модели. Эта модификация позволяет решить ряд проблем получения устойчивой и точной модели, описанных в [84, 198, 196, 183, 215].

Предлагаемый алгоритм использует счетное число порождаемых признаков при отыскании линейной регрессионной модели. Введем процедуру пошагового нахождения модели. На каждом шаге выполняются операции добавления признаков и прореживания признаков. Под сложностью понимается число элементов линейной комбинации.

Используем два набора признаков: порожденный набор $\mathcal{Z}$ и текущий набор $\mathcal{C}$. В начале работы алгоритма $\mathcal{C} = \emptyset$.

Рассмотрим $k$-й шаг алгоритма.

1. Последовательно, методом Add, добавляются признаки из объединенного набора $\mathcal{C} \cup \mathcal{Z}$ в активный набор признаков $\mathcal{A}_k$. Итерации повторяются до тех пор, пока при увеличении сложности модели правдоподобие модели не будет меньше заданного порогового $\mathcal{E}_{min}$.

2. Выполняется прореживание модели: последовательно удаляются те элементы линейной комбинации, заданной набором $\mathcal{A}_k$, для которых критерий мультиколлинеарности Белсли [33] принимает максимальное значение. Прореживание модели продолжается до тех пор пока, при уменьшении сложности модели, правдоподобие не будет меньше порогового $\mathcal{E}_{min}$. Коэффициенты полученной модели пересчитываются.

Итерации повторяются согласно критерию правдоподобия моделей. В результате получаем активный набор признаков $\mathcal{A}_k$, который на следующей итерации используется в качестве текущего набора $\mathcal{C}$.

Пороговое правдоподобие вычисляется следующим образом. Обозначим $p(\boldsymbol{\beta}) \stackrel{\text{def}}{=} p(\boldsymbol{\beta}|f)$ — априорное распределение параметров модели. Рассмотрим функцию правдоподобия $p(\mathfrak{D}|\boldsymbol{\beta}, f) \stackrel{\text{def}}{=} p(y|\{x_j\}_{j=1}^n, \boldsymbol{\beta}, f)$ — условную плотность распределения случайной величины при заданном векторе параметров.

При отыскании вектора параметров вместо максимизации функции правдоподобия $p(\mathfrak{D}|\boldsymbol{\beta}, f)$ будем максимизировать апостериорное распределение параметров

$$p(\boldsymbol{\beta}|\mathfrak{D}, f) = \frac{p(\mathfrak{D}|\boldsymbol{\beta}, f)p(\boldsymbol{\beta}|f)}{p(\mathfrak{D}|f)}. \tag{101}$$

Знаменатель $p(\mathfrak{D}|f)$ есть интеграл числителя формулы Байеса по всему пространству параметров:

$$p(\mathfrak{D}|f) = \int p(\mathfrak{D}|\boldsymbol{\beta}, f)p(\boldsymbol{\beta}|f)d\boldsymbol{\beta}. \tag{102}$$

Пусть зависимая переменная распределена нормально. Тогда функция правдоподобия принимает вид

$$p(\mathfrak{D}|\boldsymbol{\beta}, f) = \prod_{i=1}^{m} \mathcal{N}(y^i|f(\mathbf{x}^i, \boldsymbol{\beta}), \sigma_\nu^{-2}), \tag{103}$$

где $\sigma_\nu^2$ — дисперсия случайной величины $\nu$.

Пусть многомерная случайная величина — вектор параметров модели также имеет нормальное распределение с нулевым матожиданием и ковариационной матрицей

$$\alpha\mathbf{I} = \frac{1}{\sigma_{\boldsymbol{\beta}}^2}\mathbf{I}.$$

Тогда распределение вектора параметров модели

$$p(\boldsymbol{\beta}|\alpha, f) = \left(\frac{\alpha}{2\pi}\right)^n \exp\left(-\frac{\alpha}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right). \tag{104}$$

Полученный знаменатель формулы (101) называется правдоподобием модели и служит для сравнения моделей.

Сравнение моделей выполняется с помощью связанного Байесовского вывода. Обозначим распределение моделей при фиксированных данных $p(f_i|\mathfrak{D})$ и рассмотрим числитель формулы Байеса

$$p(f_i|\mathfrak{D}) = \frac{p(\mathfrak{D}|f_i)p(f_i)}{p(\mathfrak{D})}, \tag{105}$$

в котором правдоподобие моделей $p(\mathfrak{D}|f_i)$ определяется выражением (102). Будем считать априорную вероятность равной для всех моделей, $p(f_i) = p(f_j)$. Так как знаменатель выражения (105) не зависит от выбора модели, то сравнение моделей происходит через вычисление правдоподобие моделей с помощью формул (103) и (104). Порог $\mathcal{E}_{min}$ вычисляется как $\min_{i=1,\dots,M} p(\mathfrak{D}|f_i)$ для набора из $M$ моделей, имеющих максимальное правдоподобие; параметр $M$ задан.

Результатом работы алгоритма является модель удовлетворительной точности; мультикоррелирующие признаки исключены.

### 3.3.2. Выбор признаков в условиях мультикорреляции

Исследуется проблема оптимальной сложности модели в связи с ее точностью и устойчивостью. Задача состоит в нахождении наиболее информативного набора признаков в условиях их высокой мультиколлинеарности. Для выбора оптимальной модели используется модифицированный алгоритм шаговой регрессии, являющийся одним из алгоритмов добавления [66, 65] и удаления [32] признаков. Для описания работы пошагового алгоритма предложена модель $n$-мерного куба. Проанализированы величины матожидания и дисперсии функции ошибки.

Решается задача восстановления линейной регрессии при наличии большого числа мультиколлинеарных признаков [265]. Термин «мультиколлинеарность» введен Р. Фишером при рассмотрении линейных зависимостей между признаками [98]. Проблема состоит в том, что количество признаков значительно превосходит число зависимых переменных, то есть мы имеем дело с переопределенной матрицей. Для решения этот задачи необходимо исключить наиболее малоинформативные признаки. Для отбора признаков предлагается использовать модифицированный метод шаговой регрессии.

Ранее для решения подобных задач использовались следующие методы: метод наименьших углов LARS [82], Лассо [267], ступенчатая регрессия [80], последовательное добавление признаков с ортогонализацией FOS [66, 65], шаговая регрессия [83, 233] и другие.

**Задача выбора оптимальной модели.** Опишем, в чем состоит задача выбора оптимальной модели. Задана выборка $\mathfrak{D} = (\{\mathbf{x}_i, y_i\}), i \in \mathcal{I}$, где множество свободных переменных — вектор $\mathbf{x} = [x_1, \dots, x_j, \dots, x_n]$, проиндексировано $j \in \mathcal{J} = \{1, \dots, n\}$. Задано разбиение множества индексов элементов выборки $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$. Также задан класс линейных параметрических регрессионных моделей $f(\mathbf{w}, \mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$ — параметрических функций, линейных относительно параметров. Функция ошибки задана следующим образом

$$S = \sum_{i \in \mathcal{X}} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2, \tag{106}$$

где $\mathcal{X} \subseteq \mathcal{I}$ —некоторое множество индексов. Требуется найти такое подмножество индексов $\mathcal{A} \subseteq \mathcal{J}$, которое бы доставляло минимум функции

$$\mathcal{A}^* = \arg\min_{\mathcal{A}\subseteq\mathcal{J}} S(f_{\mathcal{A}}|\mathbf{w}^*, \mathfrak{D}_{\mathcal{C}}) \tag{107}$$

на множестве индексов $\mathcal{C}$. При этом параметры $\mathbf{w}^*$ модели должны доставлять минимум функции

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\mathfrak{D}_{\mathcal{L}}, f_{\mathcal{A}}) \tag{108}$$

на множестве индексов $\mathcal{L}$. Здесь $f_{\mathcal{A}}$ обозначает модель $f$, включающую только столбцы матрицы $\mathbf{X}$ с индексами из множества $\mathcal{A}$, а обозначение вида $S(\mathbf{w}|\mathfrak{D})$ означает, что переменная $\mathfrak{D}$ фиксирована, а переменная $\mathbf{w}$ изменяется.

**Процедура выбора оптимального набора признаков.** Процедура выбора оптимального набора признаков состоит из этапов добавления и удаления. На первом этапе последовательно добавляются признаки, доставляющие максимум правдоподобия модели. На втором этапе происходит последовательное удаление признаков с целью увеличения устойчивости модели, в случае обобщенно-линейных моделей уменьшения мультиколлинеарности признаков.

Пусть на $k$-ом шаге алгоритма имеется активный набор признаков $\mathcal{A}_k \in \mathcal{J}$. На нулевом шаге $\mathcal{A}_0$ пуст.

Этап добавления. Находим признак доставляющий максимум $p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, f_{\mathcal{A}_{k-1}})$ на обучающей выборке

$$j^* = \arg\max_{j\in\mathcal{J}\setminus\mathcal{A}_{k-1}} p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, f_{\mathcal{A}_{k-1}\cup\{j\}}).$$

Затем добавляем новый признак $j^*$ к текущему активному набору

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \{j^*\}$$

и повторяем эту процедуру до тех пор, пока $p(\mathfrak{D}|\mathbf{A}, \mathbf{B}, f_{\mathcal{A}_k})$менее своего максимальноне значение на данном этапе не более, чем на некоторое заданное значение $\Delta$.

Этап удаления. Находим индексы обусловленности и долевые коэффициенты для текущего набора признаков $\mathcal{A}_{k-1}$. Находим количество достаточно больших индексов обусловленности. Достаточно большими считаются индексы, квадрат которых превосходит максимальный индекс обусловленности $\eta_t$, где $t = |A_{k-1}|$ количество признаков в текущем наборе $\mathcal{A}_{k-1}$.

$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right]. \tag{109}$$

Находим в матрице долевых коэффициентов $\mathbf{var}(\mathbf{w})$ столбец $j^*$ с максимальной суммой по последним $i^*$ долевым коэффициентам

$$j^* = \arg\max_{j\in\mathcal{A}_{k-1}} \sum_{g=t-i^*+1}^{t} q_g^j. \tag{110}$$

Удаляем $j^*$-ый признак из текущего набора

$$\mathcal{A}_k = \mathcal{A}_{k-1}\setminus j^*$$

и повторяем эту процедуру до тех пор, пока $p(\mathfrak{D}|\mathbf{A},\mathbf{B},f_{\mathcal{A}_k})$ менее своего максимального значения на данном этапе не более, чем на некоторое заданное значение $\Delta$. Повторение этапов добавления и удаления осуществляется до тех пор, пока значение $p(\mathfrak{D}|\mathbf{A},\mathbf{B},f_{\mathcal{A}_k})$ не стабилизируется.

**Удаление признаков.** Рассмотрим матрицу признаков $\mathbf{X}$. Она имеет размерность $m \times n$. Выполним ее сингулярное разложение:

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal,$$

где $\mathbf{U}$, $\mathbf{V}$ — ортогональные матрицы размерностью соответственно $m \times m$ и $n \times n$ и $\mathbf{\Lambda}$ — диагональная матрица с элементами (сингулярными числами) на диагонали такими, что

$$\lambda_1 > \lambda_2 > \ldots > \lambda_r,$$

где $r$ — ранг матрицы $\mathbf{X}$. Заметим, что в нашем случае $r = n$. Это связано с тем, что в алгоритме шагового выбора на каждом шаге мы имеем мультиколлинеарный набор признаков. Столбцы матрицы $\mathbf{V}$ являются собственными векторами, а квадраты сингулярных чисел — собственными значениями матрицы $\mathbf{X}^\intercal\mathbf{X}$.

$$\mathbf{X}^\intercal\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{U}^\intercal\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\intercal,$$

$$\mathbf{X}^\intercal\mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2.$$

Отношение максимального сингулярного числа к $j$-му сингулярному числу назовем индексом обусловленности с номером $j$

$$\eta_j = \frac{\lambda_{\max}}{\lambda_j}.$$

Если матрица $X$ не полного ранга, то значительная часть индексов обусловленности не определена. В нашем случае, как говорилось выше, матрица признаков $X$ является матрицей полного ранга.

Так как модель линейна, то $\mathbf{w} = \mathbf{B}\mathbf{y}$, где $\mathbf{w}$ — вектор параметров модели. То есть $w_i = \mathbf{b}_i^\intercal\mathbf{y}$, где

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_1^\intercal \\ \ldots \\ \mathbf{b}_n^\intercal \end{pmatrix}.$$

Мы ищем несмещенную оценку параметров

$$\mathrm{E}(\mathbf{w}) = \mathbf{w} = \mathbf{B}\mathbf{X}\mathbf{w},$$

то есть $\mathbf{B}\mathbf{X} = \mathbf{I}$, где $\mathbf{I}$ — единичная матрица.

Тогда ковариация параметров $w_i$ и $w_j$ равна

$$cov(w_i, w_j) = \mathrm{E}(\mathbf{b}_i^\intercal\mathbf{y} - \mathbf{b}_i^\intercal\mathbf{X}\mathbf{w})(\mathbf{b}_j^\intercal\mathbf{y} - \mathbf{b}_j^\intercal\mathbf{X}\mathbf{w}) = \mathbf{b}_i^\intercal\mathrm{E}((\mathbf{y} - \mathbf{X}\mathbf{w})(\mathbf{y} - \mathbf{X}\mathbf{w})^\intercal))\mathbf{b}_j =$$

$$= \mathrm{E}(\xi_i\xi_j^\intercal)\mathbf{b}_i^\intercal\mathbf{b}_j = \sigma^2\mathbf{b}_i^\intercal\mathbf{b}_j,$$

где $\xi_i$ — $i$-ый регрессионный остаток, а $\sigma^2$ — дисперсия регрессионных остатков.

Мы хотим найти несмещенную оценку параметров, минимизирующую дисперсию параметров по каждой компоненте

$$\begin{cases} \sigma^2 \mathbf{b}_i^\intercal \mathbf{b}_i \longrightarrow \min\limits_{\mathbf{B}} \\ \mathbf{b}_i^\intercal \mathbf{X} = \mathbf{e}_i^\intercal \end{cases},$$

где $\mathbf{e}_i^\intercal$ — $i$-ая строка единичной матрицы. Составим фунцию Лагранжа

$$L = \mathbf{b}_i^\intercal \mathbf{b}_i + \mathbf{\Lambda}_i^\intercal (\mathbf{X}^\intercal \mathbf{b}_i - \mathbf{e}_i),$$

где $\mathbf{\Lambda} = (\mathbf{\Lambda}_1 \dots \mathbf{\Lambda}_n)$. Продифференцировав по $\mathbf{b}_i$, получим

$$\begin{cases} 2\mathbf{b}_i + \mathbf{X}\mathbf{\Lambda}_i \\ \mathbf{X}^\intercal \mathbf{b}_i - \mathbf{e}_i = 0 \end{cases}$$

Из первого уравнения $\mathbf{b}_i = -\frac{1}{2}\mathbf{X}\mathbf{\Lambda}_i$, тогда $-\frac{1}{2}\mathbf{X}^\intercal\mathbf{X}\mathbf{\Lambda}_i = \mathbf{e}_i$. То есть $\mathbf{\Lambda} = -2(\mathbf{X}^\intercal\mathbf{X})^{-1}$, и, окончательно, для $\mathbf{B}$ получим

$$\mathbf{B} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal.$$

Для ковариационной матрицы $\mathbf{A}$ получим

$$\mathbf{A} = \sigma^2\mathbf{B}\mathbf{B}^\intercal = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{X}((\mathbf{X}^\intercal\mathbf{X})^{-1})^\intercal = \sigma^2\mathbf{X}^{-1}(\mathbf{X}^\intercal)^{-1}\mathbf{X}^\intercal\mathbf{X}((\mathbf{X}^\intercal\mathbf{X})^{-1})^\intercal =$$

$$= \sigma^2((\mathbf{X}^\intercal\mathbf{X})^{-1})^\intercal = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}.$$

Выражение $\sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$ является несмещенной оценкой ковариационной матрицы признаков, а в случае линейной модели оно в точности совпадает с ковариационной матрицей, то есть $\mathbf{A}^{-1} = \sigma^{-2}\mathbf{X}^\intercal\mathbf{X}$.

Используя сингулярное разложение, дисперсия параметров, найденных методом наименьших квадратов $\mathbf{w} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, может быть записана как

$$\mathbf{var}(\mathbf{w}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{V}^\intercal)^{-1}\mathbf{\Lambda}^{-2}\mathbf{V}^{-1} = \sigma^2\mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^T.$$

Таким образом, дисперсия $j$-го регрессионного коэффициента — это $j$-й диагональный элемент матрицы $\mathbf{var}(\mathbf{w})$.

Для обнаружения мультиколлинеарности признаков построим таблицу, в которой каждому индексу обусловленности $\eta_j$ соответствуют значения $q_{ij}$ — долевые коэффициенты. Сумма долевых коэффициентов по индексу $j$ равна единице.

$$\sigma^{-2}\mathbf{var}(w_i) = \sum_{j=1}^{n} \frac{\upsilon_{ij}^2}{\lambda_j^2} = (q_{i1} + q_{i2} + \ldots + q_{in}) \sum_{j=1}^{n} \frac{\upsilon_{ij}^2}{\lambda_j^2},$$

где $q_{ij}$ — отношение соответствующего слагаемого в разложении вектора $\sigma^{-2}\mathbf{var}(w_i)$ ко всей сумме, а $\mathbf{V} = (\upsilon_{ij})$. Чем больше значение долевого коэффициента $q_{ij}$ тем больший вклад вносит $j$-ый признак в дисперсию $i$-го регрессионного коэффициента.

Из таблицы (9) определяется мультиколлинеарность: большие величины $\eta_j$ означают, что, возможно, есть зависимость между признаками. Если присутствует только один достаточно большой индекс обусловленности, тогда возможно определение участвующих в зависимости

Таблица 9. Долевые коэффициенты.

| Индекс обусловленности | $\mathbf{var}(w_1)$ | $\mathbf{var}(w_2)$ | ... | $\mathbf{var}(w_n)$ |
|---|---|---|---|---|
| $\eta_1$ | $q_{11}$ | $q_{21}$ | $\cdots$ | $q_{n1}$ |
| $\eta_2$ | $q_{12}$ | $q_{22}$ | $\cdots$ | $q_{n2}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $\eta_n$ | $q_{1n}$ | $q_{2n}$ | $\cdots$ | $q_{nn}$ |

признаков из долевых коэффициентов: признак считается вовлеченным в зависимость, если его долевой коэффициент связанный с этим индексом превышает выбранный порог (обычно 0.25). Если же присутствует несколько больших индексов обусловленности, то вовлеченность признака в зависимость определяется по сумме его дисперсионных долей, отвечающих большим значениям индекса обусловленности: если сумма превышает выбранный порог, то признак участвует как минимум в одной линейной зависимости. Для нахождения мультиколлиниарных признаков решаются задачи (109) и (110).

Проиллюстрируем метод Белсли на примере. Используется неизменные признаки $x_1$, $x_5$ и зависящие от параметра $k$ признаки $x_2$, $x_3$, $x_4$. При $k = 0$ все признаки ортогональны, при увеличении $k$ признаки $x_2$, $x_3$ приближаются к $x_1$, а $x_4$ — к $x_5$ вплоть до полной коллинеарности при $k = 1$. На рис.36a-36d приведены матрицы долевых коэффициентов в зависимости от $k$.

В таблице (10) приведены значения индексов обусловленности в зависимости от $k$.

Таблица 10. Индексы обусловленности.

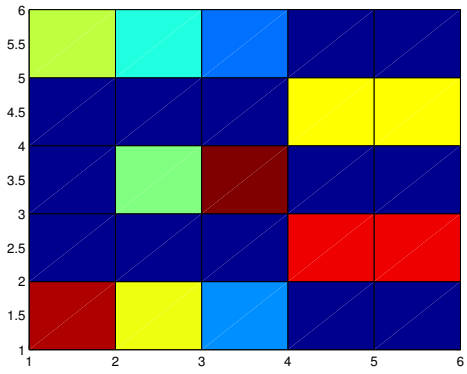
| $j, k$ | 0.15 | 0.25 | 0.4 | 0.9 |
|---|---|---|---|---|
| 1 | 1.0 | 1.0 | 1.0 | 1.0 |
| 2 | 1.0 | 1.0 | 1.1 | 1.2 |
| 3 | 1.1 | 1.2 | 1.5 | 21.5 |
| 4 | 1.2 | 1.4 | 2.0 | 22.1 |
| 5 | 1.2 | 1.5 | 2.1 | 24.0 |

Наблюдается две основных зависимости — первая между признаками $x_1$, $x_2$, $x_3$ и вторая между признаками $x_4$, $x_5$.

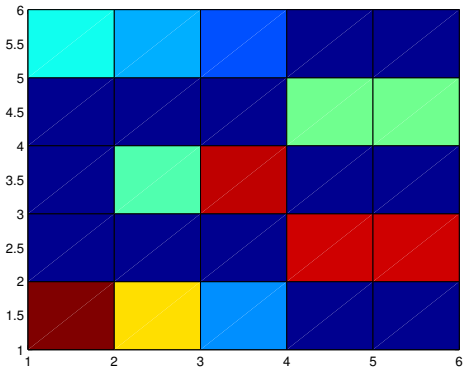
### 3.3.3. Оценка дисперсии функции ошибки

Функцию ошибки $S$ можно при фиксированном наборе признаков $\mathcal{A} \in \mathcal{J}$ считать случайной величиной. Мы хотим минимизировать ее математическое ожидание и дисперсию при фиксированной сложности модели.

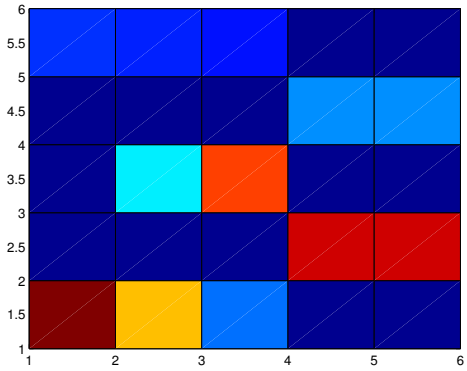
Для данного набора признаков $\mathcal{A} \in \mathcal{J}$ будем многократно разбивать выборку на обучение $\mathcal{L}$ и контроль $\mathcal{C}$. Полученные значения функции ошибки $S$ можно считать реализациями случайной величины. Тогда математическое ожидание и дисперсия оцениваются следующим



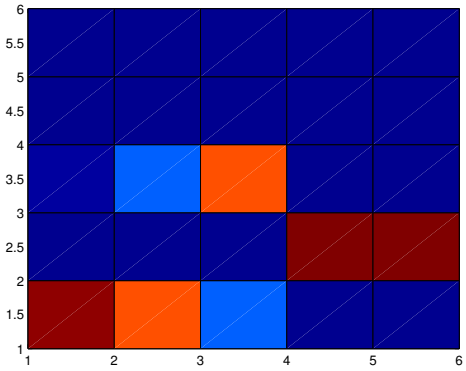

(a) $k = 0.15$




(b) $k = 0.25$




(c) $k = 0.4$




(d) $k = 0.9$

Рис. 36. Значения индексов обусловленности в зависимости от порога $k$.

образом

$$ES = \frac{1}{m}\sum_{i=1}^{m} S_i,$$

$$DS = \frac{1}{m}\sum_{i=1}^{m} (S_i - ES)^2,$$

где $m$ — число разбиений выборки, а $S_i$ — значение функции ошибки при $i$-ом разбиении.

Ниже представлены графики полученные по данным прогрессирования заболевания у больных диабетом. На нем отмечены все $2^n$ точек, где $n = 10$ — число признаков. По вертикали отложена дисперсия в логарифмическом маштабе, а по горизонтали количество признаков в наборе. При каждом значении числа признаков (сложности модели) найден набор с минимальным математическим ожиданием функции ошибки — эти точки отмечены красным.

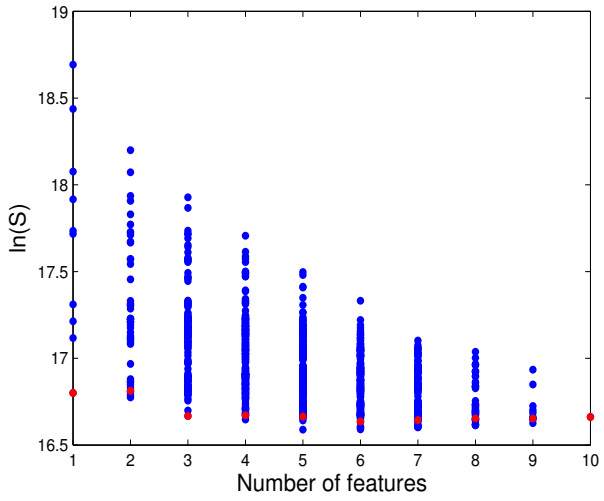



Рис. 37. Зависимость логарифма дисперсии функции ошибки от числа признаков при leave-one-out.

По графикам видно, что у наборов с малым математическим ожиданием функции ошибки дисперсия тоже мала.

**Путь в $n$-мерном кубе.** В нашей задаче мы имеем дело с $n$ признаками, то есть существует $2^n$ возможных наборов признаков, из которых мы пытаемся найти оптимальный. Все эти $2^n$ наборов можно представить как вершины $n$-мерного куба. В данной работе используется шаговый алгоритм поиска оптимального набора, то есть пошагового движения по вершинам этого куба.

Приведем пример движения по вершинам куба при работе предложенного алгоритма. Всего использовалось 6 признаков $x_1, \ldots, x_6$, они изображены на рис. 35. Также на нем показан вектор ответов $y$.

На рис. 39 показан путь по вершинам куба для описанных данных. По вертикали отложен номер признака, по горизонтали — номер итерации. Красная клетка означает, что

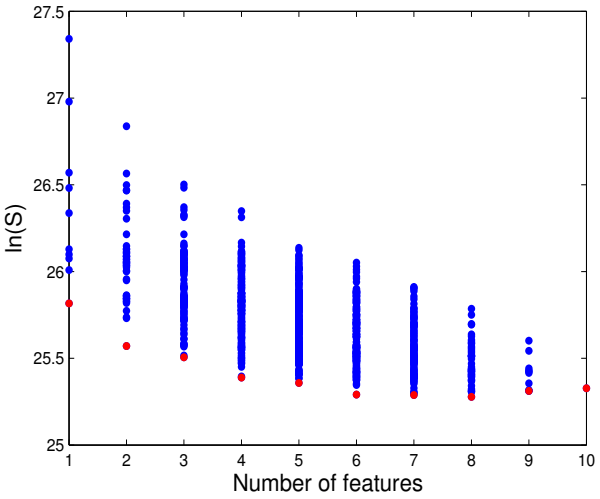



Рис. 38. Зависимость логарифма дисперсии функции ошибки от числа признаков при случайном разбиении выборки.

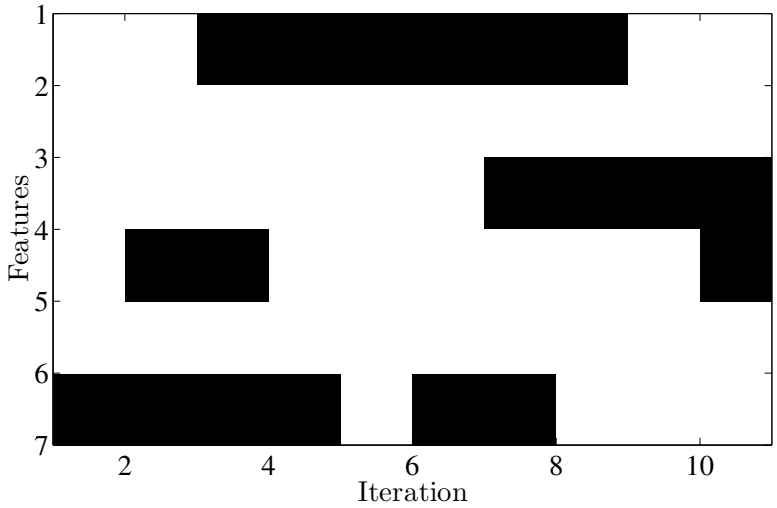



Рис. 39. Иллюстрация пути в кубе.

признак на данной итерации вошел в набор, синяя — не вошел. Например признак номер 6 присутствовал в наборе с 3 по 8 итерацию, но в конечный набор не вошел.

Предложен метод поиска оптимальной модели, основанный на комбинации двух стратегий: отбор признаков и выбор модели. Особенно полезен предложенный метод в случае, когда данные содержат большое число мультиколлинеарных признаков. Предложенный алгоритм позволяет получать хорошо обусловленные наборы порожденных признаков.

В работе теоретически обоснованно, что математическое ожидание функции ошибки $S$ достигает минимума при заданной сложности модели на том же наборе, на котором дисперсия достигает минимума. Этот результат так же подтвержден экспериментально на реальных данных.

## 3.4. Сравнение и анализ методов выбора признаков

Результаты сравнения алгоритмов приведены в табл. 11. Сравнение выполнялось на задаче поиска модели волатильности опционов. Использовались исторические данные торгов опционом Brent Crude Oil [358]. В таблицу входят значения функционала качества на обучающей и контрольной выборке, информационный критерий Акаике, число переменных модели. Исходя из значений критериев делается вывод об эффективности работы алгоритмов.

Таблица 11. Результаты работы алгоритмов выбора признаков.

| Алгоритм | $S_{\mathcal{L}}$ | $S_{\mathcal{C}}$ | AIC | BIC | $C_p$ | $\lg \kappa$ | $k$ |
|---|---|---|---|---|---|---|---|
| Генет. | 0,073 | 0,107 | -1152 | -1072 | 337 | 13 | 26 |
| МГУА | 0,146 | 0,194 | -1076 | -1045 | 745 | 6 | 10 |
| Шаг. рег. | 0,128 | 0,154 | -1092 | -1055 | 644 | 7 | 12 |
| Гребн. | 0,111 | 0,146 | -819 | -330 | 832 | 33 | 160 |
| Лассо | 0,121 | 0,147 | -1089 | -1034 | 611 | 5 | 18 |
| Ступ. | 0,071 | 0,096 | -1157 | -1077 | 324 | 9 | 26 |
| FOS | 0,106 | 0,135 | -1105 | -1044 | 527 | 7 | 20 |
| LARS | 0,098 | 0,095 | -1102 | -1017 | 492 | 7 | 28 |
| Предл. | 0,097 | 0,123 | -1118 | -1054 | 469 | 5 | 21 |

Для каждого алгоритма вычислены значения ошибок $S_{\mathcal{L}}$ и $S_{\mathcal{C}}$ на обучении и контроле (36), значение информационных критериев Акаике

$$\mathrm{AIC} = m(\ln \frac{S}{m}) + 2k,$$

Байеса

$$\mathrm{BIC} = m(\ln \frac{S}{m}) + k \ln m,$$

Маллоуза, десятичный логарифм числа обусловленности $\kappa$ матрицы значений отобранных признаков и сложность модели $k$.

На рис. 79 показана одна из полученных моделей. По оси $K$ отложена цена исполнения опциона, по оси $t$ отложено время до исполнения. Точками показаны исходные данные. Полученная модель является адекватной и удовлетворительно приближает исторические данные.

# 4. Выбор моделей

## 4.1. Связанный байесовский вывод при выборе моделей

*Связанный байесовский вывод* — метод сравнения регрессионных моделей, основанный на анализе свойств функций распределения параметров. Этот метод использует классический байесовский вывод дважды: для вычисления апостериорного распределения параметров модели и для вычисления апостериорной вероятности самой модели. Связанность заключается в том, что оба вывода используют общий сомножитель, называемый *правдоподобием модели*. Неотъемлемой частью этого метода является анализ пространства параметров модели и зависимости целевой функции от значений параметров. Результатом такого анализа является возможность оценить, насколько важны отдельные параметры модели для аппроксимации данных. Связанный байесовский вывод используется как в задачах регрессии, так и в задачах классификации.

### 4.1.1. Порождающие и разделяющие модели

Пересмотрим задачу восстановления регрессии (1),

$$\mathsf{E}(y|\mathbf{x}) = f(\mathbf{w}, \mathbf{x}),$$

поставленную в первом разделе, следующим образом. Согласно гипотезе порождения данных (6),

$$y \sim \mathcal{N}(f, \beta^{-1}),$$

и условиям (37), математическое ожидание случайной величины $y$, которое находится при восстановлении регрессии, зависит от неслучайной величины $\mathbf{x}$. В задачах регрессии считается, что величина $y$ лежит на оси действительных чисел, $y \in \mathbb{R}$. В задачах классификации считается, что величина $y$ принадлежит конечному множеству меток классов, например, $y \in \{0, 1\}$. Для предсказания значения случайной величины $y$ при новом значении $\mathbf{x}$ строится параметрическая модель $f(\mathbf{w}, \mathbf{x})$, параметры $\mathbf{w}$ которой оцениваются по обучающей выборке $\mathfrak{D} = \{y_i, x_i\}, i \in \mathcal{I}$.

В случаях, когда процедура оценивания параметров модели вместе с восстановлением математического ожидания $\mathsf{E}(y|\mathbf{x})$ также включает и восстановление условной плотности распределения $p(y|\mathbf{x})$, регрессионная модель называется *разделяющей* (англ. discriminative [47, 269, 48]). Восстановленная непрерывная плотность распределения используется для предсказания значений зависимой переменной $y$ при новых значениях независимой переменной $\mathbf{x}$. В качестве примеров разделяющих моделей приведем модели линейной и логистической регрессии, функции радиального базиса, нейронные сети или машины опорных векторов [29, 46].

Альтернативный подход к решению задачи восстановления регрессии заключается в восстановлении плотности совместного распределения $p(y, \mathbf{x})$, описанной, например, с помощью параметрической модели. Данное распределение используется для оценки параметров плотности условного распределения $p(y|\mathbf{x})$ с целью предсказания значения зависимой переменной $y$ для новых значений независимой переменной $\mathbf{x}$. Этот подход называется *порожда-*

*ющим*, (англ. generative), так как с помощью восстановленного совместного распределения $p(y, \mathbf{x})$ можно породить значения переменной $y$, вектора $\mathbf{x}$ или пары $(\mathbf{x}, y)$ в зависимости от гипотезы порождения данных. В качестве примеров порождающих моделей приведем гауссовские смеси моделей и скрытые марковские модели [92, 89].

На практике обобщающая способность порождающих моделей зачастую хуже чем у разделяющих из-за разницы между распределением, задаваемым моделью и реальным неизвестным распределением данных [48, 269]. Когда обучающая выборка велика, разделяющие техники широко используются, так как они дают хорошую обучающую способность. Однако сбор измеряемых данных, особенно подготовка значений зависимой переменной регрессионной выборки, в ряде практических приложений может стоить весьма дорого [325]. Поэтому в случаях, когда получить регрессионную выборку достаточной величины дорого, предлагается использовать порождающие методы [108, 250, 108].

Рассмотрим эвристическую процедуру, использующую комбинацию порождающих и разделяющих моделей для сочетания преимуществ обоих подходов. Порождающая модель может быть определена функцией плотности совместного распределения $p(\mathbf{x}, y|\boldsymbol{\theta})$ вектора независимых переменных $\mathbf{x}$ и зависимой переменной $y$, зависящей от набора параметров $\boldsymbol{\theta}$. Такая модель задается априорной вероятностью $p(f|\pi)$ вместе с внутриклассовым распределением вероятности для каждого класса $p(\mathbf{x}|y, \boldsymbol{\zeta})$ таким образом, что:

$$p(\mathbf{x}, y|\boldsymbol{\theta}) = p(y|\pi)p(\mathbf{x}|y, \boldsymbol{\zeta}),$$

где $\boldsymbol{\theta} = [\pi, \boldsymbol{\zeta}]^\intercal$. Так как, согласно методу наибольшего правдоподобия, элементы выборки рассматриваются как независимые случайные величины, функция плотности их совместного распределения определяется как

$$L_{\text{gn}}(\boldsymbol{\theta}) = p(X, \mathbf{y}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} p(\mathbf{x}_i, y_i|\boldsymbol{\theta}).$$

Для оценки вектора параметров $\boldsymbol{\theta}$ необходимо максимизировать функцию $L_{\text{gen}}$,

$$\hat{\boldsymbol{\theta}} = \arg\max \left( p(\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} p(\mathbf{x}_i, y_i|\boldsymbol{\theta}) \right).$$

Так как $p(\mathbf{X}, \mathbf{y}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})p(\mathbf{X}, \mathbf{y})$, данное условие эквивалентно максимизации функции плотности апостериорного распределения $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$.

В работе [52] утверждается, что при максимизации функции правдоподобия разделяющей модели

$$L_d s(\boldsymbol{\theta}) = p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} p(y_i|x_i, \boldsymbol{\theta}),$$

где

$$p(y|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}, y|\boldsymbol{\theta})}{\sum\limits_{i \in \mathcal{I}} p(\mathbf{x}_i, y_i|\boldsymbol{\theta})},$$

обобщающая способность разделяющей модели выше, чем порождающей.

### 4.1.2. Интегральная функция правдоподобия

Интегральная функция правдоподобия является функцией правдоподобия, в которой переменные-параметры имеют условное распределение.

Рассмотрим параметр $\boldsymbol{\theta} = [\mathbf{x}, \mathbf{w}]^\intercal$. Для этой пары интегральная функция правдоподобия имеет вид

$$L(\mathbf{x}; y) = p(y|\mathbf{x}) = \int_{\mathbf{w}\in\mathbb{W}} p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathbf{x})d\mathbf{w}.$$

Как взятие интеграла этой функции может стать вычислительно сложной задачей. В таком случае используются сэмплирующие методы, в первую очередь метод Монте-Карло или гиббсовские методы сэмплирования. В частных случаях используется аппроксимация Лапласа или EM-алгоритмы [52].

В байесовском сравнении моделей условные переменные являются параметрами моделей. Условное правдоподобие является вероятностью появления данных для некоторой фиксированной модели и не предполагает фиксацию параметров этой модели. Считая, как и ранее, вектор $\mathbf{w}$ параметрами модели, записываем интегральное правдоподобие модели в виде

$$p(y, f) = \int_{\mathbb{R}^n} p(\mathfrak{D}|\mathbf{w}, f)p(\mathbf{w}|f)d\mathbf{w}.$$

Для двух моделей отношение этих интегралов называется байесовским множителем:

$$\frac{p(f_1|\mathfrak{D})}{p(f_2|\mathfrak{D})} = \frac{p(f_1)}{p(f_2)}\,\frac{p(\mathfrak{D}|f_1)}{p(\mathfrak{D}|f_2)}.$$

### 4.1.3. Частотный и байесовский подход

Рассмотрим различия традиционного, частотного подхода и байесовского подхода к оценке параметров и к выбору разделяющих моделей. Пусть для некоторой модели $f(\mathbf{w}, \mathbf{x})$ по выборке $\mathfrak{D}$ путем максимизации функции правдоподобия $p(\mathfrak{D}|\mathbf{w})$ получена оценка параметров $\hat{\mathbf{w}}$. Такие параметры называются наиболее правдоподобными. Традиционный подход не использует понятия статистической сложности модели при оценке параметров. Поэтому, с целью исключения переобучения выбираемых моделей используются процедуры скользящего контроля.

Помимо функции правдоподобия в байесовском подходе рассматривается функция плотности распределения вектора параметров $p(\mathbf{w})$. При этом регрессионная выборка $\mathfrak{D}$ не используется, и данное распределение считается априорным. Апостериорная плотность распределения параметров, согласно теореме Байеса, имеет вид

$$p(\mathbf{w}|\mathfrak{D}) = \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w})}{p(\mathfrak{D})}.$$

Параметры, полученные путем максимизации функции плотности апостериорного распределения $p(\mathbf{w}|\mathfrak{D})$, называются наиболее вероятными. Знаменатель вышеприведенной формулы имеет вид

$$p(\mathfrak{D}) = \int p(\mathfrak{D}|\mathbf{w}')p(\mathbf{w}')d\mathbf{w}'$$

и рассматривается как нормировочный коэффициент, который необходим для того, чтобы интеграл апостериорного распределения $p(\mathbf{w}|\mathfrak{D})$ был равен единице.

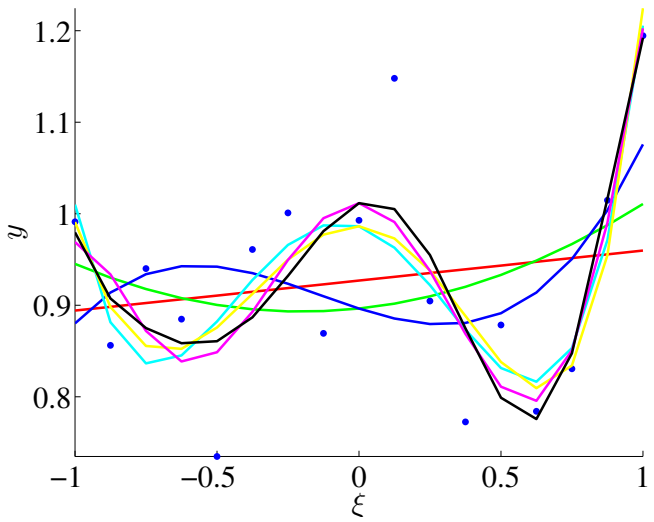



Рис. 40. Регрессионная выборка и её приближения полиномами.

Так как модель $f$, выбираемая из набора моделей $\mathfrak{F}$, зависит от значения вектора параметров $\mathbf{w}$, представим правдоподобие моделей в виде интеграла по пространству параметров

$$p(\mathfrak{D}|f) = \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, f)p(\mathbf{w}|f)d\mathbf{w}. \tag{111}$$

Априорная плотность распределения параметров $\mathbf{w}$ модели $f$ на выборке $\mathfrak{D}$ равна

$$p(\mathbf{w}|\mathfrak{D}, f) = \frac{p(\mathfrak{D}|\mathbf{w}, f)p(\mathbf{w}|f)}{p(\mathfrak{D}|f)}, \tag{112}$$

где $p(\mathbf{w}|f)$ — априорно заданная плотность распределения параметров и $p(\mathfrak{D}|\mathbf{w}, f)$ — функция правдоподобия параметров. Выражения (111) и (112) называются формулами байесовского вывода первого и второго уровня.

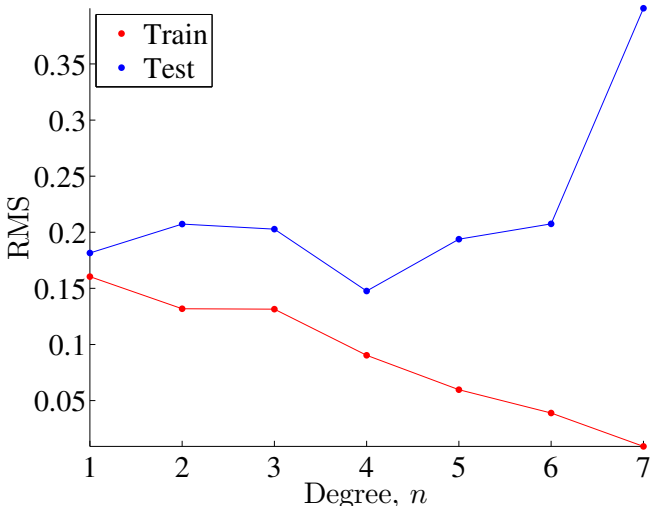



Рис. 41. Ошибка на тестовой и на обучающей выборке для полиномов различной степени.

Проиллюстрируем вышеописанное различие примером оценки параметров полиномиальной регрессионной модели, записанной в следующем виде:

$$f(\mathbf{x}, \mathbf{w}) = w_1 + w_2\xi + w_3\xi^2 + \ldots + w_n\xi^{n-1} = \sum_{j=1}^{n} w_j\xi^{j-1}.$$

Рассмотрим традиционный подход. На рисунке 40 показана синтетическая регрессионная выборка и искомая функция, по которой эта выборка была порождена. Также показаны несколько полиномиальных моделей, параметры которых получены методом наименьших квадратов. В случае, когда степень полинома слишком мала, $n = 0, 1$, результат является слабым приближением к синусоподобной кривой. В то же время при слишком большой степени полинома, $n = 9$, результат снова оказывается неудачным, так как модель переобучена. Наилучшая аппроксимация получена для модели «средней» структурной сложности, $n = 3$. Рисунок 41 подтверждает вышеприведенное предположение; по оси абсцисс отложена степень полинома, по оси ординат — среднеквадратичная ошибка для обучающей и тестовой подвыборок:

$$\text{RMS} = \sqrt{\frac{2S(\hat{\mathbf{w}})}{N}}, S = \|\mathbf{f}(\hat{\mathbf{w}}, \mathbf{X}) - \mathbf{y}\|^2.$$

Наилучшая обобщающая способность, иначе — наименьшая ошибка на контрольной выборке, доставлена моделями «средней» структурной сложности.

Рассмотрим байесовский подход к данной задаче. Пусть зависимая переменная имеет нормальное распределение, а ее математические ожидание зависит от независимой неслучайной величины $\mathbf{x}$. Тогда функция правдоподобия имеет вид:

$$p(\mathfrak{D}|\mathbf{w}) = \prod_{i=1}^{m} \mathcal{N}\big(\mathsf{E}\big(y_i|f(\mathbf{w}, \mathbf{x}_i)\big), \beta^{-1}\big),$$

где значение функции $f(\mathbf{w}, \mathbf{x}_i)$ является математическим ожиданием, а переменная $\beta$ обратна дисперсии случайной величины $y_i$. Пусть вектор параметров модели $\mathbf{w}$ является нормально распределенной случайной величиной,

$$p(\mathbf{w}|\alpha) = \mathcal{N}\big(\mathsf{E}(\mathbf{w}|\mathbf{0}), \alpha^{-1}\mathbf{I}\big) = \left(\frac{\alpha}{2\pi}\right)^{\frac{n}{2}} \exp\left(-\frac{\alpha}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w}\right),$$

где переменная $\alpha$ обратна дисперсии каждого элемента многомерной случайной величины $\mathbf{w}$. Используя теорему Байеса, можно вычислить апостериорное распределение параметра $\mathbf{w}$, которое также подчинено нормальному закону.

Для оценки обобщающей способности восстановим функцию плотности условного распределения зависимой переменной

$$p(y|\mathbf{x}, \mathfrak{D}) = \int p(y|x, \mathbf{w})p(\mathbf{w}|\mathfrak{D})d\mathbf{w} = \mathcal{N}\big(y|\mathsf{E}(\mathbf{w}^{\mathsf{T}}\mathbf{w}), \sigma^2(\mathbf{w}^{\mathsf{T}}\mathbf{w})\big),$$

в котором математическое ожидание $\mathsf{E}(x)$ и дисперсия $\sigma^2(x)$ (при предположении о равенстве дисперсий каждого элемента $x$ вектора $\mathbf{x}$) заданы выражениями

$$\mathsf{E}(x) = \beta\mathbf{x}^{\mathsf{T}}\mathbf{C}\sum_{i\in\mathcal{I}}\mathbf{x}_i y_i$$

и

$$\sigma^2(x) = \beta^{-1} + \mathbf{x}^\intercal \mathbf{C} \mathbf{x}.$$

Элементы вектора $\mathbf{x} = [x_1, \ldots, x_j, \ldots, x_n]^\intercal$ соответствуют значениям мономов полинома (принятой модели), $x_j = \xi^{j-1}$. Согласно гипотезе порождения данных, матрица $\mathbf{C}$ в этих выражениях, имеет вид:

$$\mathbf{C}^{-1} = \alpha \mathbf{I} + \beta \sum_{i \in \mathcal{I}} \mathbf{x}_i \mathbf{x}_i^\intercal.$$

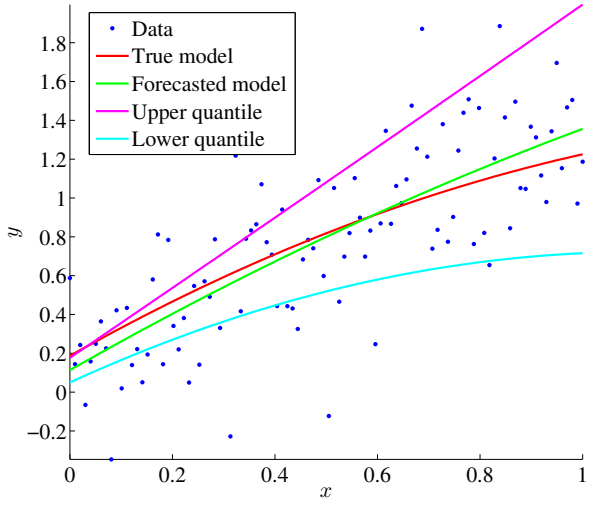



Рис. 42. Распределение зависимой переменной для модели оптимальной сложности.

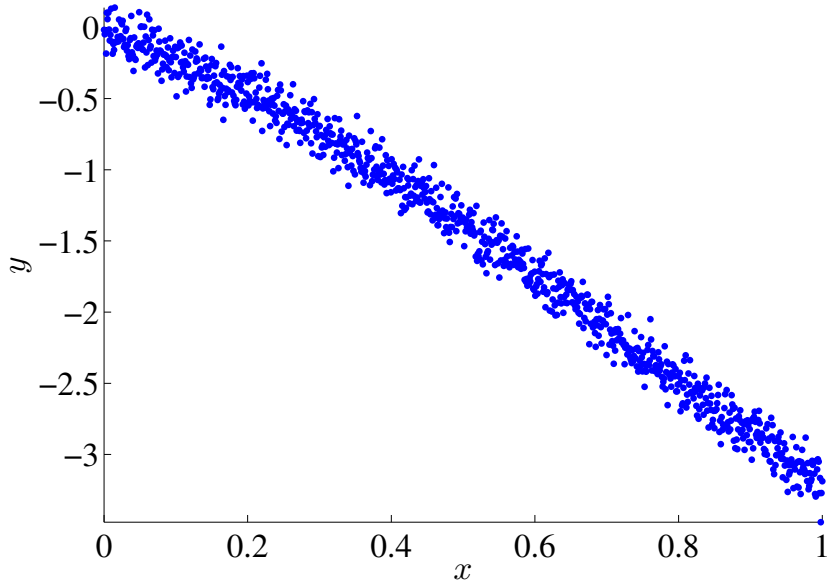
На рисунке 42 показан график плотности распределения зависимой переменной $y$ при изменяющейся независимой переменной $\xi$. Дисперсия этого условного распределения также зависит от независимой переменной $\xi$. Линиями показаны функция, породившая выборку и восстановленная функция регрессии $f(\hat{\mathbf{w}}, \xi)$ — математическое ожидание зависимой переменной $y$.

Рассмотрим традиционный подход к оценке параметров модели с помощью максимизации апостериорного распределения. Так как логарифм — монотонная функция, то для получения оценок максимизируем логарифм

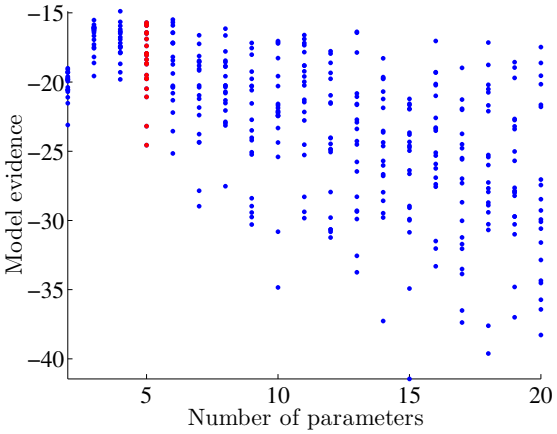
$$\ln p(\mathbf{w}|\mathfrak{D}) = -\frac{\beta}{2} \sum_{i \in \mathcal{I}} (\mathbf{x}_i^\intercal \mathbf{w} - y_i)^2 - \frac{\alpha}{2} \mathbf{w}^\intercal \mathbf{w},$$

равный сумме функции среднеквадратичной ошибки и штрафа за большие значения параметров. Второе слагаемое может рассматриваться как регуляризующий параметр [61]. Таким образом, традиционный подход является частным случаем байесовского.

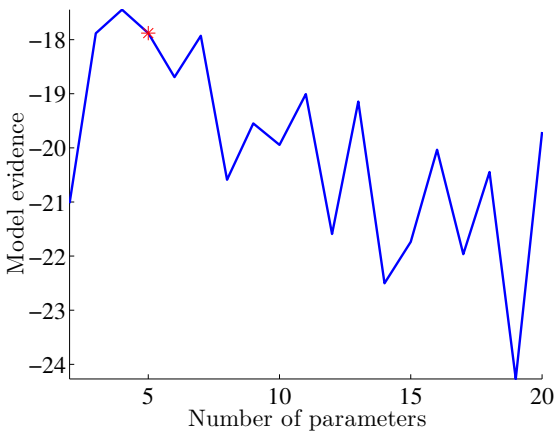
На рис. 43 показан график правдоподобия моделей для полиномов различной степени. На рис. 43a изображены данные, сгенерированные полиномиальной моделью степени 5. Рис. 43b и 43c иллюстрируют значения правдоподобия модели для различных подвыборок и в среднем.




(a) Сгенерированные данные: полиномиальная модель степени 5




(b) Правдоподобие для различных выборок




(c) Среднее правдоподобие

Рис. 43. Правдоподобие полиномов различной степени в качестве регрессионных моделей.

### 4.1.4. Второй уровень связанного байесовского вывода

При сравнении моделей используется правило бритвы Оккама в следующей формулировке: «Совместный байесовский вывод автоматически количественно выполняет правило Оккама». Бритва Оккама — принцип предпочтения простых моделей (теорий, гипотез) сложным. Если несколько моделей одинаково хорошо описывают наблюдения, принцип Оккама рекомендует выбор простейшей модели.

Теорема Байеса говорит о том, что наиболее вероятными будут те модели, которые наиболее точно прогнозируют появление некоторых данных.

Эта вероятность определяется нормализованной функцией распределения на пространстве данных $\mathfrak{D}$. Вероятность $p(\mathfrak{D}|f_k)$ появления данных $\mathfrak{D}$ при фиксированной модели $f_k$ называется правдоподобием модели $f_k$.

Найдем правдоподобие двух альтернативных моделей $f_1$ и $f_2$, описывающих данные $\mathfrak{D}$. По теореме Байеса мы связываем правдоподобие $\mathrm{p}(f_1|\mathfrak{D})$ модели $f_1$ при фиксированных данных, то есть, вероятность $p(\mathfrak{D}|f_1)$ получения данных с этой моделью и априорное правдоподобие $p(f_1)$ модели $f_1$. Так как значение нормирующего множителя

$$p(\mathfrak{D}) = \sum_{k=1}^{K} p(\mathfrak{D}|f_k)p(f_k)$$

для обеих моделей (здесь $K = 2$) одинаково, то отношение правдоподобия моделей $f_1$ и $f_2$ имеет вид

$$\frac{p(f_1|\mathfrak{D})}{p(f_2|\mathfrak{D})} = \frac{p(f_1)p(\mathfrak{D}|f_1)}{p(f_2)p(\mathfrak{D}|f_2)}. \tag{113}$$

Отношение $\frac{p(f_1)}{p(f_2)}$ в правой части указывает на то, насколько велико априорное предпочтение модели $p(f_1)$ ее альтернативе $p(f_2)$. Отношение $\frac{p(\mathfrak{D}|f_1)}{p(\mathfrak{D}|f_2)}$ указывает насколько модель $f_1$ соответствует наблюдаемым данным лучше, чем модель $f_2$.

Выражение (113) вводит правило Оккама следующим образом. Во-первых, можно задать отношение $\frac{p(f_1)}{p(f_2)}$ соответствующее отношению предпочтения моделей или критерию отбора моделей, на основании некоторой дополнительной информации. Во-вторых, независимо от предыдущего способа задания критерия отбора моделей, это отношение автоматически выполняет правило Оккама. Действительно, если $f_2$ — более сложная модель, ее плотность распределения $p(\mathfrak{D}|f_2)$ имеет меньшие значения при том условии, что ее дисперсия больше. Если невязки, доставляемые обеими моделями равны, простая модель $f_1$ будет более вероятна, чем сложная модель $f_2$. Таким образом, независимо от априорных предпочтений, вводится правило Оккама, согласно которому при равных априорных предпочтениях и равном соответствии предполагаемых моделей измеряемым данным простая модель более вероятна, чем сложная.

### 4.1.5. Функции правдоподобия моделей и данных

При создании моделей различают два уровня байесовского вывода [49, 46, 188]. На *первом уровне* предполагается, что рассматриваемая модель адекватна. Находится оценка параметров моделей по регрессионной выборке. В результате получаются наиболее правдоподобные

значения параметров и значения ошибок моделей при этих параметрах. Эта процедура повторяется для каждой модели. Задача, решаемая на *втором уровне вывода* — сравнение моделей, см. рис. 44. Результатом является упорядоченное множество моделей.

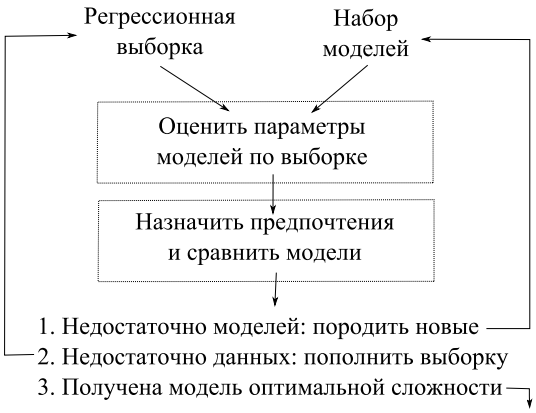



Рис. 44. Использование байесовского вывода при выборе моделей; первый и второй уровень вывода обведены пунктирной линией.

Каждая модель $f_k$ имеет вектор параметров $\mathbf{w}$. Задача первого уровня — получить оценку параметров $\mathbf{w}$ модели при полученных данных $\mathfrak{D}$. Согласно теореме Байеса, апостериорная вероятность параметров $\mathbf{w}$ равна

$$p(\mathbf{w}|\mathfrak{D}, f_k) = \frac{p(\mathfrak{D}|\mathbf{w}, f_k)p(\mathbf{w}|f_k)}{p(\mathfrak{D}|f_k)}. \tag{114}$$

Нормирующая константа $p(\mathfrak{D}|f_k)$ обычно не принимается во внимание на первом уровне вывода. Однако она становится весьма важной на втором уровне вывода. Эта константа называется правдоподобие модели (англ. «evidence», дословно «достоверность»).

При оценке параметров на практике обычно применяют оптимизационные методы, например, метод сопряженных градиентов, чтобы получить наиболее вероятные параметры $\mathbf{w}_{\text{MP}}$. При этом различают наиболее вероятные параметры $\mathbf{w}_{\text{MP}}$, которые выводятся на первом уровне как аргумент функции вероятности, и наиболее правдоподобные параметры $\mathbf{w}_{\text{ML}}$, которые оцениваются как аргумент функции наибольшего правдоподобия.

Обобщающая способность (иногда называемая прогностической способностью) модели $f$ оценивается с помощью функции апостериорного распределения параметров модели. Для оценки используется разложение в ряд Тейлора логарифма апостериорного распределения функции $p(\mathbf{w}|\mathfrak{D}, f_k)$

$$p(\mathbf{w}|\mathfrak{D}, f_k) \approx p(\mathbf{w}_{\text{MP}}|\mathfrak{D}, f_k) \exp\left(-\frac{1}{2}\Delta\mathbf{w}^{\intercal} A \Delta\mathbf{w}\right),$$

где $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\text{MP}}$, и отыскивается значение гессиана при значении параметров максимального правдоподобия $\mathbf{w}_{\text{MP}}$ в окрестности $\mathbf{w}_{\text{MP}}$:

$$A = -\nabla^2 \ln p(\mathbf{w}|\mathfrak{D}, f_k)|_{\mathbf{w}_{\text{MP}}}.$$

Таким образом, функция апостериорного распределения параметров модели $f_k$ может быть локально приближена с помощью матрицы $A^{-1}$, которая является оценкой ковариационной матрицы параметров в окрестности $\mathbf{w}_{\text{MP}}$.

На *втором уровне* байесовского вывода требуется определить, какая модель наиболее адекватно описывает данные. Апостериорная вероятность $k$-й модели задана как

$$p(f_k|\mathfrak{D}) \propto p(\mathfrak{D}|f_k)p(f_k). \tag{115}$$

Следует отметить, что сомножитель $p(\mathfrak{D}|f_k)$, зависящий от регрессионной выборки $\mathfrak{D}$, есть правдоподобие модели $f_k$, которое было названа ранее, в выражении (114), нормирующим множителем. Правдоподобие модели может быть получена путем интегрирования функции правдоподобия по всему пространству параметров модели:

$$p(\mathfrak{D}|f_k) = \int p(\mathfrak{D}|\mathbf{w}, f_k)p(\mathbf{w}|f_k)d\mathbf{w}.$$

Второй сомножитель $p(f_k)$ в выражении (115) — априорная вероятность на множестве моделей, определяет, насколько адекватной является модель до того, как появились данные. Основной проблемой байесовского вывода является отсутствие объективных методов назначения априорной вероятности $p(f_k)$. Пусть априорные вероятности $p(f_k)$ всех моделей равны. Тогда модели ранжируются по значениям $p(\mathfrak{D}|f_k)$ их правдоподобия.

Важное предположение, которое необходимо сделать для решения задачи вычисления правдоподобия, — предположение о том, что распределение $p(\mathbf{w}|\mathfrak{D}, f_k) \propto p(\mathfrak{D}|\mathbf{w}, f_k)p(\mathbf{w}|f_k)$ имеет выраженный максимум в окрестности наиболее вероятного значения параметров $\mathbf{w}_{\text{MP}}$.

На рис. 45 показана оценка матрицы ковариаций распределения параметров модели. Зеленым цветом показана оценка методом Монте-Карло, красным — методом скользящего контроля.

Функцию распределения параметров модели приближают гауссианой, определенной в пространстве параметров. Для этого используют аппроксимацию Лапласа. Согласно данному методу, эта функция приближенно равна высоте пика подынтегрального выражения $p(\mathfrak{D}|\mathbf{w}, f_k)p(\mathbf{w}|f_k)$ умноженной на ширину пика, $\sigma_{\mathbf{w}|\mathfrak{D}}$:

$$p(\mathfrak{D}|f_k) \approx p(\mathfrak{D}|\mathbf{w}_{\text{MP}}, f_k)p(\mathbf{w}_{\text{MP}}|f_k)\sigma_{\mathbf{w}|\mathfrak{D}},$$

$$\text{правдоподобие модели} \approx \text{правдоподобие данных} \cdot \text{множитель Оккама}.$$

Таким образом, правдоподобие модели находится с помощью оценок наибольшего правдоподобия параметров модели и множителя Оккама, принимающего значения на отрезке $[0, 1]$, который штрафует модель $f_k$ за ее параметры $\mathbf{w}$. Чем точнее проведена априорная оценка параметров, тем меньше штраф.

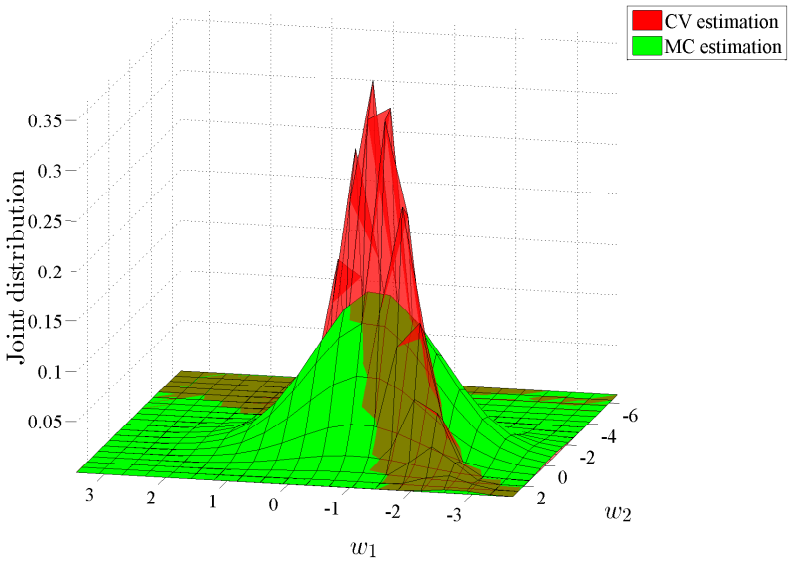
При аппроксимации Лапласа множитель Оккама может быть получен с помощью определителя ковариационной матрицы параметров

$$p(\mathfrak{D}|f_k) \approx \frac{p(\mathfrak{D}|\mathbf{w}_{\text{MP}}, f_k)p(\mathbf{w}_{\text{MP}}|f_k)}{\sqrt{\det\left(\frac{1}{2\pi}A\right)}},$$

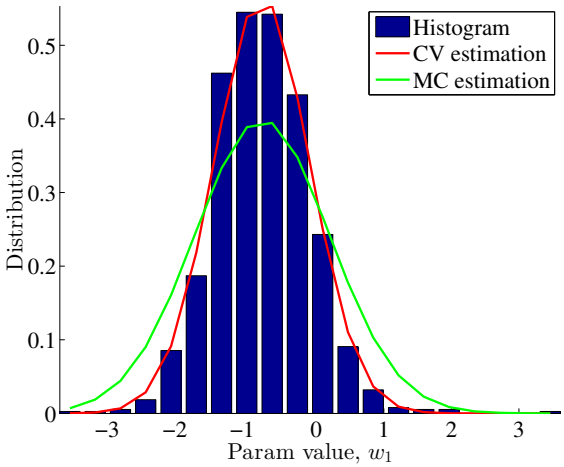
где

$$A = -\nabla^2 \ln p(\mathbf{w}|\mathfrak{D}, f_k)|_{\mathbf{w}=\mathbf{w}_{\text{MP}}}$$

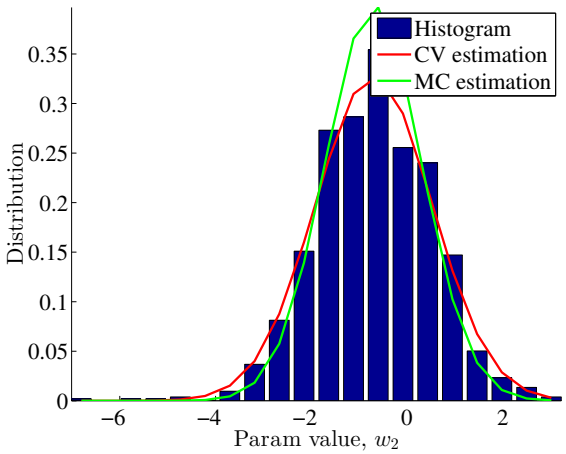
есть гессиан ковариационной матрицы параметров, вычисленный в точке $\mathbf{w}_{\text{MP}}$. Алгоритмически байесовский метод сравнения моделей посредством вычисления из правдоподобия не сложнее, чем задача оценки параметров каждой модели и оценки матрицы Гессе.




(a) Оценки совместного распределения параметров




(b) Распределение параметра $w_1$ и его оценки




(c) Распределение параметра $w_2$ и его оценки

Рис. 45. Оценка ковариации параметров модели.

Итак, для того чтобы отранжировать альтернативные модели $f_k$ по предпочтению, необходимо, воспользовавшись байесовским выводом, вычислить правдоподобие $p(\mathfrak{D}|f_k)$. Байесовское сравнение моделей — это расширение метода наибольшего правдоподобия. Правдоподобие можно вычислить как для параметрических, так и для непараметрических моделей.

**Пример интерпретации множителя Оккама.** Переменная $\sigma_{\mathbf{w}|\mathfrak{D}}$ является апостериорной неопределенностью вектора параметров $\mathbf{w}$. Пусть априорное распределение $p(\mathbf{w}|f_k)$ является равномерным на некотором большом интервале, например, $(\mathbf{w}_{\text{MP}} + 3\sigma_{\mathbf{w}}, \mathbf{w}_{\text{MP}} - 3\sigma_{\mathbf{w}})$, и принимает множество значений, которые возможны априори согласно модели $f_k$. Тогда $p(\mathbf{w}_{\text{MP}}|f_k) = \sigma_{\mathbf{w}}^{-1}$, и множитель Оккама равен $\sigma\ {}^{-1}_{\mathbf{w}|\mathfrak{D}}\sigma_{\mathbf{w}}$. Множитель Оккама есть степень сжатия пространства параметров модели при появлении данных [189, 191, 190]. Модель $f_k$ может быть представлена семейством параметрических функций, из которых фиксируется одна, как только появляются данные. Множитель Оккама есть число, обратное количеству таких функций (для конечного их числа). Логарифм множителя Оккама есть мера количества информации о параметрах модели, которая будет получена при появлении данных.

### 4.1.6. Использование байесовского вывода при выборе моделей

Воспользуемся двухуровневым байесовским выводом для оценки степени предпочтения порождаемых регрессионных моделей. Рассмотрим конечное множество моделей $f_1, \ldots, f_M$, приближающих данные $\mathfrak{D}$, обозначим априорную вероятность $i$-й модели $p(f_k)$. При заданной регрессионной выборке апостериорная вероятность модели $p(f_k|\mathfrak{D})$ равна

$$p(f_k|\mathfrak{D}) = \frac{p(\mathfrak{D}|f_k)p(f_k)}{\sum_{j=1}^{M} p(\mathfrak{D}|f_j)p(f_j)}, \tag{116}$$

где $p(\mathfrak{D}|f_k)$ — функция правдоподобия моделей, определяющая, насколько хорошо модель $f_k$ описывает данные $\mathfrak{D}$. Знаменатель дроби обеспечивает выполнение условия $\sum_{i=1}^{M} p(f_k|\mathfrak{D}) = 1$.

Сравним две модели с помощью апостериорных вероятностей

$$\frac{p(f_k|\mathfrak{D})}{p(f_j|\mathfrak{D})} = \frac{p(\mathfrak{D}|f_k)p(f_k)}{p(\mathfrak{D}|f_j)p(f_j)}. \tag{117}$$

Левая часть выражения называется отношением правдоподобия моделей. Отношение $p(f_k)/p(f_j)$ называется отношением апостериорных предпочтений моделей. Полагая априорные вероятности моделей одинаковыми, используем функции правдоподобия для выбора моделей.

Так как рассматриваемые модели $f$ зависят от параметров, представим их правдоподобие в виде интеграла по пространству параметров

$$p(\mathfrak{D}|f) = \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, f)p(\mathbf{w}|f)d\mathbf{w}. \tag{118}$$

Априорная плотность распределения параметров $\mathbf{w}$ модели $f$ на выборке $\mathfrak{D}$ равна

$$p(\mathbf{w}|\mathfrak{D}, f) = \frac{p(\mathfrak{D}|\mathbf{w}, f)p(\mathbf{w}|f)}{p(\mathfrak{D}|f)}, \tag{119}$$

где $p(\mathbf{w}|f)$ — априорное распределение параметров, а $p(\mathfrak{D}|\mathbf{w},f)$ — функция правдоподобия параметров. Выражения (116) и (119) называются формулами Байесовского вывода первого и второго уровня.

## 4.2. Методы аналитической оценки гиперпараметров

Ниже предложен ряд методов оптимизации структурных параметров регрессионной модели. Описан метод аппроксимации Лапласа функции ошибки для оценки правдоподобия модели. Предложен метод Монте-Карло оценки правдоподобия модели. Предложен метод оценки оптимальных параметров модели с помощью процедуры скользящего контроля. Исследованы свойства предлагаемых методов. Проведен вычислительный эксперимент на модельных и реальных данных. Проведены анализ и сравнение предлагаемых методов.

Предположим, что ненормированное распределение параметров, полученное в предыдущем разделе, имеет единственную моду. Используя аппроксимацию Лапласа (145), приблизим функцию плотности нормального распределения эмпирическим распределением для того, чтобы оценить ковариационные матрицы $\mathbf{A},\mathbf{B}$ совместно с параметрами $\mathbf{w}$ регрессионной модели. Перепишем в удобном виде функцию правдоподобия (16), функцию априорного и функцию апостериорного распределения параметров (17), (18), введенные в разделе 1. При этом полагаем, что модель фиксирована и является обобщенно-линейной. Это позволяет нам опустить символ $f$ в аргументах функций. Функция правдоподобия данных имеет вид

$$p(\mathfrak{D}|\mathbf{w},\beta) = \frac{\exp(-E_{\mathfrak{D}})}{Z_{\mathfrak{D}}(\mathbf{B})} = \frac{\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\intercal}\mathbf{B}(\mathbf{y}-\mathbf{f})\right)}{(2\pi)^{\frac{m}{2}}\det^{\frac{1}{2}}(\mathbf{B})}, \tag{120}$$

функция априорного распределения параметров, при предположении о том, что оценка матожидания вектора параметров равна $\mathbf{w}_0$ имеет вид

$$p(\mathbf{w}|\mathbf{A}) = \frac{\exp(-E_{\mathbf{w}})}{Z_{\mathbf{w}}(\mathbf{A})} = \frac{\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\intercal} = \mathbf{A}(\mathbf{w}-\mathbf{w}_0)\right)}{(2\pi)^{\frac{n}{2}}\det^{\frac{1}{2}}(\mathbf{B})}, \tag{121}$$

а функция апостериорного распределения параметров —

$$p(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B}) = \frac{p(\mathfrak{D}|\mathbf{w},\mathbf{B})p(\mathbf{w}|\mathbf{A})}{p(\mathfrak{D}|\mathbf{A},\mathbf{B})} = \frac{\exp\big(-S(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B})\big)}{Z_S}. \tag{122}$$

Зафиксируем значение вектора $\mathbf{w}_0$, предполагая что он доставляет локальный максимум (122). Для нахождения матриц $\mathbf{A},\mathbf{B}$ приблизим фукнцию ошибки $S(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B})$ методом Лапласа. Для этого построим ряд Тейлора второго порядка логарифма числителя ((122)) в окрестности $\mathbf{w}_0$:

$$\ln\exp\big(-S(\mathbf{w})\big) = \ln\exp\left(S(\mathbf{w}_0) + 0 + \frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}\Delta\mathbf{w} + o(||\mathbf{w}||^3)\right),$$

где $\Delta\mathbf{w} = \mathbf{w}-\mathbf{w}_0$. При упрощении и отбрасывании малой величины получим

$$-S(\mathbf{w}) \approx -S(\mathbf{w}_0) - \frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}\Delta\mathbf{w}. \tag{123}$$

В выражении (123) нет слагаемого первого порядка, так как предполагается, что $\mathbf{w}_0$ доставляет локальный минимум функции ошибки

$$\left.\frac{\partial S(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_0} = \mathbf{0}.$$

Матрица $\mathbf{H}$ — матрица Гессе функции ошибок

$$\mathbf{H} = -\nabla\nabla S(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}. \tag{124}$$

**Замечание 2.** *Так как производная первого члена – функции ошибки $E_{\mathfrak{D}}$, заданная функцией правдоподобия $p(\mathfrak{D}|\mathbf{w},\mathbf{B})$, не зависит от параметров модели, то оценка ковариационной матрицы, полученная с помощью аппроксимации Лапласа будет справедлива для любой гипотезы порождения данных рассматривающей распределение из экспоненциального семейства.*

Применяя экспоненту к обеим частям выражения (123) получаем требуемое приближение числителя (122)

$$\exp\big(-S(\mathbf{w})\big) \approx \exp\big(-S(\mathbf{w}_0)\big) \exp\left(-\frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}\Delta\mathbf{w}\right). \tag{125}$$

Таким образом, апостериорное распределение параметров при фиксированных значениях ковариационных матриц $\mathbf{A},\mathbf{B}$ принимает вид

$$p(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B}) \approx \frac{\exp\big(-S(\mathbf{w}_0)\big)\exp\left(-\frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}\Delta\mathbf{w}\right)}{Z_S(\mathbf{A},\mathbf{B})}. \tag{126}$$

Так как интеграл выражения апостериорного распределения параметров должен равняться единице,

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B})d\mathbf{w} = 1,$$

то нормирующий множитель, полученный посредством аппроксимации Лапласа, равен

$$Z_S = \exp\big(-S(\mathbf{w}_0)\big)(2\pi)^{\frac{n}{2}}\det{}^{-\frac{1}{2}}(\mathbf{H}). \tag{127}$$

Для нахождения гиперпараметров максимизируем нормирующую функцию $p(\mathfrak{D}|\mathbf{A},\mathbf{B})$ из выражения (122) относительно $\mathbf{A}$ и $\mathbf{B}$. Запишем ее в виде

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \int p(\mathfrak{D}|\mathbf{w},\mathbf{A},\mathbf{B})p(\mathbf{w}|\mathbf{A})d\mathbf{w}. \tag{128}$$

Используя выражения (120) и (121) перепишем (122) в виде

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \frac{1}{Z_{\mathbf{w}}(\mathbf{A})}\frac{1}{Z_{\mathfrak{D}}(\mathbf{B})}\int \exp\big(-S(\mathbf{w})\big)d\mathbf{w}.$$

Последнее выражение также можно переписать, используюя нормирующую константу $Z_S$ апостериорного распределения:

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = Z_{\mathbf{w}}^{-1}(\mathbf{A})Z_{\mathfrak{D}}^{-1}(\mathbf{B})Z_S.$$

Поставим нормирующие выражения в переменные $Z_{\mathfrak{D}}$ и $Z_{\mathbf{w}}$ в формулу (127)

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = Z_{\mathbf{w}}^{-1}(\mathbf{A}) Z_{\mathfrak{D}}^{-1}(\beta) \exp\bigl(-S(\mathbf{w}_0)\bigr)(2\pi)^{\frac{n}{2}} \det{}^{-\frac{1}{2}}(\mathbf{H}).$$

и прологарифмируем

$$\begin{aligned}\ln p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = &\underbrace{-\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det(\mathbf{A})}_{Z_{\mathbf{w}}^{-1}(\mathbf{A})} - \\ &\underbrace{-\frac{m}{2}\ln 2\pi + \frac{1}{2}\ln\det(\mathbf{B})}_{Z_{\mathfrak{D}}^{-1}(\mathbf{B})} \underbrace{- S(\mathbf{w}_0) + \frac{n}{2}\ln 2\pi - \frac{1}{2}\ln\det(\mathbf{H})}_{Z_S}.\end{aligned} \tag{129}$$

При упрощении данного выражения, с учетом того, что

$$2S(\mathbf{w}_0) = \mathbf{w}_0^{\intercal}\mathbf{A}\mathbf{w}_0 + \bigl(\mathbf{y} - \mathbf{f}(\mathbf{w}_0,\mathbf{X})\bigr)^{\intercal}\mathbf{B}\bigl(\mathbf{y} - \mathbf{f}(\mathbf{w}_0,\mathbf{X})\bigr),$$

получаем

$$\begin{aligned}\ln p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = &-\frac{1}{2}\ln\det(\mathbf{A}) - \frac{m}{2}\ln 2\pi + \frac{m}{2}\ln\det(\mathbf{B}) - \\ &\underbrace{-\mathbf{w}_0^{\intercal}\mathbf{A}\mathbf{w}_0 + (\mathbf{y} - \mathbf{f}_{\mathbf{w}_0})^{\intercal}\mathbf{B}(\mathbf{y} - \mathbf{f}_{\mathbf{w}_0})}_{-S(\mathbf{w}_0)} - \frac{1}{2}\ln\det(\mathbf{H}).\end{aligned} \tag{130}$$

Последнее слагаемое включает гессиан $\mathbf{H}$, определенный в выражении (124).

### 4.2.1. Процедура оценивания параметров и гиперпараметров

Для оценки структурных параметров необходимо провести процедуру максимизации правдоподобия модели. Именно эта процедура является наиболее вычислительно затратной. Оптимальные структурные параметры $\mathbf{A},\mathbf{B}$ максимизируют правдоподобие модели

$$p\left(\mathfrak{D}|\mathbf{A},\mathbf{B}\right) = \int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w},\mathbf{B})p(\mathbf{w}|\mathbf{A})d\mathbf{w} \to \max_{\mathbf{A}\in\mathbb{M}^n,\mathbf{B}\in\mathbb{M}^m}, \tag{131}$$

где $\mathbb{M}^n$ обозначает множество положительно полуопределенных матриц размерности $n \times n$.

Оптимальные значения гиперпараметров $\alpha$ и $\beta$ — элементов матриц $\mathbf{A}$ и $\mathbf{B}$ вычисляются итеративно следующим образом. При фиксированных параметрах $\mathbf{w}_0$ находятся оптимальные значения $\alpha$. С использованием $\alpha$ находятся оптимальные значения $\beta$. Далее новые $\beta$ определяют новые значения вспомогательной переменной $\lambda$. Цикл повторяется до тех пор, пока изменение значений $\alpha,\beta$ на соседних шагах не станет менее заранее заданной границы.

Оптимальные значения параметров $\mathbf{w}$ переоцениваются с использованием функции ошибки $S(\mathbf{w}|\mathfrak{D},\mathbf{A},\mathbf{B})$, определенной числителем (122) при фиксированных на данном шаге значениях матриц $\mathbf{A},\mathbf{B}$. Таким образом, параметры $\mathbf{w}$ и гиперпараметры $\mathbf{A},\mathbf{B}$ регрессионной модели $f$ оцениваются по отдельности. На каждой итерации сначала при фиксированных значениях гиперпараметров отыскиваются параметры путем оптимизации функции $S(\mathbf{w})$. При этом используется алгоритм Левенберга-Марквардта или его модификации, описанные в разделе 1. Затем по формулам, указанным выше, оцениваются матрицы гиперпараметров $\mathbf{A},\mathbf{B}$.

Найдем максимум выражения (129) относительно элементов обратных ковариационных матриц, приравняв его производную поочередно по $\mathbf{A}$ и по $\mathbf{B}$ к нулю. Три этом рассмотрим три варианта: матрицы $\mathbf{A}, \mathbf{B}$ — неотрицательно определенные, общего вида, матрицы $\mathbf{A} = \operatorname{diag}(\boldsymbol{\alpha})\mathbf{I}$ и $\mathbf{B} = \operatorname{diag}(\boldsymbol{\beta})\mathbf{I}$ — диагональные, матрицы $\mathbf{A} = \alpha\mathbf{I}$ и $\mathbf{B} = \beta\mathbf{I}$ имеют на диагоналях равные элементы.

### 4.2.2. Аналитическая оценка ковариационных матриц общего вида

Для того, чтобы оценить структурные параметры $\mathbf{A}, \mathbf{B}$ совместно с параметрами $\mathbf{w}$ регрессионной модели, воспользуемся методом аппроксимации Лапласа функции правдоподобия модели.

В данном параграфе примем нормальную гипотезу распределения зависимой переменной и априорного распределения параметров модели. Таким образом, для нахождения оптимальных структурных параметров $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ выражение (131) преобразуется следующим образом:

$$\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})}\int\limits_{\mathbf{w}\in\mathbb{W}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{f})^{\intercal}\mathbf{B}(\mathbf{y}-\mathbf{f})\right)\exp\left(-\frac{1}{2}\mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}\right)d\mathbf{w} \to \max_{\mathbf{A}\in\mathbb{M}^n,\mathbf{B}\in\mathbb{M}^m}. \tag{132}$$

Примем за функцию ошибки $S(\mathbf{w}, \mathbf{A}, \mathbf{B})$ показатель экспоненты выражения (132) с отрицательным знаком:

$$S(\mathbf{w}, \mathbf{A}, \mathbf{B}) = \frac{1}{2}(\mathbf{y}-\mathbf{f})^{\intercal}\mathbf{B}(\mathbf{y}-\mathbf{f}) + \frac{1}{2}\mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}, \tag{133}$$

тогда оптимизационная задача (132) перепишется в более удобном виде:

$$\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})}\int\limits_{\mathbf{w}\in\mathbb{W}}\exp\big(-S\left(\mathbf{w}, \mathbf{A}, \mathbf{B}\right)\big)d\mathbf{w} \to \max_{\mathbf{A},\mathbf{B}}.$$

Отметим, что оптимальными параметрами $\hat{\mathbf{w}}$ модели $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X})$ являются те, которые максимизируют апостериорное распределение параметров или, в нашем случае, минимизируют функцию ошибки

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{W}} S(\mathbf{w}|\hat{\mathbf{A}}, \hat{\mathbf{B}}),$$

где $\hat{\mathbf{A}}, \hat{\mathbf{B}}$ — оптимальные структурные параметры, максимизирующие выражение (132).

Метод аппроксимации Лапласа состоит в разложении функцию ошибки $S(\mathbf{w})$ вокруг оптимального значения $S(\hat{\mathbf{w}})$ для аппроксимации выражения (132):

$$S(\mathbf{w}) = S(\hat{\mathbf{w}}) + \frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}\Delta\mathbf{w} + o(||\mathbf{w}||^2),$$

где $\mathbf{H}$ — матрица Гессе функции ошибок

$$\mathbf{H} = \nabla\nabla S(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}$$

в точке $\mathbf{w} = \hat{\mathbf{w}}$. Здесь и далее под нормой $\|\mathbf{w}\|$ подразумевается евклидова норма $\|\mathbf{w}\| = \|\mathbf{w}\|_2$. Вместо оптимизации выражения (132), будем оптимизировать аппроксимированное выражение

$$\frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})}\exp\big(S(\hat{\mathbf{w}})\big)\int\limits_{\mathbf{w}\in\mathbb{W}}\exp\left(-\frac{1}{2}\Delta\mathbf{w}^{\intercal}\mathbf{H}\Delta\mathbf{w}\right)d\mathbf{w} \to \max_{\mathbf{A},\mathbf{B}}. \tag{134}$$

Отметим, что подынтегральное выражение в (134) является частью нормального распределения, поэтому весь интеграл в (134) можно заменить на нормировочную константу и получить оптимизационную задачу вида:

$$g(\mathbf{A},\mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}}}{(2\pi^{\frac{m}{2}})}\frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi^{\frac{n}{2}})}\exp\bigl(S(\hat{\mathbf{w}})\bigr)\frac{(2\pi^{\frac{n}{2}})}{|\mathbf{H}|^{\frac{1}{2}}} \to \max_{\mathbf{A},\mathbf{B}}. \tag{135}$$

Прологарифмируем выражение (135) и будем искать оптимум в виде:

$$-\ln g(\mathbf{A},\mathbf{B}) = -\frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{A}| + \frac{1}{2}\ln|\mathbf{B}| - S(\mathbf{w}_0) - \frac{1}{2}\ln|\mathbf{H}| \to \max_{\mathbf{A},\mathbf{B}}. \tag{136}$$

Для дальнейших рассуждений примем некоторые ограничения на вид матриц $\mathbf{A},\mathbf{B}$, позволяющие упростить вид функции $\ln g(\mathbf{A},\mathbf{B})$. В частности, везде далее будем рассматривать случай скалярной матрицы $\mathbf{B} = \beta\mathbf{I}$.

В случае скалярной матрицы $\mathbf{B} = \beta\mathbf{I}$, функция ошибки (133) записывается следующим образом:

$$S(\mathbf{w},\mathbf{A},\beta) = \frac{\beta}{2}(\mathbf{y}-\mathbf{f})^{\intercal}(\mathbf{y}-\mathbf{f}) + \frac{1}{2}\mathbf{w}^{\intercal}\mathbf{A}\mathbf{w} = \beta S_{\mathfrak{D}}(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}, \tag{137}$$

где

$$S_{\mathfrak{D}}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{f})^{\intercal}(\mathbf{y}-\mathbf{f}), \tag{138}$$

и гессиан $\mathbf{H}$ записывается в виде

$$\mathbf{H} = \beta\mathbf{H}_{\mathfrak{D}} + \mathbf{A},$$

где $\mathbf{H}_{\mathfrak{D}}$ — гессиан функции $S_{\mathfrak{D}}(\mathbf{w})$ в точке $\mathbf{w} = \hat{\mathbf{w}}$.

Функция (136) записывается следующим образом:

$$\begin{aligned}-\ln g(\mathbf{A},\beta) = -\frac{m}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{A}| + \frac{m}{2}\ln\beta - \frac{\beta}{2}\bigl(\mathbf{y}-\mathbf{f}\left(\hat{\mathbf{w}},\mathbf{X}\right)\bigr)^{\intercal}\bigl(\mathbf{y}-\mathbf{f}\left(\hat{\mathbf{w}},\mathbf{X}\right)\bigr) - \\ -\frac{1}{2}\hat{\mathbf{w}}^{\intercal}\mathbf{A}\hat{\mathbf{w}} - \frac{1}{2}\ln|\beta\mathbf{H}_{\mathfrak{D}} + \mathbf{A}| \to \max_{\mathbf{A},\mathbf{B}}.\end{aligned} \tag{139}$$

Далее, будем рассматривать частные случаи скалярной и диагональной матрицы $\mathbf{A}$, что позволит дифференцировать слагаемое

$$\frac{1}{2}\ln|\beta\mathbf{H}_{\mathfrak{D}} + \mathbf{A}| \tag{140}$$

формулы (139).

### 4.2.3. Одинаковая дисперсия элементов вектора параметров

Рассмотрим случай одинаковых элементов на диагонали ковариационной матрицы $\mathbf{A} = \alpha\mathbf{I}$. При этом упрощении посчитаем выражение (140):

$$\frac{1}{2}\ln|\beta\mathbf{H}_{\mathfrak{D}} + \alpha\mathbf{I}| = \frac{1}{2}\sum_{j=1}^{n}\ln(\beta h_j + \alpha),$$

где $h_j$ — собственное число матрицы $\mathbf{H}_{\mathfrak{D}}$.

Приравняв производные выражения (139) по $\alpha$ и $\beta$ к нулю, найдем оптимальные значения структурных параметров $\alpha$ и $\beta$.

$$\frac{\partial(-\ln g(\alpha,\beta))}{\partial\alpha} = \frac{n}{2\alpha} - \frac{\|\hat{\mathbf{w}}\|^2}{2} - \frac{1}{2}\sum_{j=1}^{n}\frac{1}{\beta h_j+\alpha} = 0,$$

$$\alpha\|\hat{\mathbf{w}}\|^2 = n - \sum_{j=1}^{n}\frac{\alpha}{\beta h_j+\alpha} = \beta\sum_{j=1}^{n}\frac{h_j}{\beta h_j+\alpha}.$$

Введем обозначение

$$\gamma = \beta\sum_{j=1}^{n}\frac{h_j}{\beta h_j+\alpha}, \tag{141}$$

тогда

$$\alpha = \frac{\gamma}{\|\hat{\mathbf{w}}\|^2}. \tag{142}$$

Аналогично, приравняв производную выражения (139) по $\beta$ к нулю, получаем

$$\beta = \frac{m-\gamma}{\|\mathbf{y}-\mathbf{f}(\hat{\mathbf{w}},\mathbf{X})\|^2}. \tag{143}$$

Поскольку $\gamma$ является функцией от $\beta$ и $\alpha$, а также от оптимального значения параметров модели $\hat{\mathbf{w}}$, уравнения (141), (142) и (143) решаются итеративно для фиксированного $\hat{\mathbf{w}}$.

### 4.2.4. Независимо-распределенные элементы вектора параметров

В случае $\mathbf{A} = \mathrm{diag}(\alpha_j)$ результаты оказываются сравнимыми с результатами из предыдущего параграфа. В частности, вместо выражения (141) примем за $\rho$ величину

$$\rho = \beta\sum_{j=1}^{n}\frac{h_j}{\beta h_j+\alpha_j},$$

тогда выражение для $\beta$ будет таким:

$$\beta = \frac{m-\rho}{\|\mathbf{y}-\mathbf{f}(\hat{\mathbf{w}},\mathbf{X})\|^2},$$

а порядок вычисления матрицы $\mathbf{A}$ будет состоять из $n$ независимых уравнений:

$$\alpha_j = \frac{\beta h_j}{2}\left(-1+\sqrt{1+\frac{4}{\beta h_j\|\hat{\mathbf{w}}\|^2}}\right).$$

Полученные выше результаты позволяют сформулировать теорему. Пусть вектор параметров $\mathbf{w}_0 = [w_{1(0)},\ldots,w_{n(0)}]^\intercal$ фиксирован.

**Теорема 8.** *В окрестности вектора параметров $\mathbf{w}_0$ оценка ковариационных матриц $\mathbf{A}^{-1},\mathbf{B}^{-1}$ для гипотезы нормального распределения зависимой переменной имеет вид*

$$\alpha_i = \frac{1}{2}\lambda_i\left(\sqrt{1+\frac{4}{(w_i-w_{i(0)})^2\lambda_i}}-1\right),\ \textit{где}\ \lambda_i = \beta\mathbf{diag}(h_i),$$

$$\beta = \frac{m-\gamma}{2(\mathbf{f}-\mathbf{y})^\intercal\mathbf{B}'(\mathbf{f}-\mathbf{y})},\ \textit{где}\ \gamma = \sum_{j=1}^{W}\frac{\lambda_j}{\lambda_j+\alpha_j}.$$

*Последовательности $\|\mathbf{A}_{k+1}-\mathbf{A}_k\|^2$ и $\|\beta_{k+1}-\beta_i\|^2$ монотонно убывают с увеличением номера шага $k$.*

### 4.2.5. Получение оценок для линейной модели

В случае линейной модели,

$$\mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{X}\mathbf{w},$$

все формулы можно выписать явно, избавив себя от необходимости решать численно оптимизационные задачи. Так, интеграл от экспоненты функции ошибки в точности равен

$$\int \exp\big(-S(\mathbf{w})\big) d\mathbf{w} = S(\hat{\mathbf{w}})(2\pi)^{\frac{n}{2}} (\det \mathbf{H}^{-1})^{\frac{1}{2}},$$

где $\hat{\mathbf{w}}$ — единственная точка максимума унимодальной функции ошибки $S(\mathbf{w})$, а гессиан

$$\mathbf{H} = \mathbf{A} + \beta \mathbf{X}^\intercal \mathbf{X}.$$

При этом для точки максимума $\hat{\mathbf{w}}$, являющейся точкой наиболее вероятных параметров

$$\hat{\mathbf{w}} = \arg\max p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}),$$

справедливо выражение

$$\hat{\mathbf{w}} = (\mathbf{A} + \beta \mathbf{X}^\intercal \mathbf{X})^{-1} \beta \mathbf{X}^\intercal y.$$

В частности, для случая диагональной матрицы $\mathbf{A} = \operatorname{diag}(\alpha_j)$ можно выписать явные формулы оценки структурных параметров:

$$\beta = \frac{m - \rho}{\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2},$$

где вспомогательная переменная $\rho$ зависит от параметров $\alpha_j$ и $\beta$:

$$\rho = \sum_{j=1}^{n} \frac{\beta h_j}{\alpha_j + \beta h_j},$$

а искомый параметр $\alpha_j$ выражен как

$$\alpha_j = \frac{\beta h_j}{2}\big(-1 + \sqrt{1 + \frac{4}{\beta h_j \|\hat{\mathbf{w}}\|^2}}\big).$$

Здесь $h_j$ является $j$-м собственным числом гессиана $\mathbf{H}$ функции ошибки $S(\mathbf{w})$, а в случае линейной модели — $j$-м собственным числом матрицы $\mathbf{X}^\intercal \mathbf{X}$.

### 4.2.6. Вычисление гессиана

В нелинейном случае гессиан приходится определять численными методами. Для этого используются два метода: метод аппроксимации вычисления вторых производных функции ошибки и метод приближения ошибки квадратичной поверхностью.

**Метод разностной аппроксимации вычисления гессиана.** Элемент $h_{jk}$ гессиана $H$ в точке $\hat{\mathbf{w}}$ вычисляется по формуле

$$h_{jk} = \frac{\partial^2 S}{\partial w_j \partial w_k} = \frac{S(\hat{\mathbf{w}} + (\mathbf{e}_j + \mathbf{e}_k)r) - S(\hat{\mathbf{w}} + \mathbf{e}_j r) - S(\hat{\mathbf{w}} + \mathbf{e}_k r) + S(\hat{\mathbf{w}})}{r^2},$$

где $\mathbf{e}_j, \mathbf{e}_k$ — единичные векторы, $r$ — малый параметр. Погрешность этой формулы имеет порядок $O(r)$. Данный метод требует вычисления функции ошибки в $\frac{n(n+1)}{2}$ точках и является вычислительно эффективным.

**Метод приближения ошибки квадратичной поверхностью.** Предлагаемый метод является менее вычислительно эффективным, однако более устойчивым. Метод основан на том, в окрестности оптимальной точки $\hat{\mathbf{w}}$ генерируется множество $\mathbb{W}$ размера $K$, состоящее из векторов $\mathbf{w}$, близких к $\hat{\mathbf{w}}$. Для каждого из этих $\mathbf{w} \in \mathbb{W}$ вычисляется значение функции ошибки $S(\mathbf{w})$.

Таким образом, составляется обучающая выборка $(\mathbb{W}, \mathbf{y}) = \{(\mathbf{w}_l, y_l)\}_{i=1}^{K}$ размера $K$, где

$$y_l = 2\left(S(\mathbf{w}_l) - S(\hat{\mathbf{w}})\right).$$

В точке оптимума $\hat{\mathbf{w}}$ функция ошибки может быть приближена квадратичной поверхностью, поэтому согласно модели:

$$y_l = (\mathbf{w}_l - \hat{\mathbf{w}})^{\intercal}\mathbf{H}(\mathbf{w}_l - \hat{\mathbf{w}}).$$

Получив таким образом $K$ уравнений, параметры $h_{jk}$ находятся методом наименьших квадратов. Отметим, что для устойчивого решения необходимо значение $K \gg n^2$, то есть метод требует больших вычислительных затрат, однако является устойчивым при больших $K$.

### 4.2.7. Аппроксимация Лапласа для оценки нормирующего коэффициента

Эмпирическая плотность распределение $p^*(\mathbf{w}|\mathfrak{D})$, описанная в предыдущем разделе, не является плотностью распределения случайной величины, поскольку ее интеграл не равен единице. Так как гипотеза порождения данных предполагает, что параметров обобщенно-линейных моделей распределены нормально, предлагается аппроксимировать эмпирическое распределение $p^*$ нормальным, теоретическим, ниже оно обозначается $\hat{p}$, найдя при этом нормирующий множитель, обеспечивающий равенство интеграла единице. Вариант этого метода для оценки маргинальных распределений опубликован в [277].

Рассмотрим абсолютно непрерывную многомерную случайную величину $\mathbf{w}$ и ее плотность распределения $p(\mathbf{w})$,

$$p(\mathbf{w}) = \frac{1}{Z_{\mathbf{w}}}p^*(\mathbf{w}),$$

включающая нормирующий множитель $Z_{\mathbf{w}}$ заданного эмпирического распределения $p^*(\mathbf{w})$,

$$Z_{\mathbf{w}} \stackrel{\text{def}}{=} \int\limits_{\mathbf{w}\in\mathbb{W}} p^*(\mathbf{w})d\mathbf{w}$$

Нормирующий множитель $Z_{\mathbf{w}}$ неизвестен, требуется его оценить. Предполагается, что $p^*(\mathbf{w})$ имеет моду многомерной случайной величины $\mathbf{w}$ в точке $\mathbf{w}_0$, см. рис 46. Для оценки константы $Z_{\mathbf{w}}$ используем приближение функции $p^*(\mathbf{w})$ нормальным распределением $\hat{p}(\mathbf{w})$, максимум которого совпадает с модой распределения $p^*(\mathbf{w})$. Найдем моду $p^*(\mathbf{w})$, а именно точку $\mathbf{w}_0$, в которой $\nabla_{\mathbf{w}}p(\mathbf{w}) = \mathbf{0}$, то есть

$$\left.\frac{\partial p(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_0} = \mathbf{0}.$$

Для нахождения $p(\mathbf{w}) = Z_{\mathbf{w}}^{-1}p^*(\mathbf{w})$ прологарифмируем и разложим $p^*(\mathbf{w})$ в ряд Тейлора в окрестности предполагаемого максимума $\mathbf{w}_0$. Так как $\ln p^*(\mathbf{w})$ есть монотонная функция от $p^*(\mathbf{w})$, то аргумент ее максимума равен аргументу максимума $p^*(\mathbf{w})$.

$$\ln p^*(\mathbf{w}) = \ln p^*(\mathbf{w}_0) + 0 - \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)A(\mathbf{w} - \mathbf{w}_0) + \cdots, \tag{144}$$

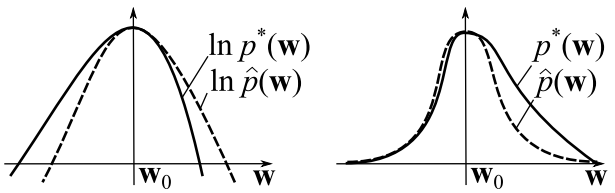



Рис. 46. Приближение эмпирического распределения теоретическим с целью нормировки.

где $(n \times n)$-матрица Гессе

$$\mathbf{A} = [\alpha_{ij}], i, j \in \mathcal{J}, \quad |\mathcal{J}| = n,$$

определена как

$$\alpha_{ij} = -\left.\frac{\partial^2 \ln p^*(\mathbf{w})}{\partial w_i \partial w_j}\right|_{\mathbf{w}=\mathbf{w}_0},$$

кратко,

$$\mathbf{A} = -\nabla\nabla \ln p^*(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_0}, \quad \text{здесь } \nabla \text{ — градиент функции.}$$

Взяв экспоненту разложения, получим

$$p^*(\mathbf{w}) \approx p^*(\mathbf{w}_0) \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^\intercal \mathbf{A}(\mathbf{w}-\mathbf{w}_0)\right).$$

Тогда нормальное распределение $\hat{p}(\mathbf{w})$, приближающее нормированное распределение $p(\mathbf{w})$ имеет вид

$$\hat{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \mathbf{A}^{-1}) = \frac{1}{(2\pi)^{\frac{n}{2}} \det^{-\frac{1}{2}} \mathbf{A}} \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^\intercal \mathbf{A}(\mathbf{w}-\mathbf{w}_0)\right),$$

а нормировочная константа для $p^*(\mathbf{w})$

$$Z_{\mathbf{w}} \approx p^*(\mathbf{w}_0) \frac{(2\pi)^{\frac{n}{2}}}{\det^{\frac{1}{2}} \mathbf{A}}. \tag{145}$$

### 4.2.8. Метод Монте-Карло сэмплирования функции ошибки

Ниже описана процедура сэмплирования параметров модели при фиксированных структурных параметрах. Аппроксимируется интеграл значений правдоподобия по сэмплированным параметрам. Оптимальными структурными параметрами считаются те, которые доставляют максимум аппроксимирующей функции.

Для того, чтобы оценить структурные параметры $\mathbf{A}$ и $\mathbf{B}$, согласно байесовскому выводу, требуется максимизировать интеграл:

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w}, \mathbf{B}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w} \to \max_{\mathbf{A}\in\mathbb{M}^n, \mathbf{B}\in\mathbb{M}^m}. \tag{146}$$

Далее, будем рассматривать случай, когда структурный параметр $\mathbf{A}$ является матрицей, обратной к матрице вторых моментов $\mathbf{\Sigma}$ случайного вектора $\mathbf{w}$, $\mathbf{A} = \mathbf{\Sigma}^{-1}$. Без ограничения общности, примем $\mathsf{E}(\mathbf{w}) = \mathbf{0}$. Этот случай обобщает предположения, введенные нами в предыдущем разделе, о гипотезе нормального распределения вектора $\mathbf{w}$.

Отметим, что в наших предположениях задано евклидово пространство случайных векторов $\mathbf{w}$ матрицей Грама $\mathbf{A}^{-1}$. Поскольку матрица $\mathbf{A}^{-1}$ является симметричной положительно определенной матрицей, для нее существует и единственно разложение Холецкого [63]:

$$\mathbf{A}^{-1} = \mathbf{R}^\intercal \mathbf{R}, \tag{147}$$

где $\mathbf{R}$ — верхняя треугольная матрица со строго положительными элементами на диагонали. Отметим также, что $\mathbf{R}$ является матрицей перехода из евклидового пространства случайных векторов $\mathbf{w}_0 \sim p(\mathbf{w}_0|\mathbf{\Sigma}_0)$ с матрицей ковариаций, или матрицей Грама, $\mathbf{\Sigma}_0 = \mathbf{I}$, в пространство векторов $\mathbf{w}$ с матрицей ковариаций $\mathbf{\Sigma}$.

В силу существования и единственности разложения Холецкого (147) матрицы $\mathbf{A}^{-1}$ будем искать оптимум (146) в виде

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w},\mathbf{B})p(\mathbf{w}|\mathbf{R})d\mathbf{w} \to \max_{\mathbf{R},\mathbf{B}}.$$

Для дальнейших упрощений ограничим общность нашей задачи, зафиксировав матрицу $\mathbf{B} = \mathbf{B}_0$. Будем искать решение в виде

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w},\mathbf{B}_0)p(\mathbf{w}|\mathbf{R})d\mathbf{w} \to \max_{\mathbf{R}}. \tag{148}$$

Поскольку интеграл (148) нельзя вычислить аналитически, применим стохастический метод интегрирования по пространству параметров $\mathbb{W}$. Для этого заметим, что выражение (148) является математическим ожиданием правдоподобия данных:

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w},\mathbf{B}_0)p(\mathbf{w}|\mathbf{R})d\mathbf{w} \ = \mathsf{E}\left(p(\mathfrak{D}|\mathbf{w},\mathbf{B}_0)\right)$$

и, согласно закону больших чисел,

$$\int\limits_{\mathbf{w}\in\mathbb{W}} p(\mathfrak{D}|\mathbf{w},\mathbf{B}_0)p(\mathbf{w}|\mathbf{R})d\mathbf{w} \approx \frac{1}{K}\sum_{\mathbf{w}\in\mathbb{W}(\mathbf{R})} p(\mathfrak{D}|\mathbf{w},\mathbf{B}_0),$$

где $\mathcal{W}(\mathbf{R})$ — множество мощности $K$ векторов $\mathbf{w}$ с матрицей ковариаций $\mathbf{R}^\intercal\mathbf{R}$, которое может быть получено в результате процедуры сэмплирования.

Запишем оценку правдоподобия модели, которую необходимо максимизировать по параметру $\mathbf{R}$:

$$\mathcal{E}(\mathbf{R}) \approx \frac{1}{K}\sum_{\mathbf{w}\in\mathbb{W}(\mathbf{R})} p(\mathfrak{D}|\mathbf{w},\mathbf{B}_0) \to \max_{\mathbf{R}}. \tag{149}$$

Таким образом, для нахождения оптимальных параметров итерируемого значения матрицы $\mathbf{R}$ оптимизационной задачи (149) для каждого $\mathbf{R}$ необходимо провести процедуру сэмплирования параметров $\mathcal{W}(\mathbf{R})$. Однако, как было отмечено ранее, матрица $\mathbf{R}$ является матрицей перехода при преобразовании евклидового пространства с матрицей Грама $\mathbf{I}$ в евклидово пространство с матрицей Грама $\mathbf{R}^\intercal\mathbf{R}$.

Это означает, что достаточно провести процедуру сэмплирования однократно перед запуском алгоритма оптимизации, получив множество

$$\mathcal{W}_0 = \mathcal{W}(\mathbf{I}) = \{\mathbf{w}_0|\mathbf{w}_0 \sim p(\mathbf{w}_0|\mathbf{I})\}.$$

Затем, на каждом шаге алгоритма, получать множество $\mathcal{W}(\mathbf{R})$ преобразованием множества $\mathcal{W}_0$ по правилу

$$\mathcal{W}(\mathbf{R}) = \{\mathbf{R}^\intercal \mathbf{w}_0 | \mathbf{w}_0 \in \mathcal{W}_0\}.$$

**Алгоритм Метрополиса-Гастингса порождения выборки.** Для порождения выборки $\mathcal{W}_0 = \{\mathbf{w}|\mathbf{w} \sim p(\mathbf{w}|\mathbf{I})\}$ используется алгоритм Метрополиса-Гастингса.

Основной идеей алгоритма является сэмплирование выборки, которая образует цепь Маркова, в которой каждый элемент выборки $\mathbf{w}_{t+1}$ коррелирует только с предыдущим элементом выборки $\mathbf{w}_t$.

Для работы алгоритма Метрополиса-Гастингса введем вспомогательное распределение $Q(\mathbf{w}|\mathbf{w}')$, выберем начальный элемент $\mathbf{w}_0$ и положим $\mathcal{W}_0 = \{\mathbf{w}_0\}$. Далее, пусть выбран элемент $\mathbf{w}_t$ согласно распределению $Q(\mathbf{w}'|\mathbf{w}_t)$. Следующий элемент $\mathbf{w}'$ генерируется случайным образом. Затем рассчитывается число $a$ — вероятность включения элемента $\mathbf{w}'$ в выборку $\mathcal{W}_0$.

$$a = \min_{\mathbf{w}' \in \mathbb{R}^n} \left( \frac{p(\mathfrak{D}|\mathbf{w}', \mathbf{B}_0) Q(\mathbf{w}_t|\mathbf{w}')}{p(\mathfrak{D}|\mathbf{w}_t, \mathbf{B}_0) Q(\mathbf{w}'|\mathbf{w}_t)}, 1 \right).$$

С вероятностью $a$ новый элемент $\mathbf{w}'$ становится элементом $t+1$ выборки $\mathcal{W}_0$, иначе элемент $\mathbf{w}'$ отклоняется, и процедура шага $t+1$ повторяется заново:

$$\mathbf{w}_{t+1} = \begin{cases} \mathbf{w}', \text{ с вероятностью } a, \\ \mathbf{w}_t, \text{ с вероятностью } 1-a. \end{cases}$$

Вспомогательное распределение $Q(\mathbf{w}|\mathbf{w}')$ примем нормальным:

$$Q(\mathbf{w}|\mathbf{w}') = Q(\mathbf{w}'|\mathbf{w}) = \frac{1}{(2\pi\alpha^{-1})^{\frac{n}{2}}} \exp\left( -\frac{\alpha}{2} (\mathbf{w} - \mathbf{w}')^\intercal (\mathbf{w} - \mathbf{w}') \right).$$

То есть, функция $Q(\mathbf{w}|\mathbf{w}')$ является симметричной, и

$$a = \frac{p(\mathfrak{D}|\mathbf{w}', \mathbf{B}_0)}{p(\mathfrak{D}|\mathbf{w}_t, \mathbf{B}_0)}.$$

Начальный элемент $\mathbf{w}_0$ выбирается случайным образом образом из распределения $P(\mathbf{w}|\mathbf{I})$.

Таблица 12. Анализ ошибок: относительное смещение оценок.

| | Scalar | | Diag | | Full | |
|---|---|---|---|---|---|---|
| | $\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$ | $\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$ | $\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$ | $\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$ | $\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$ | $\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$ |
| OLS | 0.3 | - | 0.67 | - | 0.37 | - |
| LA | 0.095 | **0.14** | 0.54 | 1.09 | - | - |
| MK | 0.078 | 0.16 | **0.52** | **0.36** | **0.34** | 0.57 |
| CV | **0.041** | 0.39 | 0.53 | 0.42 | 0.36 | **0.55** |

### 4.2.9. Оценка структурных параметров методом скользящего контроля

В рамках этого методы предполагается, что реализации случайной величины $\mathbf{w}$ заданы составом регрессионной выборки. Каждая реализация является оптимальным значением вектора параметров $\mathbf{w}$ на соответствующей подвыборке. Будем строить оценку среднего риска

$$L(\mathbf{w}) = \mathsf{E}_{\mathfrak{D}}\left(S_{\mathfrak{D}}(\mathbf{w})\right),$$

где

$$S_{\mathfrak{D}}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{f})^{\mathsf{T}}(\mathbf{y}-\mathbf{f}),$$

согласно (138). Отметим, что в данном случае функция $S_{\mathfrak{D}}(\mathbf{w})$ является частью первого слагаемого функции ошибки $S(\mathbf{w})$ в выражении (137):

$$S(\mathbf{w}) = \beta S_{\mathfrak{D}}(\mathbf{w}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w},$$

где второе слагаемое $\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w}$ отвечает за априорное распределение параметров модели $\mathbf{w}$.

Согласно [131], для оценивания среднего риска $L(\mathbf{w})$ разделим выборку $\mathfrak{D}$ на $Q$ непересекающихся блоков

$$\mathfrak{D} = \mathfrak{D}_1^{l_1} \sqcup \ldots \sqcup \mathfrak{D}_Q^{l_Q}$$

одинаковых, или почти одинаковых, мощностей $l_1, \ldots, l_Q$ соответственно. Обозначим $\hat{\mathbf{w}}_{\mathfrak{D}\setminus\mathfrak{D}_q}(\mathbf{A})$ оценку вектора параметров $\mathbf{w}$ путем минимизации функции ошибки (137) на обучающей подвыборке $\mathfrak{D}\setminus\mathfrak{D}_q$ при фиксированной матрице $\mathbf{A}$. Будем минимизировать оценку среднего риска (обозначим функцию CV — Cross-Validation)

$$\mathrm{CV}(\mathfrak{D}, \mathbf{A}) = \frac{1}{m}\sum_{i=1}^{m} S_{\mathfrak{D}_q}(\hat{\mathbf{w}}_{\mathfrak{D}\setminus\mathfrak{D}_q}(\mathbf{A})) \to \min_{\mathbf{A}\in\mathbb{M}^n},$$

где $S_{\mathfrak{D}_q}(\hat{\mathbf{w}}_{\mathfrak{D}\setminus\mathfrak{D}_q}(\mathbf{A}))$ — функция ошибки, оцененная на контрольной подвыборке $\mathfrak{D}_q$ при векторе параметров $\hat{\mathbf{w}}$, оцененном на обучающей подвыборке $\mathfrak{D}\setminus\mathfrak{D}_q$ при фиксированной матрице $\mathbf{A}$. Отметим, что в данном случае матрица $\mathbf{B}$ фиксируется, и оптимизация проводится только по элементам матрицы $\mathbf{A}$.

### 4.2.10. Анализ метода оценки ковариационных матриц

Предложенные алгоритмы протестированы на синтетических и реальных данных. Приведем список данных с подробным описанием.

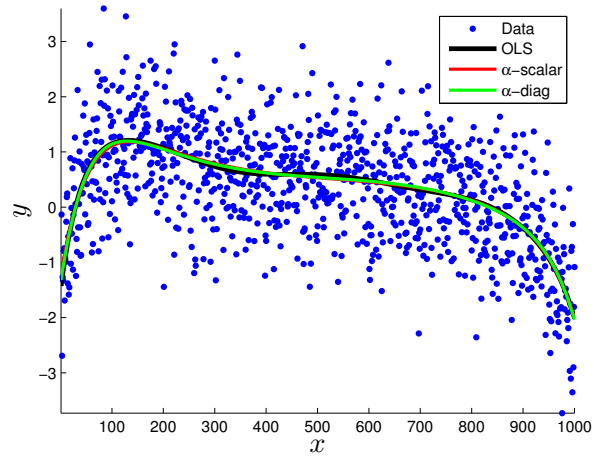
**Линейная полиномиальная модель.** Матрица $\mathbf{X}$ представляет собой набор полиномов: столбец $i$ матрицы $\mathbf{X}$ является набором точек полинома $x^{i-1}$ из отрезка $[-1, 1]$. Первый столбец матрицы состоит из всех единиц.

Используется линейная модель для генерирования зависимой переменной:

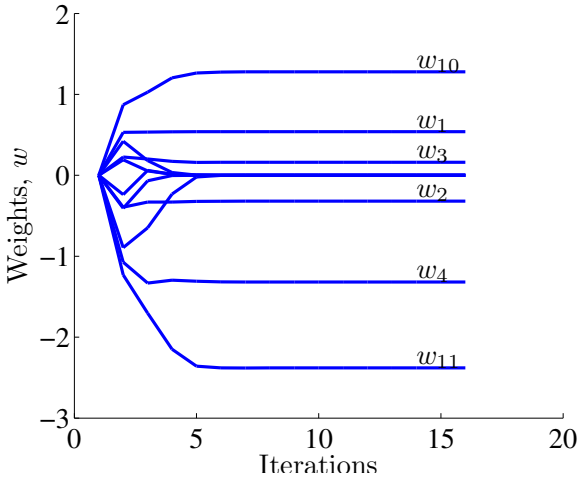
$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},$$

$$\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{B}).$$

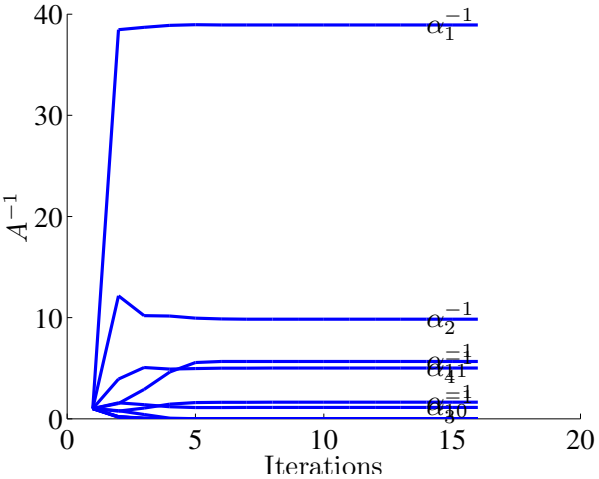
Матрицы $\mathbf{A}$ и $\mathbf{B}$ являются диагональными или скалярными матрицами и генерируются по закону гамма-распределения с параметрами (1,1).



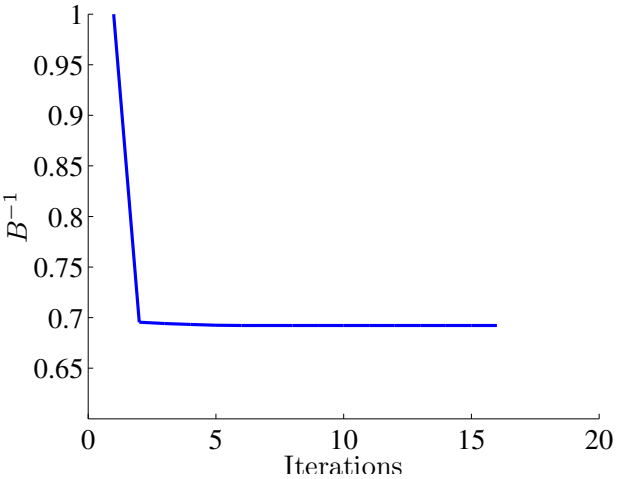

(a) Сгенерированные данные




(b) Сходимость параметров $\mathbf{w}$

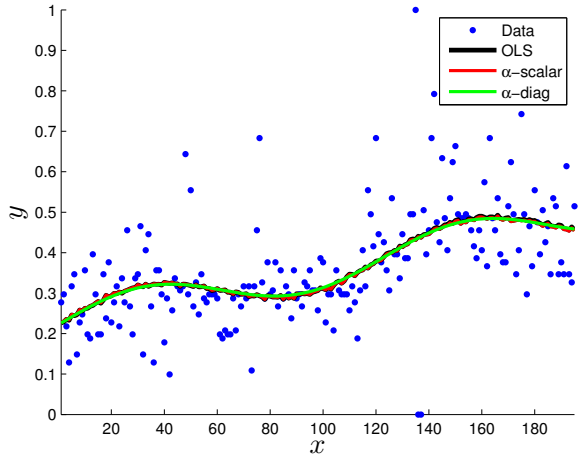



(c) Сходимость структурных параметров $\mathbf{A}^{-1}$

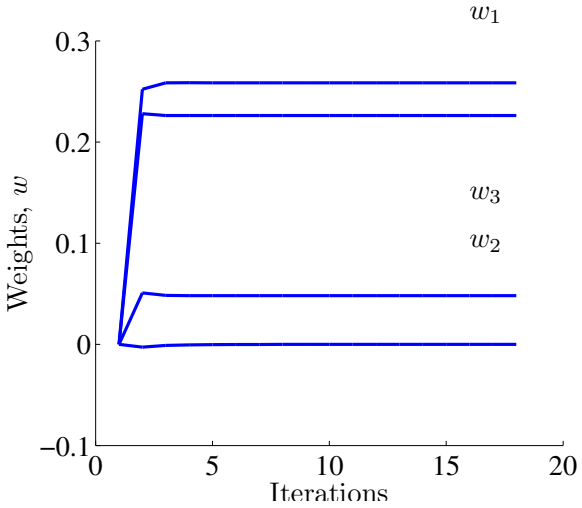



(d) Сходимость структурных параметров $\mathbf{B}^{-1}$

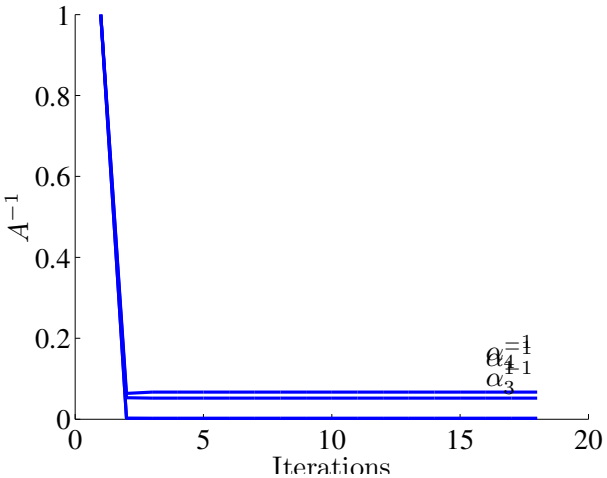
Рис. 47. Метод аппроксимации Лапласа для линейной полиномиальной модели в случае диагональной матрицы $\mathbf{A}$.



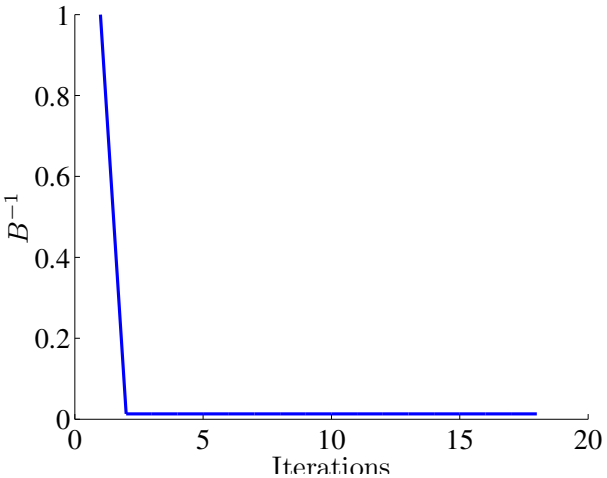

(a) Аппроксимация данных




(b) Сходимость параметров $\mathbf{w}$

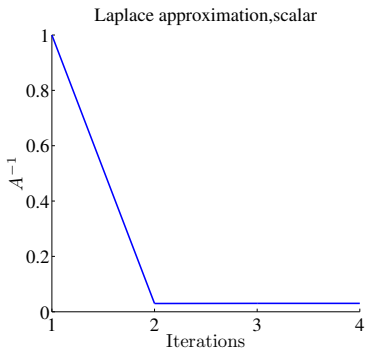



(c) Сходимость структурных параметров $\mathbf{A}^{-1}$

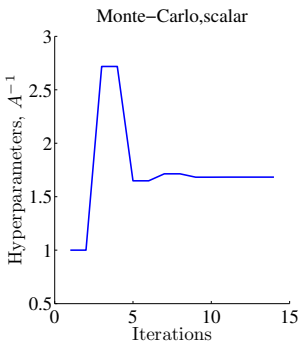



(d) Сходимость структурных параметров $\mathbf{B}^{-1}$

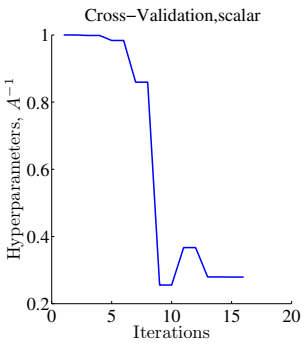
Рис. 48. Прогнозирование цен на хлеб, метод аппроксимации Лапласа в случае диагональной матрицы $\mathbf{A}$.




(a) Аппроксимация Лапласа

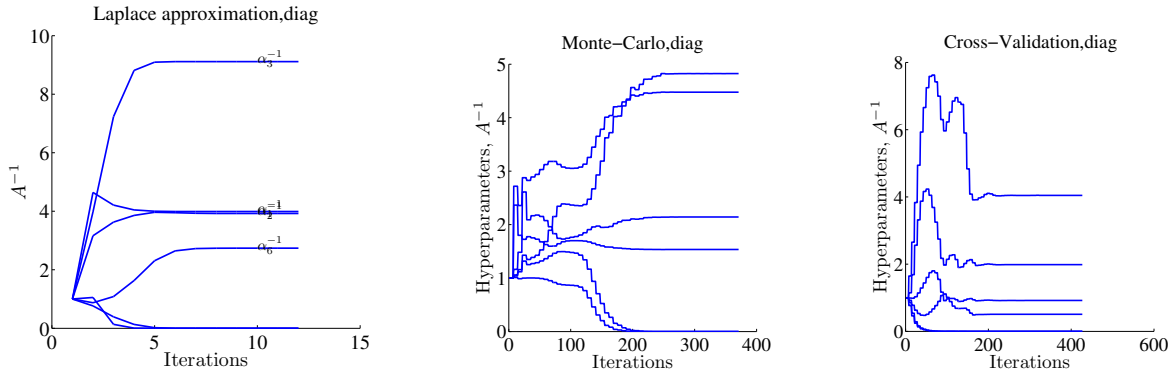



(b) Монте-Карло



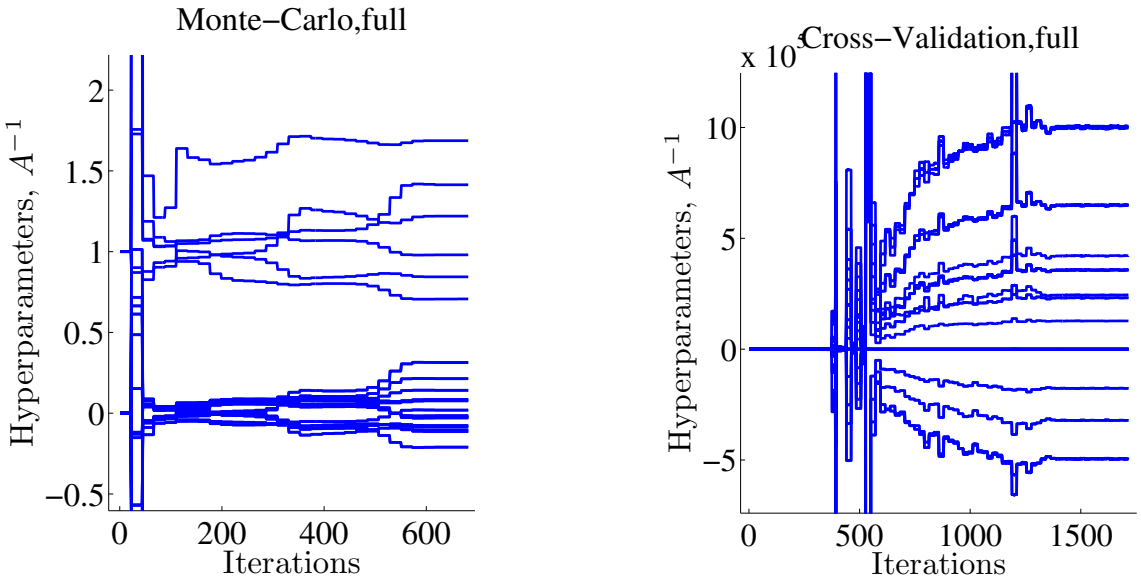

(c) Скользящий контроль

Рис. 49. Сходимость структурных параметров $\mathbf{A}^{-1}$ в скалярном случае, $\mathbf{A} = \alpha\mathbf{I}$.




(a) Аппроксимация Лапласа (b) Монте-Карло (c) Скользящий контроль

Рис. 50. Сходимость структурных параметров $\mathbf{A}^{-1}$ в диагональном случае.




(a) Монте-Карло (b) Скользящий контроль

Рис. 51. Сходимость структурных параметров $\mathbf{A}^{-1}$.

**Реальные данные: цены на хлеб.** Предложенные алгоритмы протестированы на синтетических и реальных данных. Ниже показаны графики сходимости оценок параметров и структурных параметров к оптимальным значениями $\hat{\mathbf{w}}, \hat{\mathbf{A}}$ сравнение их с истинными оценками $\mathbf{w}^*, \mathbf{A}^*$. Выполнен анализ ошибок. Синтетические данные представляют собой выборку, сгенерированную линейной полиномиальной моделью:

$$y = \sum_{j=0}^{n} w_j x^j + \varepsilon,$$

где

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^*), \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{B}^*) = \mathcal{N}(\mathbf{0}, \beta^* \mathbf{I}).$$

Элементы матриц $\mathbf{A}^*$ и $\mathbf{B}^*$ являются наперед заданными величинами.

С помощью предложенных алгоритмов получены оценки $\hat{\mathbf{A}}$ матрицы $\mathbf{A}^*$ и соответствующего ей оптимального вектора параметров $\hat{\mathbf{w}}$. В случае аппроксимации Лапласа, получена также оценка $\hat{\mathbf{B}}$ матрицы $\mathbf{B}^*$. В случае методов Монте-Карло и скользящего контроля матрица $\mathbf{B}^*$ подается в качестве входного параметра.

Результаты работы алгоритмов представлены в таблице 12 в виде нормы относительного отклонения оптимального вектора параметров от истинного значения, $\frac{\|\hat{\mathbf{w}}-\mathbf{w}^*\|}{\|\mathbf{w}^*\|}$, а также в виде нормы относительного отклонения оптимального значения матрицы $\mathbf{A}$ от ее истинного значения, $\frac{\|\hat{\mathbf{A}}-\mathbf{A}^*\|}{\|\mathbf{A}^*\|}$. В первой строке матрицы записаны результаты OLS — метода наименьших квадратов оценки вектора параметров $\mathbf{w}$. Жирными выделены значения, наиболее близкие к истинному вектору параметров $\mathbf{w}^*$ и матрице $\mathbf{A}^*$. Из таблицы видно, что алгоритмы возвращают сравнимые результаты.

Сходимости структурных параметров, элементов матрицы $\mathbf{A}^{-1}$, для всех трех алгоритмов показаны в скалярном случае на рис. 49, в диагональном случае — на рис. 50, в общем случае — на рис. 51. По оси абсцисс этих графиков отложены итерации процедуры, по оси ординат — значения элементов матрицы $\mathbf{A}$.

Из графиков видно, что в скалярном случае (на рис. 49) сходимость наступает после 10-20 итераций. Для диагонального (рис. 50) и полного (рис. 51) случаев требуется гораздо больше итераций. Особой интерес представляет собой диагональный случай (рис. 50), в котором появляются нулевые диагональные элементы матрицы $\mathbf{A}^{-1}$. Появление нулевого элемента $\alpha_j$ на диагонали матрицы $\mathbf{A}$ означает, что накладывается очень большой штраф на элемент вектора параметров $w_j$, и оптимальное значение компоненты $j$ вектора параметров $\hat{\mathbf{w}}$ аннулируется, что свидетельствует о неинформативности признака $j$. Во всех трех случаях, алгоритмы выделили два из шести признаков (четвертую и пятую степени полинома) как неинформативные, уменьшив таким образом сложность модели.

## 4.3. Оценка гиперпараметров для случая линейных моделей

Для линейных регрессионных моделей предлагается явно оценить оптимальное значение гиперпараметров, используя функцию правдоподобия модели. Полученные оценки гиперпараметров могут быть использованы для оценки параметров модели и отбора признаков [56, 191, 176, 49, 46, 347, 343]. Предложенный подход сравнивается с подходом, использу-

ющим аппроксимацию Лапласа распределения параметров модели [347] и методом наименьших углов [82]. Приведенная ниже теорема опубликована в работе [305].

**Теорема 9.** *Правдоподобие в предположениях о нормальном распределении шума* $\boldsymbol{\varepsilon}$ *и параметров модели* $\mathbf{w}$ *(см. табл. 1) имеет вид*

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}|\mathbf{K}|^{\frac{1}{2}}}\exp\left(\frac{1}{2}\mathbf{y}^{\intercal}(\mathbf{C}^{\intercal}\mathbf{K}\mathbf{C}-\mathbf{B})\mathbf{y}\right), \tag{150}$$

*а его логарифм имеет вид*

$$\ln p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = -\frac{1}{2}\left(\ln|\mathbf{K}| + m\ln 2\pi - \ln|\mathbf{B}| - \ln|\mathbf{A}| - \mathbf{y}^{\intercal}(\mathbf{C}^{\intercal}\mathbf{K}\mathbf{C}-\mathbf{B})\mathbf{y}\right). \tag{151}$$

*Здесь*

$$\mathbf{K} = \mathbf{X}^{\intercal}\mathbf{B}\mathbf{X} + \mathbf{A}, \quad \mathbf{C} = \mathbf{K}^{-1}\mathbf{X}^{\intercal}\mathbf{B}.$$

*Доказательство.* Как было сказано ранее,

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \int_{\mathbb{R}^n}\frac{1}{(2\pi)^{\frac{m}{2}}|\mathbf{B}|^{-\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^{\intercal}\mathbf{B}(\mathbf{y}-\mathbf{X}\mathbf{w})\right)\frac{1}{(2\pi)^{\frac{n}{2}}|\mathbf{A}|^{-\frac{1}{2}}}\exp\left(-\frac{1}{2}\mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}\right)d\mathbf{w}.$$

Переписав произведение двух экспонент как экспоненту от их суммы, получаем

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \int_{\mathbb{R}^n}\frac{|\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}}\exp\left(-\frac{1}{2}\left((\mathbf{y}-\mathbf{X}\mathbf{w})^{\intercal}\mathbf{B}(\mathbf{y}-\mathbf{X}\mathbf{w}) + \mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}\right)\right)d\mathbf{w}.$$

Раскроем скобки:

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \int_{\mathbb{R}^n}\frac{|\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}}\exp\left(-\frac{1}{2}\left(\mathbf{w}^{\intercal}\mathbf{X}^{\intercal}\mathbf{B}\mathbf{X}\mathbf{w} - 2\mathbf{w}^{\intercal}\mathbf{X}^{\intercal}\mathbf{B}\mathbf{y} + \mathbf{y}^{\intercal}\mathbf{B}\mathbf{y} + \mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}\right)\right)d\mathbf{w}.$$

Введем обозначения $\mathbf{K} = \mathbf{A} + \mathbf{X}^{\intercal}\mathbf{B}\mathbf{X}, \mathbf{C} = \mathbf{K}^{-1}\mathbf{X}^{\intercal}\mathbf{B}$ и выделим полный квадрат по $(\mathbf{w}-\mathbf{C}\mathbf{y})$:

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \int_{\mathbb{R}^n}\frac{|\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n+m}{2}}}\exp\left(-\frac{1}{2}\left((\mathbf{w}-\mathbf{C}\mathbf{y})^{\intercal}\mathbf{K}(\mathbf{w}-\mathbf{C}\mathbf{y}) - \mathbf{y}^{\intercal}(\mathbf{C}^{\intercal}\mathbf{K}\mathbf{C}-\mathbf{B})\mathbf{y}\right)\right)d\mathbf{w}.$$

Учитывая, что интеграл по плотности многомерного нормального распределения равен единице, получаем

$$p(\mathfrak{D}|\mathbf{A},\mathbf{B}) = \frac{|\mathbf{B}|^{\frac{1}{2}}|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{m}{2}}|\mathbf{K}|^{\frac{1}{2}}}\exp\left(\frac{1}{2}\left(\mathbf{y}^{\intercal}(\mathbf{C}^{\intercal}\mathbf{K}\mathbf{C}-\mathbf{B})\mathbf{y}\right)\right).$$

Следовательно, искомое правдоподобие модели $p(\mathfrak{D}|\mathbf{A},\mathbf{B})$ имеет вид (150), а его логарифм — вид (151). □

Рассмотрим теперь случай, когда матрица $\mathbf{A}$ — диагональная, а матрица $\mathbf{B} = \beta\mathbf{I}$.

**Следствие 1.** *Если матрица* $\mathbf{A}$ *— диагональная, а матрица* $\mathbf{B}$ *имеет вид* $\mathbf{B} = \beta\mathbf{I}$*, то логарифм правдоподобия модели* $\ln p(\mathfrak{D}|\mathbf{A},\mathbf{B})$ *имеет вид*

$$\ln p(\mathfrak{D}|\mathbf{A},\beta) = -\frac{1}{2}\left(\ln|\mathbf{K}| + m\ln 2\pi - m\ln\beta - \ln|\mathbf{A}| - \beta\mathbf{y}^{\intercal}(\beta\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^{\intercal}-\mathbf{I})\mathbf{y}\right),$$

*где* $\mathbf{K} = \mathbf{A} + \beta\mathbf{X}^{\intercal}\mathbf{X}$.

### 4.3.1. Вычисление производной функции правдоподобия модели

Для поиска максимума правдоподобия будем пользоваться градиентными методами оптимизации [327], поэтому нам понадобится выражения для производных $\ln p(\mathfrak{D}|\mathbf{A},\mathbf{B})$ по гиперпараметрам $\mathbf{A},\mathbf{B}$.

Пусть матрица $\mathbf{A}$ имеет вид $\mathbf{A} = \{\alpha_{ij}\}, i,j = \overline{1,n}$, а матрица $\mathbf{B}$ имеет вид $\mathbf{B} = \{\beta_{ij}\}, i,j = \overline{1,m}$. Обе матрицы являются симметричными и неотрицательно определенными, так как являются матрицами ковариации.

Верны следующие два свойства производных матриц [232]. Для симметричной матрицы $\mathbf{M}$ верно, что

$$\frac{\partial \ln|\mathbf{M}|}{\partial t} = \operatorname{tr}\left(\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial t}\right),$$

где $t$ — некоторый параметр, $\mathbf{M} = \mathbf{M}(t)$. Так же верно, что

$$\frac{\partial \mathbf{M}^{-1}}{\partial t} = -\mathbf{M}^{-1}\frac{\partial \mathbf{M}}{\partial t}\mathbf{M}^{-1}.$$

Введем обозначение $\mathbf{S}^{ij}$ — такая матрица, что для двух индексов $k,l$ выполнено, что

$$S^{ij}_{kl} = \begin{cases} 1, & k = i, l = j \text{ или } k = j, l = i, \\ 0, & \text{иначе.} \end{cases}$$

Запишем производную $\ln p(\mathfrak{D}|\mathbf{A},\beta)$ по $\beta_{ij}$:

$$\begin{aligned}\frac{\partial \ln p(\mathfrak{D}|\mathbf{A},\mathbf{B})}{\partial \beta_{ij}} = -\frac{1}{2}\Big(&\operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{X}^\intercal\mathbf{S}^{ij}\mathbf{X}\right) - \operatorname{tr}\left(\mathbf{B}^{-1}\mathbf{S}^{ij}\right) - \\ &\mathbf{y}^\intercal\big(\mathbf{S}^{ji}\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\intercal\mathbf{B} + \mathbf{B}^\intercal\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\intercal\mathbf{S}^{ij} - \\ &\mathbf{B}^\intercal\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\intercal\mathbf{S}^{ij}\mathbf{X}\mathbf{K}^{-T}\mathbf{X}^\intercal\mathbf{B} \\ &- \mathbf{S}^{ij}\big)\mathbf{y}\Big).\end{aligned}$$

Аналогично запишем производную $\ln p(\mathfrak{D}|\mathbf{A},\beta)$ по $\alpha_{ij}$:

$$\begin{aligned}\frac{\partial \ln p(\mathfrak{D}|\mathbf{A},\mathbf{B})}{\partial \alpha_{ij}} = -\frac{1}{2}\Big(&\operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{S}^{ij}\right) - \operatorname{tr}\left(\mathbf{A}^{-1}\mathbf{S}^{ij}\right) + \\ &\mathbf{y}^\intercal\mathbf{B}^\intercal\mathbf{X}\mathbf{K}^{-1}\mathbf{S}^{ij}\mathbf{K}^{-T}\mathbf{X}^\intercal\mathbf{B}\mathbf{y}\Big).\end{aligned}$$

Так же запишем производные в предположениях следствия 1,

$$\mathbf{A} = \operatorname{diag}(\alpha_1,\alpha_2,\ldots,\alpha_n), \quad \mathbf{B} = \frac{1}{\beta}\mathbf{I}:$$

$$\begin{aligned}\frac{\partial \ln p(\mathfrak{D}|\mathbf{A},\beta)}{\partial \beta} = -\frac{1}{2}\Big(&\operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{X}^\intercal\mathbf{X}\right) - \frac{m}{\beta} + \\ &\mathbf{y}^\intercal(2\beta\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\intercal - \mathbf{I} - \beta^2\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\intercal\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^\intercal)\mathbf{y}\Big)\end{aligned}$$

$$\frac{\partial \ln p(\mathfrak{D}|\mathbf{A},\beta)}{\partial \alpha_i} = -\frac{1}{2}\left(\operatorname{tr}\left(\mathbf{K}^{-1}\mathbf{I}^{ii}\right) - \frac{1}{\alpha_i} - \beta^2\mathbf{y}^\intercal\mathbf{X}\mathbf{K}^{-1}\mathbf{I}^{ii}\mathbf{K}^{-1}\mathbf{X}^\intercal\mathbf{y}\right).$$

Так как получены значения производных правдоподобия модели $\ln p(\mathfrak{D}|\mathbf{A},\mathbf{B})$ по гиперпараметрам $\mathbf{A},\mathbf{B}$, можно использовать любой градиентный метод оптимизации для поиска гиперпараметров $\mathbf{A},\mathbf{B}$, максимизирующх правдоподобие модели.

Будем говорить, что задана модель линейной регрессии $f$, если задано подмножество признаков $\mathcal{I}_f \subseteq \mathcal{I} = \{1, 2, \ldots, n\}$. Полученные значения гиперпараметров $\alpha_i$ для диагональной матрицы $A$ могут быть использованы для отбора признаков и выбора модели линейной регрессии. Параметры $\mathbf{w}_i$ модели $f$ сравниваются, используя оценки значений гиперпараметров $\alpha_i$. Большие значения гиперпараметра $\alpha_i$ означают большой штраф на значение параметра и, следовательно, меньшую значимость параметров модели. Малые значения $\alpha_i$ показывают большую значимость данного компонента модели для ее качества.

### 4.3.2. Отбор шумовых и коррелирующих признаков

Результатом вычислительного эксперимента является отбор шумовых и коррелирующих признаков. Тестирование алгоритма производится на временном ряде продаж нарезного хлеба в зависимости от времени. Ряд содержит 195 записей. Модель, аппроксимирующая ряд: $\mathbf{y} = 0.2256 + 0.1996\boldsymbol{\xi} + 0.0496\sin(10\boldsymbol{\xi})$, где $\boldsymbol{\xi} \in \mathbb{R}^n$ — регрессионная выборка. Введем следующие обозначения: $\boldsymbol{\xi}_0$, $\boldsymbol{\xi}^1$ — значение каждого элемента выборки в нулевой и первой степени соответственно, $\sin(10\boldsymbol{\xi})$ — поэлементное применение элементарной функции к вектору $\boldsymbol{\xi}$.

Пусть матрица плана $\mathbf{X}$ представлена в следующем виде $\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n]$, где $\boldsymbol{\chi} \in \mathbb{R}^m$. В данном случае она состоит из трёх столбцов: $\boldsymbol{\xi}_0$, $\boldsymbol{\xi}^1$, $\sin(10\boldsymbol{\xi})$.

**Отбор шумовых признаков.** Шумовая выборка сформирована при помощи добавления столбца случайных чисел с нормальным распределением. Модель, аппроксимирующая данные в эксперименте: $\mathbf{y} = w_1\boldsymbol{\chi}_1 + w_2\boldsymbol{\chi}_2 + w_3\boldsymbol{\chi}_3 + w_4\boldsymbol{\chi}_4$, где $\boldsymbol{\chi}_1 = \boldsymbol{\xi}_0$, $\boldsymbol{\chi}_2 \sim \mathcal{N}(0, 2)$, $\boldsymbol{\chi}_3 = \boldsymbol{\xi}^1$, $\boldsymbol{\chi}_4 = \sin(10\boldsymbol{\xi})$. При наличии в выборке шумового элемента процедура сходится за восемь итераций. На рис. 52 проиллюстрированы изменения матрицы Гессе $\mathbf{H}$ на каждом шаге процедуры.

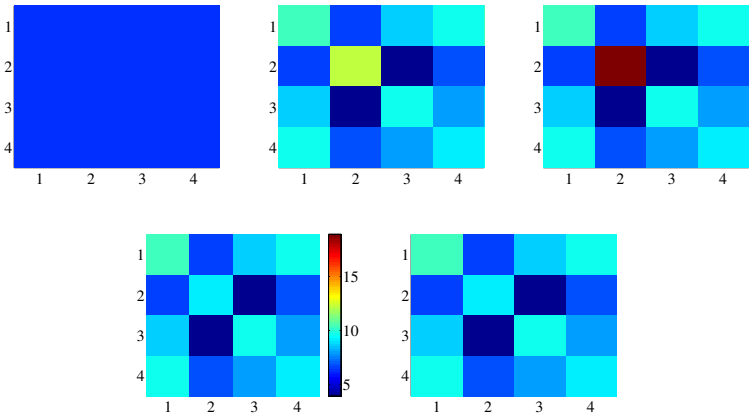
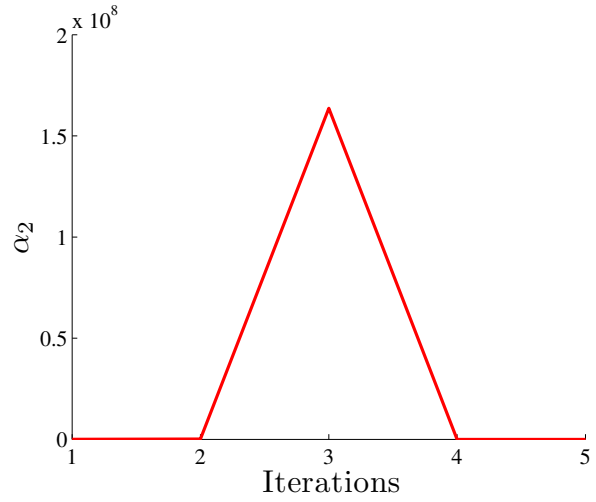


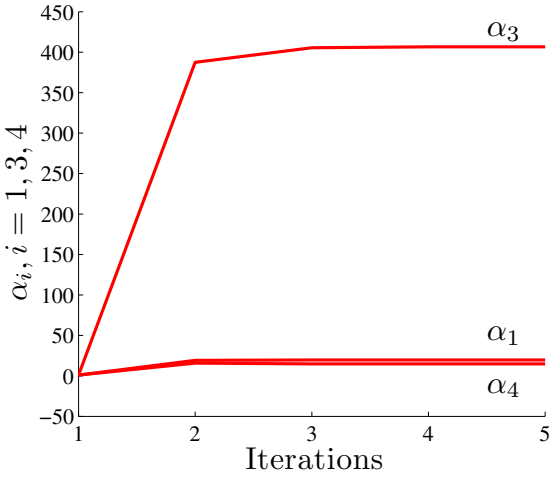

Рис. 52. Итерационная процедура вычисления матрицы Гессе, случай шумового параметра.

На второй итерации наблюдается резкое отличие диагонального элемента $(2, 2)$. В течение итераций 2 и 3 он продолжает возрастать, пока не достигает критической относитель-

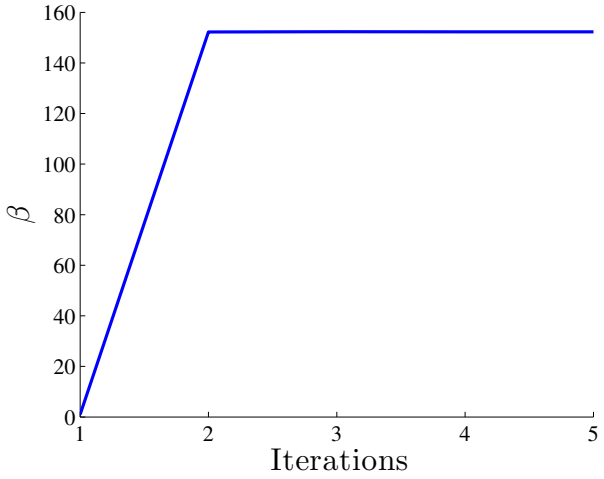
ной величины (принята эмпирическая оценка отношения максимального элемента матрицы к минимальному $10^6$). Далее на 4-й итерации выполняется его зануление. Таким образом происходит выявление шумового признака.



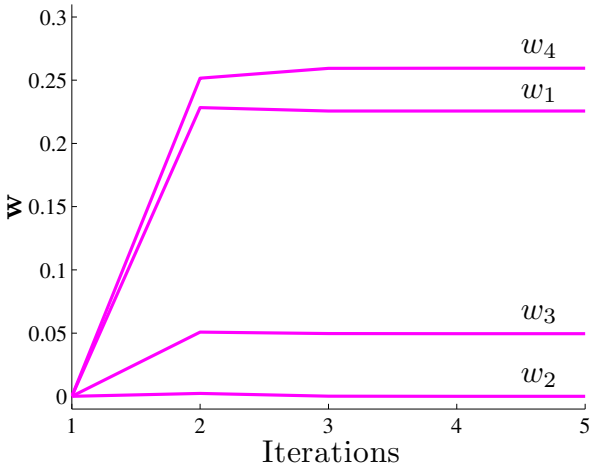

(a) Элемент матрицы **A**, соответствующий шумовому параметру модели




(b) Элементы матрицы **A**, соответствующие нешумовым параметрам модели




(c) Скалярный гиперпараметр $\beta$ (случай шумового параметра)




(d) Параметры модели **w** (случай шумового параметра)

Рис. 53. Сходимость гиперпараметров: шумовые параметры.

На рис. 53a и 53b представлены диагональные элементы матрицы **A**. Первый график иллюстрирует изменения второго диагонального элемента $\alpha_2$, который соответствует шумовому параметру модели. Резкий скачок объясняется тем, что на данной итерации алгоритм находится вблизи локального минимума $\mathbf{w}_0$ и, несмотря на возрастание диагональных элементов матрицы **H**, знаменатель формулы (135) мал. Далее происходит зануление элементов матрицы Гессе и соответствующий гиперпараметр $\alpha$ становится равным нулю.

На графиках рис. 53c и 53d представлены скалярный гиперпараметр $\beta$ и процедура изменения параметров модели $w_i$ соответственно.

**Отбор коррелирующих признаков.** Выборка с коррелирующими признаками сформирована при помощи добавления в матрицу плана столбца $1.3\boldsymbol{\chi}_2$. Таким образом, модель, аппроксимирующая данные в эксперименте: $y = w_1\boldsymbol{\chi}_1 + w_2\boldsymbol{\chi}_2 + w_3\boldsymbol{\chi}_3 + w_4\boldsymbol{\chi}_4$, где $\boldsymbol{\chi}_1 = \boldsymbol{\xi}_0$, $\boldsymbol{\chi}_2 = \boldsymbol{\xi}^1$, $\boldsymbol{\chi}_3 = 1.3\boldsymbol{\xi}^1$, $\boldsymbol{\chi}_4 = \sin(10\boldsymbol{\xi})$. На рис. 54 поэлементно проиллюстрирована матрица Гессе $\mathbf{H}$.

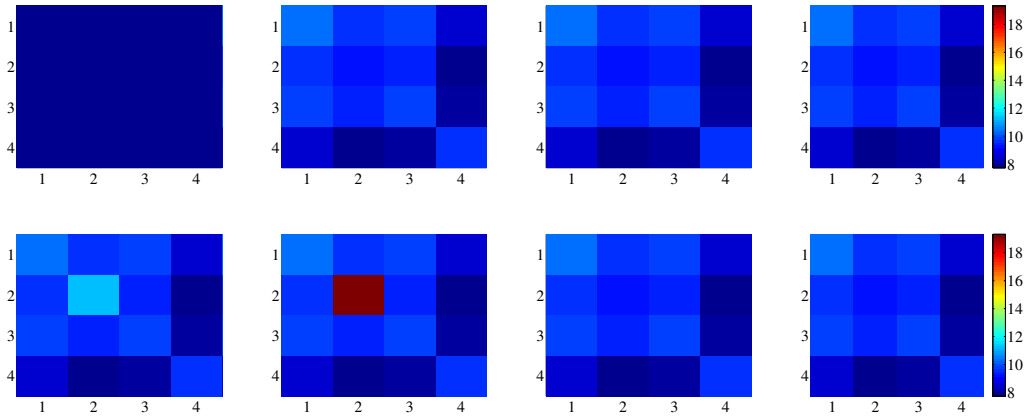


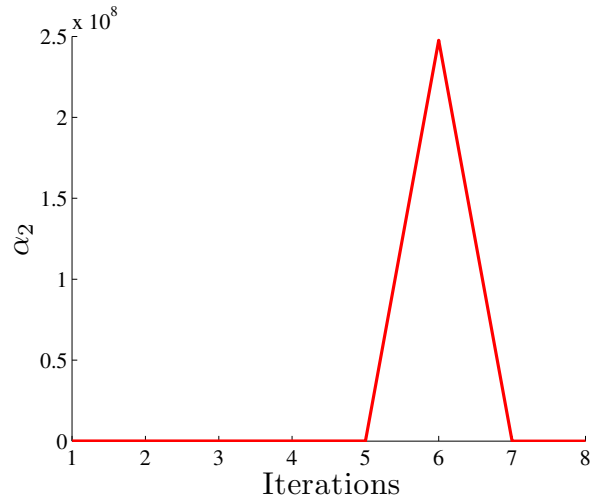
Рис. 54. Итерационный процедура вычисления матрицы Гессе, случай коррелирующих параметров.

При наличии коррелирующих признаков также наблюдается возрастание диагональных элементов. Это происходит из-за того, что алгоритм выбирает ближайший вектор $\boldsymbol{\chi}$ к вектору $\mathbf{y}$ (в пространстве векторов матрицы $\mathbf{X}$), а коррелирующий с ним считает шумовым. На графиках рис. 55a и 55b представлены диагональные элементы матрицы $\mathbf{A}$.
На рис. 55c представлены изменения скалярного гиперпараметра $\beta$. На рис. 55d представлены изменения параметров модели $w_i$ в течение итерационной процедуры. Коррелирующий параметр $w_2$ сначала возрастает, а затем стремится к нулю. Это происходит из-за того, что пространство параметров модели многоэкстремально.

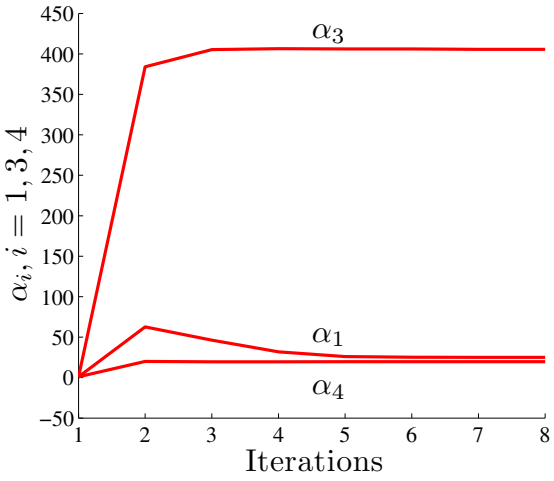
В работе предложен способ отсеивания шумовых и коррелирующих признаков, а также алгоритм оценки ковариационной матрицы параметров модели. Данный алгоритм имеет следующие преимущества перед методами, описанными во введении: 1) нет необходимости разделения данных на обучающую и контрольную выборку; 2) алгоритм не содержит никаких параметров, которые необходимо оценивать или задавать дополнительно (как, например, в методах регуляризации); 3) добиваясь сходимости как параметров, так и гиперпараметров, предложенный алгоритм повышает устойчивость выбранной регрессионной модели.

## 4.4. Выбор многоуровневых моделей

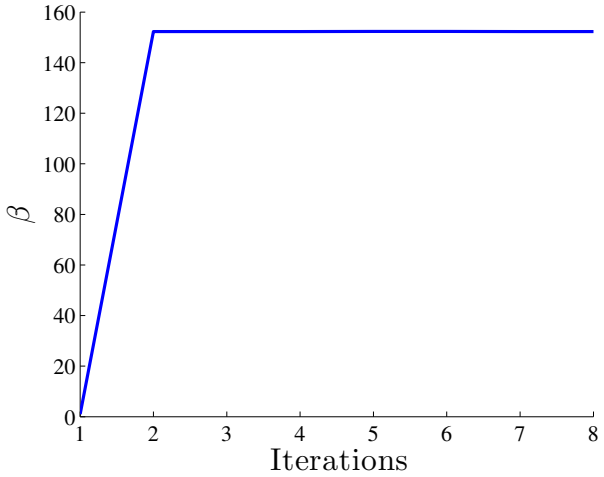
Обсуждается метод выбора активного набора признаков и фильтрации объектов выборки при восстановлении регрессии. Предполагается, что элементы рассматриваемой выборки естественным образом были разбиты на подмножества; для каждого из которых имеется своя, отличная от других, гипотеза порождения данных. Задача заключается в том, чтобы определить это разбиение и восстановить регрессионную модель для каждой подвыборки.



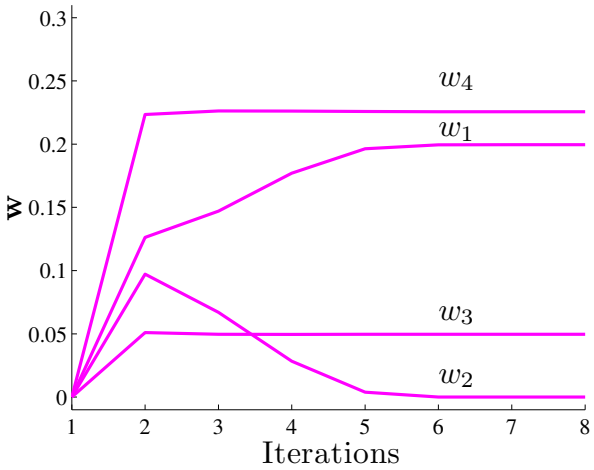

(a) Элементы матрицы **A**, соответствующие независимым параметрам модели




(b) Элемент матрицы **A**, соответствующий коррелирующему параметру модели




(c) Скалярный гиперпараметр $\beta$ (случай зависимых параметров)




(d) Вектор параметров модели **w** (случай зависимых параметров)

Рис. 55. Сходимость гиперпараметров: коррелирующие параметры.

При этом оценивается ковариационная матрица параметров каждой модели, и на основании анализа этой матрицы определяется вероятность принадлежности некоторого объекта данной подвыборке, а и некоторого признака — данной модели.

Работа опирается на следующие результаты. Предположим, что измеряемых свободных переменных недостаточно для восстановления адекватной регрессионной модели. Для пополнения их набора используем порождающие функции и вводим при этом меры их структурной сложности, аналогичные предложенным Е. Владиславлевой [273].

В работе мы исходим из того, что процедура скользящего контроля недостаточно эффективна при решении прикладных задач. В случае, когда число измеряемых или порожденных признаков многократно превосходит объем выборки, однократное разбиение выборки не исключает переобучения модели и приводит к тому, что выборку приходится разбивать на несколько подвыборок: обучающую, тестовую, контрольную и так далее, как показано С. Ватанабе [274] и С. Арло [25].

Для выбора адекватной регрессионной модели используется функция правдоподобия модели, см. Д. МакКай [188]. Эта функция является составной частью связанного байесовского вывода, см. К. Бишоп [47]. Её использование согласуется с принципом минимальной длины описания, являющимся универсальным критерием выбора модели, см. П. Грюнвальд [116, 120]. Для оценки вероятности принадлежности признаков и объектов выборки к тем или иным моделям используются методы анализа ковариационных матриц, рассмотренные Дж. Нельдером [177]. Для оценки сходства двух и более моделей используется расстояние Дженсена-Шеннона, см. [180].

Предлагаемый метод заключается в следующем. Фиксируется класс моделей; порождается множество производных признаков. Индексы элементов выборки разбиваются на подмножества. Каждое из подмножеств соответствует модели. Число моделей выбирается таким, чтобы расстояние между моделями было статистически значимым [180]. Принадлежность элемента выборки к модели определяется по результатам анализа ковариационной матрицы зависимых переменных. Структура модели определяется по результатам анализа ковариационной матрицы параметров модели.

Результатом является многоуровневая модель оптимальной сложности — набор адекватных регрессионных моделей, описывающих выборку. В качестве иллюстрации приведена задача прогнозирования периодических временных рядов.

### 4.4.1. Выбор модели и фильтрация объектов

Линейная модель $f$ однозначно задается активным множеством индексов признаков $\mathbf{A} \subseteq \mathcal{J}$. Предполагая частичную гомоскедаксичность выборки (например, среди объектов встречаются выбросы, которые должны быть исключены из рассмотрения), зададим «фильтрованную» выборку, иначе — активное множество объектов индексами $\mathbf{B} \subseteq \mathcal{I}$. Обозначим множество многомерных величин $\{\mathbf{x}^i | i \in \mathbf{B}\}$ как $\mathbf{x}^{\mathbf{B}}$. Задача выбора модели имеет вид

$$\mathfrak{F} \ni \hat{f} = \arg \max_{\mathbf{A} \subseteq \mathcal{J}, \mathbf{B} \subseteq \mathcal{I}} \mathcal{E}\left(f(\mathbf{w}_{\mathbf{A}}, \mathbf{x}^{\mathbf{B}})\right). \tag{152}$$

Способы решения этой задачи рассмотрены автором в [262]. Заметим, что для набору индексов признаков $\mathcal{J}$ мощности $n$ соответствуют $2^n$ вершин двоичного куба. Каждая вершина

задает некоторый активный набор признаков $\mathbf{A}$: считается, что $j$-й признак вошел в набор, если значение $j$-й координаты вершины единица. При решении задачи мы руководствуемся следующими предположениями:

1) среди вершин куба существует по крайней мере одна, обозначим ее $\hat{\mathbf{A}}$, доставляющая матожидание правдоподобия модели,

2) от вершины $\mathbf{A} = \emptyset$ к вершине $\hat{\mathbf{A}}$ есть путь по ребрам куба (иначе — стратегия последовательного добавления-удаления признаков), который доставляет правдоподобию модели $\mathcal{E}\left(f(\mathbf{w}_\mathbf{A}, \mathbf{x})\right)$ сходимость по вероятности.

Множество индексов $\mathbf{B}$ задает выпуклую комбинацию $\{x_i | i \in \mathbf{B}\}$ — область $\mathcal{X}_\mathbf{A}$, «по крайней мере», в которой значения дисперсии $\{\beta_i | i \in \mathbf{B}\}$ зависимых переменных $\{\mathbf{y}_i | i \in \mathbf{B}\}$ меняются «незначительно». Другими словами, третий центральный момент, или коэффициент асимметрии случайной величины $\mathbf{y}$, соответствующей области $\mathcal{X}_\mathbf{A}$ равен нулю [235].

### 4.4.2. Алгоритм выбора многоуровневых моделей

*Многоуровневой* [235, 180, 177, 120, 243, 262, 264] моделью $\mathfrak{f}$ называется набор моделей $\mathfrak{f} = \{f_k | f \in \mathfrak{F}\}$, $k = 1, \ldots, l$, такой, что

$$f_k : \mathcal{W}_k \times \mathcal{X}_{\mathbf{B}_k} \to \mathcal{Y}_{\mathbf{B}_k},$$

при разбиении $\mathcal{I} \supseteq \mathbf{B}^* = \sqcup \mathbf{B}_k$.

Введем функцию расстояния $\rho(f_k, f_l)$ между двумя моделями. Для этого используем дивергенцию Дженсена-Шеннона, в которой $\rho_{kl} \in [0, 1]$ является метрикой [180]:

$$\rho(p_k \| p_l) = 2^{-1} D_{\mathrm{KL}}\left(p_k \| p'\right) + 2^{-1} D_{\mathrm{KL}}\left(p' \| p_l\right),$$

где $p' = 2^{-1}(p_k + p_l)$ и здесь $p_k \stackrel{\text{def}}{=} (p(\mathbf{w}_\mathbf{A} | \mathfrak{D}, \mathbf{A}, \mathbf{B}, f_k)$. Несимметричная функция расстояния — дивергенция Кулльбака-Лейблера задана как

$$D_{\mathrm{KL}}\left(p \| p'\right) = \int\limits_{\mathbf{w} \in \mathbb{W}} p'(\mathbf{w}) \ln \frac{p(\mathbf{w})}{p'(\mathbf{w}) d\mathbf{w}}.$$

Отметим, что расстояние вводится только на моделях, имеющих одинаковый набор признаков $\mathbf{A}$.

Задача нахождения многоуровневых моделей ставится следующим образом:

$$\mathfrak{F} \supset \hat{\mathfrak{f}} = \arg \max_{\mathbf{B}_1, \mathbf{B}_2 \subset \mathbf{B}} \rho(f_1, f_2) \tag{153}$$

при заданном множестве индексов признаков $\hat{\mathbf{A}}$, таком, что

$$\hat{\mathbf{A}} = \arg \max_{\mathbf{A} \subseteq \mathcal{J}} \mathcal{E}\left(f_1(\mathbf{w}_\mathbf{A}, \mathbf{x}^{\mathbf{B}_1})\right) \mathcal{E}\left(f_2(\mathbf{w}'_\mathbf{A}, \mathbf{x}^{\mathbf{B}_2})\right).$$

## 4.5. Маргинальные смеси моделей

### 4.5.1. Смеси линейных моделей

Рассмотрим $K$ линейных моделей, каждая из которых имеет параметры $\mathbf{w}_k \in \mathbb{R}^n$.

Предположим, что для каждой модели дисперсия регрессионных остатков равна $\beta$. Тогда распределение зависимой переменной $y$ для смеси нормальных распределений может быть записано в виде

$$p(y|\boldsymbol{\theta}) = \sum_{k=1} K\pi_k \mathcal{N}(y|\mathbf{w}_k^\mathsf{T}\mathbf{x}, \beta^{-1}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)}.$$

Здесь вектор параметров $\boldsymbol{\theta}$ есть набор всех параметров данного приложения, присоединенных векторов

$$\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}, \beta]^\mathsf{T},$$

в котором

$\mathbf{w}_1, \ldots, \mathbf{w}_k$ — параметров каждой из $k$ моделей,

$\boldsymbol{\pi} = [\pi_1, \ldots, \pi_k]$ — весов моделей,

$\beta$ — структурного параметра.

Логарифм функции правдоподобия предыдущего выражения при заданной выборке $\mathfrak{D} = \{(y^i, \mathbf{x}^i)\} = (\mathbf{y}, \mathbf{X})$ имеет вид

$$\ln p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i=1}^{m} \ln \left( \sum_{k=1}^{K} \pi_k \mathcal{N}(y|\mathbf{w}_k^\mathsf{T}\mathbf{x}^i, \beta^{-1}) \right).$$

Для максимизации этой функции будем использовать EM-алгоритм для смесей нормального распределения (безусловных). Используем набор $Z = \{\mathbf{z}^1, \ldots \mathbf{z}^m\}$ векторов скрытых переменных, где $\mathbf{z}^i \in \{0,1\}^K$. Все компоненты вектора $\mathbf{z}^i = [z_1^i, \ldots, z_k^i]^\mathsf{T}$ равны нулю, кроме одной, например, с номером $k$. Равенство этой компоненты единице означает, что данный элемент выборки принадлежит $k$-й модели.

Логарифм функции правдоподобия совместного распределения переменных $\mathbf{y}, Z$ имеет вид

$$\ln p(\mathbf{y}, Z|\boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{k=1}^{K} z_k^i \ln \left( \pi_k \mathcal{N}(y^i|\mathbf{w}_k^\mathsf{T}\mathbf{x}^i, \beta^{-1}) \right).$$

Для оценки вектора параметров $\boldsymbol{\theta}$ и матрицы $\mathbf{Z}$, описывающей принадлежность объектов моделям, используется EM-алгоритм. Назначаются начальные параметры $\boldsymbol{\theta}_0$ На E-шаге алгоритма эти параметры используются для вычисления вероятности принадлежности каждого элемента выборки одной из $K$ моделей. Введем матрицу $\boldsymbol{\Gamma}$, состоящую из элементов $\gamma_k^i$, которые интерпретируются как (математическое ожидание принадлежности $i$-ого элемента выборки $j$-й модели,

$$\gamma_k^i = \mathsf{E}(z_k^i) = p(k|\mathbf{x}^i, \boldsymbol{\theta}_0) = \frac{\pi_k \mathcal{N}(y^i|\mathbf{w}_k^\mathsf{T} x^i, \beta^- 1)}{\sum_{k'} \pi'_k \mathcal{N}(y^i|\mathbf{w}_k^\mathsf{T} x^i, \beta^- 1).}$$

Полученный результат является апостериорное вероятностями того, что каждый $i$-й элемент выборки порожден $k$-й моделью.

Используем матрицу $\mathbf{\Gamma} = [\gamma_k^i]$ для определения принятого апостериорное распределения $p(Z|\mathbf{y}, \boldsymbol{\theta}_0)$ функции правдоподобия общего вида, которая записывается как

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mathsf{E}_Z(\ln p(\mathbf{y}, Z|\boldsymbol{\theta})) = \sum_{i\in\mathcal{I}}\sum_{k=1}^{K} \gamma_k^i \left(\ln \pi_k + \ln \mathcal{N}(y^i|\mathbf{w}_k^\intercal \mathbf{x}^i, \beta^{-1}).\right)$$

На М-шаге алгоритма максимизируем функцию $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ относительно $\boldsymbol{\theta}$ при фиксированных значениях матрицы $\mathbf{\Gamma}$. При оптимизации относительно коэффициентов $\pi_k$, включенных в вектор $\boldsymbol{\theta}$, требуется соблюдение условия нормировки коэффициентов $\sum_{k=1}^{K} \pi_k = 1$, которое может быть получено путем введения множителя Лагранжа. При этом весовые коэффициенты моделей заданы в виде

$$\pi_k = \frac{1}{n}\sum_{i=1}^{m} \gamma_k^i.$$

Максимизируем предыдущее выражение относительно вектора параметров $\mathbf{w}_k$ модели с номером $k$. Делая подстановку нормального распределения, видим, что функция $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ принимает вид

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \sum_{i\in\mathcal{I}} \gamma_k^i \left(-\frac{\beta}{2}(y^i - \mathbf{w}_k^\intercal \mathbf{x}^i)^2\right) + \text{const}.$$

Константа в данном выражении означает вклад в функцию $Q$ параметров $\mathbf{w}_k \neq \mathbf{w}_j$ моделей с индексами, отличными от $k$. Таким образом, максимизируется взвешенная сумма квадратов регрессионных остатков одной-единственной модели. При этом каждому элементу выборки с номером $i$ соответствует весовой коэффициент $\beta\gamma_k^i$

### 4.5.2. Смеси обобщенно-линейных моделей

Оценка принадлежности каждого элемента выборки одной из моделей производится аналогично предыдущему примеру, с учетом иной гипотезы порождения данных. Изменим предположение о нормальном распределении многомерной случайной величины — зависимой переменной на предположение о биномиальном распределении. Тогда условное распределение этой переменной для вероятностной смеси из $K$ моделей будет иметь вид

$$p(y|\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \sigma(\mathbf{w}_k^\intercal \mathbf{x})_k^y (1 - \sigma(\mathbf{w}_k^\intercal \mathbf{x})_k)^{1-y}$$

Здесь вектор параметров $\boldsymbol{\theta}$ есть набор всех параметров данного приложения, присоединенных векторов

$$\boldsymbol{\theta} = [\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}]^\intercal,$$

в котором

$\mathbf{w}_1, \ldots, \mathbf{w}_k$ — параметров каждой из $k$ моделей,

$\boldsymbol{\pi} = [\pi_1, \ldots, \pi_k]$ — весов моделей.

Для заданной выборки $\{y^i, \mathbf{x}^i\}$ функция правдоподобия имеет вид

$$p(\mathbf{y}|\theta) = \prod_{i\in\mathcal{I}} \left( \sum_{k=1}^{K} \pi_k \sigma_i k^y (1-\sigma_{ij}^{1-y^i}), \right)$$

здесь $\sigma_i k = \sigma(\mathbf{w}_k^\intercal \mathbf{x}^i)$ — логистическая функция активации в данной модели, $\sigma(\cdot) = 1/(1 + \exp(-1\cdot)$

Максимизируем эту функцию правдоподобия итеративно с помощью EM-алгоритма, использую введенную ранее скрытую переменную $z_{nk}$. Полная функция правдоподобия для смеси из $K$ моделей имеет вид

$$p(\mathbf{y}, Z|\boldsymbol{\theta}) = \prod_{i\in\mathcal{I}} \prod k = 1^K \left( \pi_k \sigma_{ik}^{y^i} (1 - \sigma_{ik}\ 1 - y^i) \right)\ z_k^i,$$

здесь $Z$ — матрица латентных переменных с элементами $z_k^i$. Зададим начальное значение вектора параметров $\boldsymbol{\theta}_0$. На E-шаге этот вектор используется для нахождения логарифмической функции правдоподобия, которая от этого вектора зависит, заданной как

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \mathsf{E}_Z \left( \ln p(\mathbf{y}, Z|\boldsymbol{\theta}) = \right) \sum_{i\in\mathcal{I}} \sum_{k=1}^{K} \gamma_{nk} \left( \ln \pi_k + y_i \ln \sigma_{ik} + (1-y_i) \ln(1-\sigma_{ik}) \right).$$

На M-шаге функция правдоподобия максимизируется относительно $\boldsymbol{\theta}$ при заданном $\boldsymbol{\theta}_0$, значения матрицы Г зафиксированы. Как и ранее, максимизация функции относительно $\boldsymbol{\theta}_k$ может быть выполнена с использованием множителя Лагранжа, для выполнения условия нормировки $\sum_{k=1}^{K} \pi_k = 1$. При этом значения весового коэффициента $\pi$ вычисляются как

$$\pi_k = \frac{1}{m} \sum_{i\in\mathcal{I}} \gamma_i k.$$

Оценим набор параметров $\{w_k\}, k = 1, \ldots, K$ смеси моделей. Заметим, что логарифмическая функция правдоподобия $Q(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ включает для каждого индекса $k$ только один из векторов $\{\mathbf{w}_k\}$. То есть, различные векторы $\mathbf{w}_k$ не связаны на M-шаге алгоритма. На этом шаге решение может быть получено методом итеративного перевзвшивания наименьших квадратов.

Градиент и гессиан вектора параметров $\mathbf{w}_k$ задан выражением

$$\nabla_k Q = \sum_{i\in\mathcal{I}} \gamma_{ik} (y_i - \sigma_{ik}) \mathbf{x}^i,$$

$$\mathbf{H}_k = -\nabla_k \nabla_k Q = \sum_{i\in\mathcal{I}} \gamma_{ik} \sigma_{ik} (1-\sigma_{ik}) \mathbf{x}^i \mathbf{x}^{i\intercal}.$$

Здесь $\nabla_k$ обозначает градиент для вектора параметров $\mathbf{w}_k$. Для фиксированного значения $\gamma_{ik}$ градиенты не зависят от параметров $\{\mathbf{w}_l\}$, $k \neq l$, то есть имеется возможность получить решение для каждого вектора $\mathbf{w}_k$ с помощью алгоритма итеративного перевзвешивания. Это означает, что на M-шаге происходит оценка параметров каждой из моделей логистической регрессии, независимо от остальных. При этом каждому элементу выборки поставлен в соответствие вес $\gamma_{ik}$.

### 4.5.3. Иллюстрация: прогнозирование периодических временных рядов

Примем следующую гипотезу порождения данных: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B})$ из которой следует $\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\text{МР}}, \mathbf{A})$. Тогда, при отсутствии гипотезы гомоскедаксичности регрессионных остатков и независимости элементов многомерной случайной величины $\mathbf{y}$, оптимизируемая функция $S$ будет иметь вид

$$\begin{aligned} 2S(\mathbf{w}|\mathfrak{D}, f) =& (\mathbf{w} - \mathbf{w}_{\text{МР}})^{\intercal}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\text{МР}}) + \\ &+ (\mathbf{f} - \mathbf{y})^{\intercal}\mathbf{B}(\mathbf{f} - \mathbf{y}). \end{aligned} \tag{154}$$

Учитываются также следующие предположения:

1) существуют несколько типов периодов, каждый из которых должен быть спрогнозирован своей собственной моделью,

2) не все фазы периода должны быть включены в модель.

Опишем предлагаемый алгоритм решения задачи (152).

1. Задаются единичные ковариационные матрицы $\mathbf{A}, \mathbf{B}$.
2. Для фиксированных значений матриц $\mathbf{A}, \mathbf{B}$ оцениваются параметры $\mathbf{w}_{\text{МР}}$ модели $f$. При этом оптимизируется функция (154).
3. Оцениваются ковариационные матрицы $\mathbf{A}, \mathbf{B}$ согласно гипотезе порождения данных.
4. Последние два шага повторяются до сходимости: пока изменение элементов матриц $A, B$ не будут меньше заданных.
5. Выбираются те признаки $\mathbf{A}$ и объекты $\mathbf{B}$, которым соответствует наибольшие значения диагональных элементов матриц $\mathbf{A}, \mathbf{B}$ соответственно.
6. Мощности множеств $\mathbf{A}, \mathbf{B}$ выбираются такими, чтобы они доставляли максимум функции правдоподобия (35).

Алгоритм решения задачи (153) состоит из двух основных шагов. Модели, включенные в $\mathfrak{f}$ заданы разбиением множества индексов $\mathbf{B}_1 \sqcup \mathbf{B}_2$, имеют различные ковариационные матрицы $\mathbf{B}_1, \mathbf{B}_2$ и общий набор признаков $\mathbf{A}$.

1. Решается задача максимизации правдоподобия $\mathfrak{f}$ на множестве $\mathbf{A}$ как в предыдущем алгоритме; разбиение $\mathbf{B}$ фиксировано.
2. Решается задача максимизации расстояния $\rho(f_1, f_2)$. Для этого значения диагональных элементов $\mathbf{B}_1, \mathbf{B}_2$ упорядочиваются по убыванию. Выполняется обмен $b_1, b_2$ индексами из разбиения $\mathbf{B}_1, \mathbf{B}_2$, соответствующими наименьшим значениям диагональных элементов. Числа $b_1, b_2$ выбираются такими, что расстояние $\rho(f_1, f_2)$ между двумя моделями было максимально.

Рассмотренный метод позволяет решать задачу совместного выбора признаков и объектов как для одной регрессионной модели, так и для их набора. При этом особое внимание уделяется принятию статистических гипотез и, как следствие, корректности использования функций качества, с помощью которых отыскиваются оптимальные, а данном случае наиболее вероятные параметры моделей, а также их матрица их ковариаций.

# 5. Выбор моделей для данных в разнородных шкалах и экспертных оценок

В этом разделе описан способ построения интегральных индикаторов качества [290] сложных объектов с использованием экспертных оценок. Интегральные индикаторы вычисляются как линейная комбинация показателей объектов [356]. Используются экспертные оценки качества объектов и важности показателей, которые корректируются в процессе вычисления. Для сравнения с предлагаемым методом приведены известные методы построения интегрального индикатора «без учителя» и «с учителем». Для построения интегральных индикаторов необходимы как экспертные оценки качества объектов, так и объективные, измеряемые показатели — описания объектов. Роль экспертной оценки в данной работе велика. Эксперт устанавливает критерий, по которому оценивается объект, определяет множество сопоставимых по данному критерию объектов и выставляет оценки каждому объекту. Проблемам получения адекватных экспертных оценок посвящена работа [328].

Для принятия решений при администрировании объектов управления, например, государственных заповедников, регионов Российской Федерации или объектов финансирования, часто используются интегральные оценки качества или оценки эффективности управления объектами [315, 304, 297, 296]. «Качество — совокупность свойств объекта, обусловливающих его способность удовлетворять определенные потребности в соответствии с его назначением» [318]. Интегральный индикатор число, поставленное в соответствие объекту и рассматриваемое как оценка его качества.

При построении интегральных индикаторов, во-первых, выбирается критерий качества объектов [287]. «Критерий — признак, на основании которого производится оценка (например, оценка качества системы, ее функционирования), сравнение альтернатив (т.е. эффективности различных решений), классификация объектов и явлений» [318]. Во-вторых, формируется набор объектов, сравнимых в контексте выбранного критерия. В-третьих, формируется набор из тех показателей, которые эксперт считает необходимыми для описания этого критерия. После этого составляется таблица «объект-признак». Предполагается, что в этой таблице нет объектов-выбросов (способы их обнаружения описаны в работе [287]) и пропущенных значений. Показатели приведены к единой шкале и соответствуют принципу «чем больше, тем лучше»: большему значению показателя (при прочих равных) соответствует большее значение интегрального индикатора. Предполагается, что мультиколлинеарность показателей отсутствует или невысока [345, 80, 32]. Ранее было предложено несколько подходов к построению интегральных индикаторов [332, 324, 315, 306, 200]. Подход «без учителя» заключается в нахождении интегральных индикаторов с помощью описаний объектов и выбранного метода их построения. Приведем в качестве примера построение интегрального индикатора методом главных компонент, согласно которому интегральный индикатор является проекцией векторов-описаний объектов на первую главную компоненту матрицы «объект-признак» [334, 154, 148]. В статье [260] рассматриваются также такие методы построения интегральных индикаторов «без учителя», как Парето-расслоение и метрический метод. Подход «с учителем» использует кроме описаний объектов экспертные оценки их качества или оценки важности показателей и заключается в нахождении компромисса между ними

и вычисленными индикаторами. Ранее был предложен подход, в котором восстанавливается регрессия описаний объектов на экспертные оценки качества объектов [342]. Экспертные оценки могут быть выставлены в линейных или в ранговых шкалах. В некоторых случаях [328] эксперты не могут выставить оценки в линейных шкалах. Поэтому данная работа посвящена уточнению экспертных оценок, выставленных в ранговых шкалах. Предлагаемый метод заключается в следующем. Принята линейная модель зависимости интегрального индикатора от весов показателей. Экспертные оценки весов показателей задают выпуклый многогранный конус, а матрица «объект – признак» — линейное отображение этого конуса из пространства показателей в пространство интегральных индикаторов. Полученный в результате отображения конус может пересекаться с конусом, заданным экспертными оценками интегрального индикатора. В этом случае экспертные оценки показателей и объектов считаются непротиворечивыми и отыскивается наиболее устойчивый интегральный индикатор. В противном случае выполняется описанная ниже процедура рангового уточнения оценок. Данный метод рассматривает оценки, выставленные одним экспертом. Если оценки выставлены группой экспертов, их следует привести к согласованному виду [317], например, посредством вычисления медианы Кемени.

## 5.1. Регрессионная модель согласования экспертных оценок

Задано множество $\Upsilon = \{\upsilon_1, ..., \upsilon_m\}$ объектов и множество показателей $\Psi = \{\psi_1, ..., \psi_n\}$. Произвольный объект $\upsilon_i$ описывается с помощью вектора-строки $\mathbf{x}_i = \langle x_{i1}, x_{i2}, ..., x_{in}\rangle : \mathbf{x}_i \in \mathbb{R}^n$. Множество измерений представляется в виде матрицы исходных данных, обозначаемой $\mathbf{X} = \{x_{ij}\}_{i,j=1}^{m,n}$ в пространстве действительных чисел: $\mathbf{X} \in \mathbb{R}^{m\times n}$. Элемент $x_{ij}$ — значение $j$-го показателя $\psi_j$ для $i$-го объекта $\upsilon_i$.

*Интегральным индикатором объекта* $\upsilon_i \in \Upsilon$ с номером $i$ называется скаляр $y_i$, поставленный в соответствие набору $\mathbf{x}_i$ описаний объекта. При рассмотрении множества объектов $\Upsilon$ вектор $\mathbf{y} = \langle y_1, ..., y_m\rangle^\intercal : \mathbf{y} \in \mathbb{R}^m$ считается интегральным индикатором множества объектов, описанных матрицей $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^m : \mathbf{X} \in \mathbb{R}^{m\times n}$.

Объект $\upsilon_i$, имеющий максимальный по значению интегральный индикатор (наибольшую экспертную оценку, если она рассматривается в качестве интегрального индикатора) $y_i = \max\{y_1, ..., y_m\}$ считается *наилучшим*. Показатель $\psi_j$, имеющий максимальный по значению вес (наибольшую экспертную оценку, если она рассматривается в качестве веса показателя) $w_j = \max\{w_1, ..., w_n\}$ считается *наиважнейшим* при нахождении интегрального индикатора. Таким образом выполнено предложение

$$\begin{gathered} x_{\xi\zeta} = \max\{x_{i\zeta}\}_{i=1}^m \Rightarrow y_\xi = \max\{y_1, ..., y_m\}, \\ x_{\eta\vartheta} = \min\{x_{i\vartheta}\}_{i=1}^m \Rightarrow y_\eta = \min\{y_1, ..., y_m\}. \end{gathered} \tag{155}$$

Векторы $\boldsymbol{\chi}_j = \langle x_{1j}, ..., x_{mj}\rangle^\intercal : \boldsymbol{\chi}_j \in \mathbf{X}$ нормированы так, что выполняется равенство

$$x_{ij} = 1 - \frac{|x_{ij} - x_j^{opt}|}{\max([x_j^{opt} - \min(\boldsymbol{\chi}_j)], [\max(\boldsymbol{\chi}_j) - x_j^{opt}])}, \quad i = 1, ..., m,\ j = 1, ..., n, \tag{156}$$

где оптимальное значение $\mathrm{opt}(\boldsymbol{\chi}_j) : \min(\boldsymbol{\chi}_j) \leqslant \mathrm{opt}(\boldsymbol{\chi}_j) \leqslant \max(\boldsymbol{\chi}_j)$ задано.

### 5.1.1. Базовая модель построения интегральных индикаторов

Основная идея методов нахождения интегрального индикатора «без учителя» заключается в том, что наилучшим считается $i$-й объект с максимальными значениями показателей (обозначим ее "*tbtb*" — the bigger the better). Объект с наибольшим интегральным индикатором при выполнении условия (156) имеет значения показателей $\bar{\mathbf{x}}_i = \{1, 1, ..., 1\}$. Сильная сторона данной идеи в ее простоте и универсальности. Слабая сторона идеи заключается в том, что она предполагает определенную линейную зависимость между столбцами матрицы $\mathbf{X}$. Например, оценивая объекты, которым соответствует матрица $\mathbf{X} = \{x_{ij}\}_{i,j=1}^{2,m}$ при коэффициенте корреляции между ее столбцами $r_{1,2} = -1$, эксперт, который ориентируется на гипотезу "*tbtb*", скажет, что данные противоречивы.

Задана таблица описаний объектов — матрица $\mathbf{X} = \{x_{ij}\}_{i,j=1}^{m,n}$, $\mathbf{X} \in \mathbb{R}^{m \times n}$. Элемент матрицы $x_{ij}$ — значение $j$-го показателя $i$-го объекта. Вектор $\mathbf{x}_i = \langle x_{i1}, ..., x_{in} \rangle$ — описание $i$-го объекта. Для краткости объектом далее будет называется непосредственно сам вектор $\mathbf{x}_i$.

Интегральный индикатор объекта — линейная комбинация вида

$$y_i = \sum_{j=1}^{n} w_j g_j(x_{ij}), \tag{157}$$

где $g_j$ — функция приведения показателей в единую шкалу:

$$g_j : x_{ij} \mapsto (-1)^{s_j} \frac{x_{ij} - \min_i x_{ij}}{\max_i x_{ij} - \min_i x_{ij}} + s_j. \tag{158}$$

Модификатор $s_j$ назначается равным единице, если оптимальное значение $j$-го показателя минимально, и нулю, если оптимальное значение показателя максимально. Если знаменатель дроби (158) равен нулю для некоторых значений индекса $j$, то соответствующий признак исключается из дальнейшего рассмотрения.

При выполнении условия (158) интегральный индикатор будет иметь вид

$$\mathbf{y} = \mathbf{X}\mathbf{w},$$

где вектор интегральных индикаторов $\mathbf{y} = \langle y_1, \ldots, y_m \rangle^\intercal$ и вектор весов важности показателей $\mathbf{w} = \langle w_1, \ldots, w_n \rangle^\intercal$. Для построения интегрального индикатора требуется найти веса важности показателей.

### 5.1.2. Критерий наибольшей информативности

Рассмотрим алгоритм получения интегрального индикатора «без учителя». Метод главных компонент, используемый для вычисления интегральных индикаторов [263], заключается в том, что к множеству описаний объектов применяется преобразование вращения, которое соответствует критерию *наибольшей информативности* С. Р. Рао [231]. Согласно этому критерию, наибольшая информативность есть минимальное значение суммы квадратов расстояния от описаний объектов до их проекций на первую главную компоненту.

**Теорема 10** (Рао). *Наилучшим выбором линейных функций, для которых остаточная дисперсия, предсказания с помощью линейного предиктора, минимальна, является выбор первых $k$ главных компонент случайной величины $\mathbf{X}$.*

Для нахождения первой главной компоненты требуется найти такие линейные комбинации $\mathbf{Z}^\intercal = \mathbf{W}\mathbf{X}^\intercal$ векторов-столбцов матрицы $\mathbf{X}$, что векторы-столбцы $\mathbf{z}_1, ..., \mathbf{z}_n$ матрицы $\mathbf{Z}$ обладали бы наибольшей дисперсией: $\max \sum_{j=1}^{n} \mathsf{D}\mathbf{z}_j$ при ограничениях нормировки $\mathbf{W}\mathbf{W}^\intercal = \mathbf{I}$ — единичная матрица. Рао было показано, что строки матрицы $\mathbf{W}$ есть собственные векторы ковариационной матрицы $\mathbf{\Sigma} = \mathbf{X}^\intercal\mathbf{X}$. Значение интегрального индикатора $\mathbf{y}$ вычисляется как проекция векторов-строк матрицы $\mathbf{X}$ на первую главную компоненту, $\mathbf{y} = \mathbf{X}\mathbf{w}$, где $\mathbf{w}$ — вектор-столбец матрицы $\mathbf{W}^\intercal$, соответствующий наибольшему собственному значению матрицы $\mathbf{\Sigma}$.

### 5.1.3. Метрический метод построения модели

При нахождении интегрального индикатора с помощью данной процедуры вычисляется расстояние от вектора-столбца $\mathbf{x_i}$., описывающего каждый объект, либо до наихудшего объекта с показателями, принимающими значение $\mathbf{x}_i = \{0, 0, ..., 0\}$, $y_i = (\sum_{j=1}^{n} \bar{x}_{ij}^k)^{\frac{1}{k}}$; либо до наилучшего объекта с показателями, принимающими значение $\mathbf{x}_i = \{1, 1, ..., 1\}$, при соблюдении условия (158), $y_i = (\sum_{j=1}^{n}(1 - x_{ij})^k)^{\frac{1}{k}}$. При значении $k = 1, 2$ расстояния вычислены соответственно в манхэттенской и евклидовой метрике. При $k \geq 3$ полученное расстояние называется расстоянием Минковского. В частности, для нахождения интегральных индикаторов использовалась взвешенная сумма $\mathbf{y} = \mathbf{X}\mathbf{w}_0$, где веса $\mathbf{w}_0$ назначались экспертами.

### 5.1.4. Расслоение Парето

Интегральный индикатор, полученный методом расслоения Парето [12,13] инвариантен к любым преобразованиям исходных данных, сохраняющих порядок значений объектов внутри данного показателя. Это дает возможность опустить предварительную обработку данных [14].

Имеем исходные данные, представленные матрицей $\mathbf{X} = \{x_{ij}\}_{i,j=1}^{m,n}$, и $\mathbf{x}_i, \mathbf{x}_\xi \in \mathbb{R}^n$ — векторы данной матрицы, описывающие $i$-й и $\xi$-й объекты. Вектор $\mathbf{x}_\xi = \langle x_{\xi j} \rangle_{j=1}^{n}$ называется недоминируемым, если не найдется ни одного вектора $\mathbf{x}_i$, такого что

$$x_{ij} > x_{\xi j}, \quad i = 1, ..., m, j = 1, ..., n. \tag{159}$$

Для некоторого вектора $\mathbf{x}_\xi \in W$ пространство $W = \mathbb{R}^n$, в котором он находится, является объединением двух областей $W = W_1 \cup W_2$. Недоминируемые векторы $\mathbf{x}_\xi \in W_1$, остальные доминируемые векторы $\mathbf{x}_i \in W_2$. При совпадении векторов $\mathbf{x}_\xi = \mathbf{x}_i$ считается, что оба вектора находятся при соблюдении условия (159) в недоминируемой области $\mathbf{x}_\xi, \mathbf{x}_i \in W_1$. Произвольный вектор $\mathbf{x}_i$ сам себя не доминирует.

Введем отношение порядка на множестве объектов $\{\mathbf{x}_i\}$. Объект $\mathbf{x}_i$ доминирует объект $\mathbf{x}_k$, если все элементы вектора $\mathbf{x}_i$ не меньше соответствующих элементов вектора $\mathbf{x}_k$,

$$\mathbf{x}_i \succeq \mathbf{x}_k, \quad \text{если } x_{ij} \geqslant x_{kj} \quad \text{для } j = 1, \ldots, n.$$

Рассмотрим Парето-оптимальный фронт $P_1$ — множество недоминируемых объектов: для каждого объекта $\mathbf{x}_k \in P_1$ нет объекта $\mathbf{x}_i$ такого, что $\mathbf{x}_i \succeq \mathbf{x}_k$. Считая множество $P_0 = \emptyset$ Определим множество $P_s$ следующим образом. Парето-оптимальный фронт $P_s$, соответствующий

слою с номером 1, есть множество недоминируемых объектов из набора $\left\{\{\mathbf{x}_i\}_{i=1}^m \setminus \{\emptyset \cup P_1 \cup \ldots \cup P_{s-1}\}\right\}$. Множество Парето-оптимальных фронтов строится индуктивно до тех пор, пока объединение всех фронтов не совпадет со множеством объектов.

Определим интегральный индикатор $i$-го объекта $y_i = S - s(i)$, где $s$ — индекс Парето-оптимального фронта $P_s$, которому принадлежит объект $\mathbf{x}_i$, и $S$ — число всех полученных фронтов.

Для выполнения процедуры расслоения Парето требуется найти все недоминируемые векторы для каждого слоя. Определим исходные множества $S$ и $T$ как $S = \{\mathbf{x}_i\}_{i=1}^m$ и $T = \emptyset$. Для $\zeta$-го слоя множество

$$S_\zeta = \{\mathbf{x}_\xi : x_{\xi j} > x_{ij}, \mathbf{x}_\xi = \mathbf{x}_i, \xi \in \{1...n\}\}_{i,j=1}^{m,n}.$$

Все найденные векторы $\{\mathbf{x}_\xi\} \in S_\zeta$ находятся в одном слое. Добавляем множество $S_\zeta$ в множество $T$. Исключаем множество векторов $S_\zeta$ из дальнейшего рассмотрения и повторяем процедуру для множества векторов $S \setminus T$ до тех пор, пока в этом множестве не останется ни одного вектора. В результате расслоения получаем множество $T$, состоящее из $l$ слоев $S_\zeta$:

$$T = \bigcup_{\zeta=1}^{l} S_\zeta. \tag{160}$$

Для получения интегрального индикатора поставим в соответствие каждому вектору $\mathbf{x}_\xi$, $\xi = 1, ..., m$, индекс $\zeta$ множества $S_\zeta$, которому принадлежит вектор $\mathbf{x}_\xi$. Полученное множество $\Xi = \{\zeta_\xi\}_{\xi=1}^m$ приведем к виду, удовлетворяющему начальным условиям $\mathbf{y} = \{\max(\Xi) - \zeta_\xi\}_{\xi=1}^m$. Очевидно, что данном случае $\mathbf{y} \in \mathbb{Z}^m$.

Существенным недостатком расслоения Парето является то, что при большой размерности пространства показателей или при отрицательной корреляционной зависимости показателей значение $l$ становится равным единице. Этот недостаток может быть обойден, если в качестве набора входных показателей взять только те показатели, вклад которых при нахождении первой главной компоненты матрицы $\mathbf{X}$ наиболее значителен.

Для множества базовых показателей объектов $\Psi = \{\psi_j\}_{j=1}^n$ найдем такое подмножество $\Psi^* = \{\psi_{j_1}, ..., \psi_{j_l}\}$, $j_k \in \{1, ..., n\}$, $k = 1, ..., l$, для которого число $\nu^*$ различных элементов множества $\mathbf{y}_p = \{y_1, ..., y_m\}$, полученного посредством процедуры расслоения Парето, будет наиболее близко к $\nu$. Значение $\nu$ определяется экспертом на основании сведений о числе ожидаемых кластеров, на которые разбивается множество объектов $\Upsilon$, или на основании результатов процедуры кластеризации.

Подмножество базовых показателей $\Psi^*$ строится следующим образом. Пусть $\mathbf{y}_m$ — интегральный индикатор, полученный методом главных компонент и пусть $\boldsymbol{\chi}_j$ — вектор-столбец матрицы $\mathbf{X}$, соответствующий показателю $\psi_j$. Поставим в соответствие каждому показателю $\psi_j$ коэффициент корреляции $r_j = r(\mathbf{y}_m, \boldsymbol{\chi}_j)$ и получим множество $\mathbf{r} = \langle r_1, ..., r_j \rangle$. Из множества $\Psi$ последовательно выберем подмножества $\Psi^{(1)}, \Psi^{(2)}, \Psi^{(3)}, ...$, которые состоят из одного, двух, трех и т. д., элементов — показателей, имеющих наибольший коэффициент $r_j$ корреляции с первой главной компонентой $\mathbf{y}_m$.

Для каждого такого подмножества $\Psi^{(i)} =$ найдем интегральный индикатор $\mathbf{y}_p^{(i)}$ методом расслоения Парето. Искомым множеством $\Psi^*$ будем считать такое множество $\Psi^{(i)}$, которому

соответствует расслоение Парето с числом слоев, иначе, числом $\nu^*$ различных элементов множества $\mathbf{y}_p^*$, наиболее близким к заданному числу $\nu$.

Вклад показателей $\{\psi_{j_1}, ..., \psi_{j_l}\}$ подмножества базовых показателей при нахождении первой главной компоненты вычисляется как отношение

$$\rho^* = \frac{r_1^* + ... + r_l^*}{r_1 + ...r_n},$$

где $r_1^*, ..., r_l^*$ — множество коэффициентов корреляции между столбцами $\boldsymbol{\chi}_j$ матрицы $\mathbf{X}$, соответствующими построенному множеству $\Psi^*$, и первой главной компонентой $\mathbf{y}_m$; и $r_1, ..., r_n$ — множество коэффициентов корреляции между всеми столбцами $\boldsymbol{\chi}_j$ матрицы $\mathbf{X}$ и первой главной компонентой $\mathbf{y}_m$. Значение $\rho^*$ характеризует качество интегрального индикатора, полученного процедурой расслоения Парето с использованием подмножества $\Psi^*$ базовых показателей.

## 5.2. Криволинейные линейные методы согласования экспертных оценок

### 5.2.1. Экспертно-статистический метод

В работе [288] предложен экспертно-статистический метод нахождения интегральных индикаторов «с учителем», использующий экспертные оценки качества объектов. Он заключается в нахождении таких весов $\mathbf{w}$, при которых достигался бы минимум функционала невязки экспертных интегральных индикаторов $\mathbf{y}_0$ и вычисленных индикаторов: $\mathbf{w} = \arg\min_{\mathbf{w}} ||\mathbf{y}_0 - \mathbf{X}\mathbf{w}||_2^2$. Полученный индикатор $\mathbf{y}_{\text{э-с}} = \mathbf{X}\mathbf{w}$.

Рассмотрим задачу, в которой эксперты способны выставить адекватные интегральные индикаторы $\mathbf{y}_0$ и веса показателей $\mathbf{w}_0$. Тогда каждый объект можно оценить двумя путями: непосредственно через экспертную оценку и через взвешенную сумму значений показателей объекта, где веса определяются экспертными оценками показателей. В общем случае эти оценки различны.

Пусть задан вектор $\mathbf{y}_0 = \langle y_{01}, ..., y_{0m}\rangle^\intercal$, $\mathbf{y}_0 \in \mathbb{R}^m$ экспертных оценок интегральных индикаторов $m$ объектов и вектор $\mathbf{w}_0 = \langle w_{01}, ..., w_{0n}\rangle^\intercal$, $\mathbf{w}_0 \in \mathbb{R}^n$ экспертных оценок весов $n$ показателей. Задана матрица $\mathbf{X}$.

Согласно принятой модели, по исходным экспертным оценкам весов показателей $\mathbf{w}_0$ можно вычислить значения вектора интегрального индикатора $\mathbf{y}_1 = \mathbf{X}\mathbf{w}_0$, также, по исходным экспертным оценкам значения вектора интегрального индикатора $\mathbf{y}_0$ можно вычислить веса показателей $\mathbf{w}_1 = \mathbf{X}^+\mathbf{y}_0$, где $\mathbf{X}$ — линейный оператор, представляемый при помощи данной матрицы, $\mathbf{X}^+$ — оператор, псевдообратный [299] оператору $\mathbf{X}$. В общем случае вектор экспертной оценки $\mathbf{y}_0$ объектов и вектор взвешенной суммы значений показателей объектов $\mathbf{y}_1$ различны: $\mathbf{y}_0 \neq \mathbf{y}_1$, также $\mathbf{w}_0 \neq \mathbf{w}_1$.

*Согласованными значениями интегрального индикатора и весов показателей* называются такие значения $\hat{\mathbf{y}}$ и $\hat{\mathbf{w}}$, при которых выполняется условие

$$\begin{cases} \hat{\mathbf{y}} &= \mathbf{X}\hat{\mathbf{w}}, \\ \hat{\mathbf{w}} &= \mathbf{X}^+\hat{\mathbf{y}}. \end{cases} \tag{161}$$

Требуется найти оператор согласования $\Phi$, переводящий тройку $(\mathbf{y}_0, \mathbf{w}_0, \mathbf{X})$ в согласованную тройку $(\hat{\mathbf{y}}, \hat{\mathbf{w}}, \mathbf{X})$.

### 5.2.2. Линейное согласование экспертных оценок

Предлагаемый подход имеет целью согласовать экспертные оценки и заключается в поиске компромиссного решения. Согласно этому подходу, экспертам предоставляется возможность разрешить противоречие между интегральными индикаторами объектов, весами показателей и измеряемыми данными.

Введем следующую процедуру согласования. Пусть $\mathbf{X}$ — матрица линейного оператора, отображающего пространство весов показателей $W \ni \mathbf{w}_0$ в пространство интегральных индикаторов объектов $Q \ni \mathbf{y}_0$, $\mathbf{X} : W \longrightarrow Q$, и пусть для $\mathbf{X}$ существует псевдообратный оператор $\mathbf{X}^+$, отображающий пространство интегральных индикаторов в пространство весов показателей $\mathbf{X}^+ : Q \longrightarrow W$. То есть, $\mathbf{X}^+\mathbf{X} = \mathbf{I}_n$, $\mathbf{X}\mathbf{X}^+ = \mathbf{I}_m$, $\mathbf{X}^+\mathbf{X}\mathbf{X}^+ = \mathbf{X}^+$, $\mathbf{X}\mathbf{X}^+\mathbf{X} = \mathbf{X}$.

Сингулярное разложение [352] невырожденной матрицы $\mathbf{X}$ имеет вид $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$, где $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, ..., \lambda_R)$ — диагональная матрица, $R = \min(m, n)$ и $\mathbf{U}^\intercal\mathbf{U} = \mathbf{I}_m, \mathbf{V}\mathbf{V}^\intercal = \mathbf{I}_n$ — ортогональные матрицы. Матрица $\mathbf{X}^+ = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^\intercal$ является для матрицы $\mathbf{X}$ псевдообратной. Действительно, $\mathbf{X}^+\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^\intercal\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal = \mathbf{I}_n$, $\mathbf{X}\mathbf{X}^+ = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^\intercal = \mathbf{I}_m$.

Найдем отображение вектора $\mathbf{w}_0$ из пространства $W$ в пространство $Q$: $\mathbf{y}_1 = \mathbf{X}\mathbf{w}_0$ и отображение вектора $\mathbf{y}_0$ из пространства $Q$ в пространство $W$ : $\mathbf{w}_1 = \mathbf{X}^+\mathbf{w}_0$, см. рис. 56. Мы получили два отрезка — $[\mathbf{y}_1, \mathbf{y}_0] \subset Q$ и $[\mathbf{w}_1, \mathbf{w}_0] \subset W$. Их Евклидова длина $\|\mathbf{y}_0 - \mathbf{y}_1\|$ и $\|\mathbf{w}_0 - \mathbf{w}_1\|$ характеризует несогласованность экспертных оценок. Найдем согласованные оценки на этих отрезках. Для этого введем параметры $\alpha$ и $\beta$.

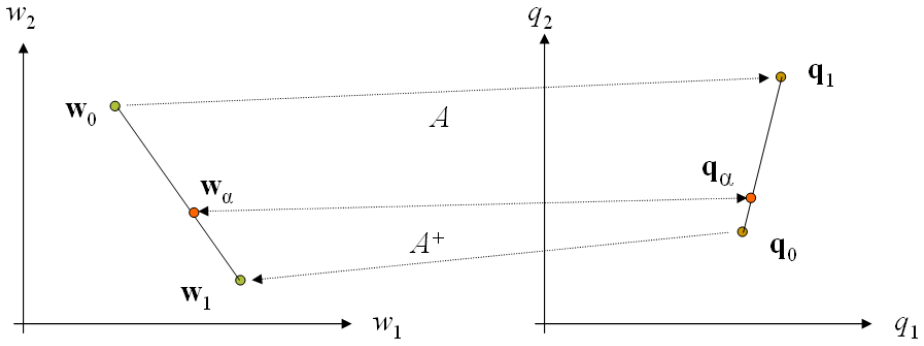



Рис. 56. Уточненные векторы экспертных оценок весов и индикаторов при $\alpha$-согласовании.

**Теорема 11.** *Существуют такие $\alpha, \beta$, при которых значения векторов $\mathbf{w}_\alpha, \mathbf{y}_\beta$*

$$\begin{array}{rcl} \{\mathbf{w}_\alpha : \mathbf{w}_\alpha & = & (1-\alpha)\mathbf{w}_0 + \alpha\mathbf{w}_1\} \in [\mathbf{w}_0, \mathbf{w}_1], \\ \{\mathbf{y}_\beta : \mathbf{y}_\beta & = & \beta\mathbf{y}_0 + (1-\beta)\mathbf{y}_1\} \in [\mathbf{y}_0, \mathbf{y}_1], \end{array} \tag{162}$$

*где $\alpha, \beta \in [0, 1]$, удовлетворяют требованиям согласования, то есть, $\mathbf{X}\mathbf{w}_\alpha = \mathbf{y}_\beta$, причем $\alpha = 1 - \beta$.*

*Доказательство.* Так как $\mathbf{y}_1 = \mathbf{X}\mathbf{w}_0$, $\mathbf{w}_1 = \mathbf{X}^+\mathbf{y}_0$, и линейный оператор $\mathbf{X} : [\mathbf{w}_0, \mathbf{w}_1] \longrightarrow [\mathbf{y}_0, \mathbf{y}_1]$, то равенство $(1-\alpha)\mathbf{X}\mathbf{w}_0 + \alpha\mathbf{X}\mathbf{w}_1 = (1-\beta)\mathbf{y}_0 + \beta\mathbf{y}_1$ справедливо при $\alpha = 1 - \beta$. $\square$

Перепишем выражение (162) с одним параметром:

$$\begin{array}{rcl} \mathbf{y}_\alpha & = & \alpha\mathbf{y}_0 + (1-\alpha)\mathbf{X}\mathbf{w}_0, \\ \mathbf{w}_\alpha & = & (1-\alpha)\mathbf{w}_0 + \alpha\mathbf{X}^+\mathbf{y}_0. \end{array} \tag{163}$$

Таким образом, согласованные экспертные оценки находятся с помощью выражения

$$\begin{array}{rcl} \mathbf{w}_\alpha & = & (1-\alpha)\mathbf{w}_0 + \alpha\mathbf{X}^+\mathbf{y}_0, \\ \mathbf{y}_\alpha & = & \alpha\mathbf{y}_0 + (1-\alpha)\mathbf{X}\mathbf{w}_0, \end{array} \tag{164}$$

где $\alpha : \alpha \in [0,1]$ — параметр доверия экспертным оценкам интегральных индикаторов объектов либо экспертным оценкам весов показателей. При значении $\alpha = 0$ мы игнорируем экспертные оценки объектов, учитывая оценки весов; при значении $\alpha = 1$ мы игнорируем экспертные оценки весов, учитывая оценки объектов.

Очевидно, что процедура $\alpha$-согласования дает согласованный результат.

**Теорема 12.** *Тройка* $(\mathbf{y}_\alpha, \mathbf{w}_\alpha, \mathbf{X})$*, полученная процедурой* $\alpha$*-согласования (164) удовлетворяет требованиям согласования (161).*

*Доказательство.* Подставив в равенство $\mathbf{X}\mathbf{w}_\alpha = \mathbf{y}_\alpha$ выражения для $\mathbf{y}_\alpha$ и $\mathbf{w}_\alpha$, получаем

$$\alpha\mathbf{y}_0 + (1-\alpha)\mathbf{X}\mathbf{w}_0 = \mathbf{X}\left((1-\alpha)\mathbf{w}_0 + \alpha\mathbf{X}^+\mathbf{y}_0\right), \text{или}$$
$$\alpha\mathbf{y}_0 + (1-\alpha)\mathbf{X}\mathbf{w}_0 = (1-\alpha)\mathbf{X}\mathbf{w}_0 + \alpha\mathbf{X}\mathbf{X}^+\mathbf{y}_0.$$

Так как $\mathbf{X}\mathbf{X}^+\mathbf{y}_0 = \mathbf{y}_0$, то $\alpha\mathbf{y}_0 + (1-\alpha)\mathbf{X}\mathbf{w}_0 = (1-\alpha)\mathbf{X}\mathbf{w}_0 + \alpha\mathbf{y}_0$. □

Таким образом, посредством параметра $\alpha$ становится возможно уточнять экспертные оценки $\mathbf{w}_0, \mathbf{y}_0$, получая новые оценки $\mathbf{w}_\alpha, \mathbf{y}_\alpha$.

Оценим невязку при выбранном параметре $\alpha$. Евклидово расстояние между исходными векторами $\mathbf{y}_0, \mathbf{w}_0$ и полученными векторами $\mathbf{y}_\alpha, \mathbf{w}_\alpha$ в пространстве интегральных индикаторов и в пространстве весов соответственно равны $\varepsilon^2 = \|\mathbf{y}_\alpha - \mathbf{y}_0\|^2$и $\delta^2 = \|\mathbf{w}_\alpha - \mathbf{w}_0\|^2$. В качестве критерия выбора параметра $\alpha$ возьмем условие минимального расстояния между начальными и согласованными экспертными оценками в обоих пространствах $Q$ и $W$. Учитывая, что размерности этих пространств соответственно равны $m$ и $n$, нормируем квадраты расстояний и находим такие согласованные значения векторов $\mathbf{y}_\alpha$ и $\mathbf{w}_\alpha$, что они удовлетворяют условию

$$\frac{\varepsilon^2}{m-1} = \frac{\delta^2}{n-1}. \tag{165}$$

На практике эксперты сами могут выбирать значение параметра $\alpha$ в зависимости от предпочтений важности оценок объектов или оценок показателей. Полученные результаты удобно предложить экспертам на обсуждение в следующем виде:

$$\left[\begin{array}{c|c|c} \text{Начальные} & & \mathbf{w}_0^\mathsf{T} \\ \hline & \text{Конечные} & \mathbf{w}_\alpha^\mathsf{T} \\ \hline \mathbf{y}_0 & \mathbf{y}_\alpha & \mathbf{X} \end{array}\right]. \tag{166}$$

При изменении параметра доверия экспертов $\alpha$ к экспертным оценкам объектов и показателей или при изменении самих экспертных оценок вышеописанную процедуру можно повторить и передать на обсуждение экспертов вновь полученные результаты.

### 5.2.3. Квадратичное согласование экспертных оценок

Определим согласованное решение как решение удовлетворяющее условию (161), при котором расстояние от согласованных векторов $\mathbf{y}_\gamma$ и $\mathbf{w}_\gamma$ таких, что $\mathbf{y}_\gamma = \mathbf{X}\mathbf{w}_\gamma$ до соответственно векторов экспертных оценок $\mathbf{y}_0$ и $\mathbf{w}_0$ будет минимальным. Пусть

$$\begin{array}{rcl} \varepsilon^2 & = & \|\mathbf{X}\mathbf{w} - \mathbf{y}_0\|^2, \\ \delta^2 & = & \|\mathbf{w} - \mathbf{w}_0\|^2. \end{array} \tag{167}$$

Решение задачи нахождения минимального расстояния от согласованных векторов до векторов экспертных оценок имеет вид

$$\mathbf{w}_\gamma = \arg\min_{\mathbf{w}\in W}(\varepsilon^2 + \gamma^2\delta^2), \tag{168}$$

где весовой множитель $\gamma^2 \in (0, \infty)$ — определяет степень компромисса между оценкой объектов и показателей. При малых значениях $\gamma^2$ в большей степени учитывается экспертная оценка объектов, а при больших значениях $\gamma^2$ в большей степени учитывается экспертная оценка показателей.

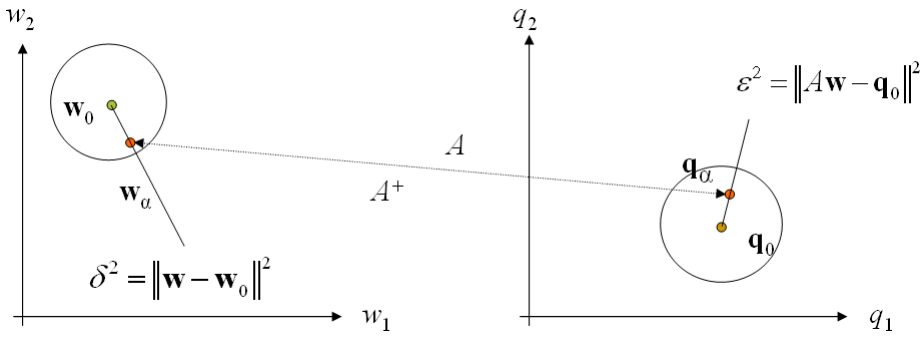



Рис. 57. Уточненные векторы экспертных оценок весов и индикаторов при $\gamma$-согласовании.

Выпуклый функционал $(\varepsilon^2 + \gamma^2\delta^2)$ достигает единственного глобального минимума на множестве $\mathbf{w}_\gamma \in W$ в точке

$$\mathbf{w}_\gamma = (\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})^{-1}(\mathbf{X}^\intercal\mathbf{y}_0 + \gamma^2\mathbf{w}_0). \tag{169}$$

Так же, как и предыдущем методе, параметр $\gamma^2$ для получения согласованных векторов $\mathbf{y}_\gamma = \mathbf{X}\mathbf{w}_\gamma$ и $\mathbf{w}_\gamma$ выбирается исходя из условия $\frac{\varepsilon^2}{m-1} = \frac{\delta^2}{n-1}$ или назначается экспертами.

**Теорема 13.** *Функционал $(\varepsilon^2 + \gamma^2\delta^2)$ достигает единственного глобального минимума на множестве $\mathbf{w}_\gamma \in W$ в точке*

$$\mathbf{w}_\gamma = (\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})^{-1}(\mathbf{X}^\intercal\mathbf{y}_0 + \gamma^2\mathbf{w}_0). \tag{170}$$

*Доказательство.* Функционал $(\varepsilon^2 + \gamma^2\delta^2)$ есть строго выпуклая функция, поэтому точка минимума выражения (168) существует и единственна. Найдем эту точку. Подставляя правые части выражения (167) в выражение (168) получаем

$$\mathbf{w}_\gamma = \arg\min_{\mathbf{w}\in W}\left(\|\mathbf{X}\mathbf{w} - \mathbf{y}_0\|^2 + \gamma^2\|\mathbf{w} - \mathbf{w}_0\|^2\right).$$

Используем обозначение нормы вектора $\|\mathbf{x}\|^2 = \sum_i x_i^2$ через скалярное произведение $(\mathbf{x}, \mathbf{x})$. Представим функционал $(\varepsilon^2 + \gamma^2\delta^2)$ в виде

$$\begin{aligned}
&\|\mathbf{X}\mathbf{w} - \mathbf{y}_0\|^2 + \gamma^2\|\mathbf{w} - \mathbf{w}_0\|^2 = \\
&(\mathbf{X}\mathbf{w} - \mathbf{y}_0, \mathbf{X}\mathbf{w} - \mathbf{y}_0) + \gamma^2(\mathbf{w} - \mathbf{w}_0, \mathbf{w} - \mathbf{w}_0) = \\
&(\mathbf{X}^\intercal\mathbf{X}\mathbf{w} - 2\mathbf{X}^\intercal\mathbf{y}_0 + \gamma^2\mathbf{w} - 2\gamma^2\mathbf{w}_0, \mathbf{w}) + (\mathbf{y}_0, \mathbf{y}_0) + \gamma^2(\mathbf{w}_0, \mathbf{w}_0).
\end{aligned}$$

Полученное выражение имеет минимум по $\mathbf{w}$ при значении $\nabla_{\mathbf{w}} = 0$, где

$$\nabla_{\mathbf{w}} = 2(\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})\mathbf{w} - 2(\mathbf{X}^\intercal\mathbf{y}_0 + \gamma^2\mathbf{w}_0),$$

здесь $\mathbf{I}$ — единичная матрица, размерность которой равна размерности матрицы $\mathbf{X}^\intercal\mathbf{X}$. Из предыдущего выражения находим вектор $\mathbf{w}_\gamma \in \mathbb{R}^n$, который удовлетворяет условию (168):

$$\mathbf{w}_\gamma = (\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})^{-1}(\mathbf{X}^\intercal\mathbf{y}_0 + \gamma^2\mathbf{w}_0).$$

□

**Теорема 14.** *Тройка* $(\mathbf{y}_\gamma, \mathbf{w}_\gamma, \mathbf{X})$*, полученная процедурой* $\gamma^2$*-согласования*

$$\mathbf{w}_\gamma \;\; = \;\; (\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})^{-1}(\mathbf{X}^\intercal\mathbf{y}_0 + \gamma^2\mathbf{w}_0),$$

*удовлетворяет требованиям согласования (161).*

*Доказательство.* Так как $\mathbf{y}_\gamma = \mathbf{X}\mathbf{w}_\gamma$, то $\mathbf{X}(\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})^{-1}(\mathbf{X}^\intercal\mathbf{y}_0 + \gamma^2\mathbf{w}_0) = \mathbf{X}\mathbf{w}_\gamma$. □

Выбор параметра доверия к экспертным оценкам интегрального индикатора или к экспертным оценкам весов показателей проиллюстрируем следующим образом. На рис. 58 показано изменение векторов $\mathbf{w}_\alpha$ и $\mathbf{y}_\alpha$ для различных значений параметра $\alpha$. По оси абсцисс отложены номера компонент векторов, а по оси ординат отложены значения векторов. Значения параметра $\alpha$ на графиках сверху вниз соответственно равны $\{0, 0.36, 1\}$. Каждая горизонтальная пара графиков показывает состояние согласованной пары векторов $\mathbf{y}_\alpha$ и $\mathbf{w}_\alpha$ при данном значении $\alpha$. По оси абсцисс отложены номера компонент векторов, по оси ординат отложены значения данных компонент векторов.

При минимальном значении параметра $\alpha$, близки исходная оценка индикатора $\mathbf{y_0}$ и согласованная оценка $\mathbf{y}_\alpha$, см. верхний правый график. При максимальном значении параметра $\alpha$, близки исходная оценка весов показателей $\mathbf{w_0}$ и согласованная оценка $\mathbf{w}_\alpha$, см. нижний левый график. При значении $\alpha = 0.36$ расстояния обоих согласованных векторов до соответствующих им исходных векторов становятся одинаковы: $\frac{\varepsilon^2}{m} = \frac{\delta^2}{n}$. Изменение расстояний $\varepsilon, \delta$ при выборе параметров $\alpha, \gamma^2$ показаны на рис. 59. Здесь по оси абсцисс отложены значения $\alpha, \gamma^2$, а по оси ординат значения $\varepsilon, \delta$. При увеличении $\alpha$ расстояние $\varepsilon$ между векторами $\mathbf{y_0}$ и $\mathbf{y}_\alpha$ увеличивается, а расстояние $\delta$ между векторами $\mathbf{w_0}$ и $\mathbf{w}_\alpha$ уменьшается.

Для оценки работ процедур согласования воспользуемся суммарным расстоянием от векторов экспертных оценок до согласованных векторов $\frac{\varepsilon^2}{m} + \frac{\delta^2}{n}$. Для процедуры $\alpha$-согласования оно равно 0.67, для процедуры $\gamma^2$-согласования — 0.62. Расстояние, полученное с помощью

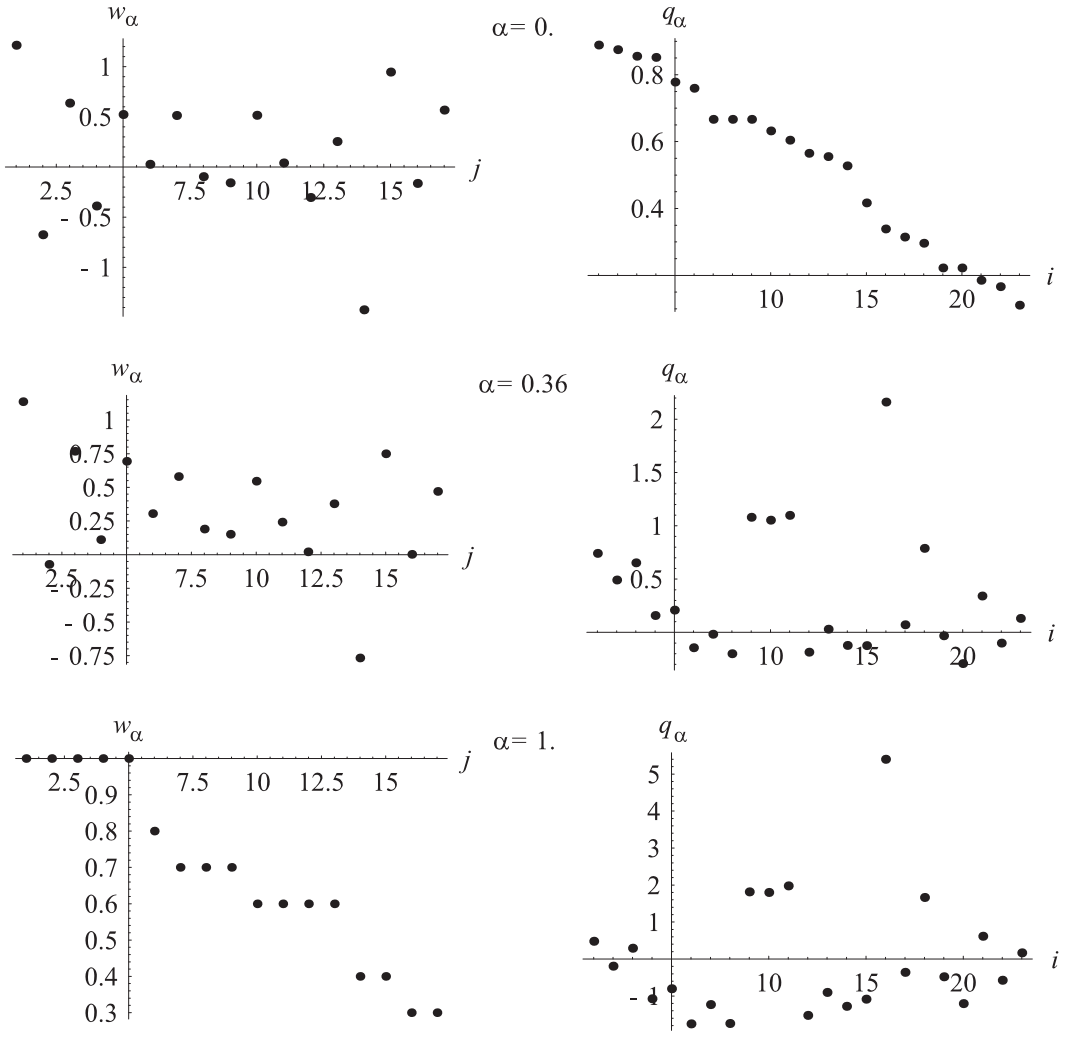



Рис. 58. Изменение весов показателей и интегрального индикатора при различных значениях параметра $\alpha$.

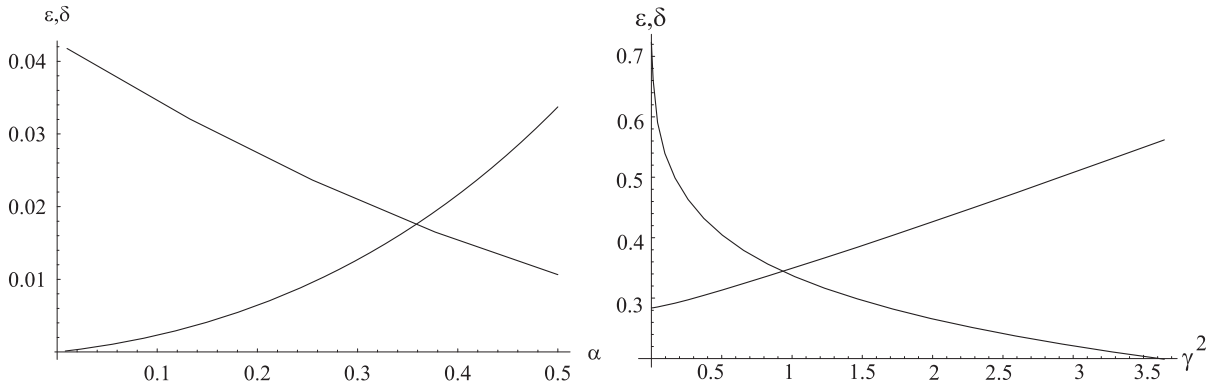



Рис. 59. Зависимость расстояний между векторами от параметров $\alpha$ и $\gamma^2$.

процедуры $\gamma^2$-согласования меньше, чем расстояние, полученное с помощью процедуры $\alpha$-согласования, так как во втором случае согласованные векторы $\mathbf{y}_\alpha, \mathbf{w}_\alpha$ принадлежат соответственно отрезкам $[\mathbf{y}_0, \mathbf{y}_1]$ и $[\mathbf{w}_0, \mathbf{w}_1]$, а в первом случае согласованные векторы $\mathbf{y}_\gamma, \mathbf{w}_\gamma$ лежат в окрестности соответственно векторов $\mathbf{y}_0$ и $\mathbf{w}_0$.

Существует множество способов получения интегральных индикаторов с использованием измеряемых данных. Но после выбора алгоритма и получения результатов встает вопрос: как показать, что полученные индексы верны? Для ответа на этот вопрос аналитики приглашают экспертов. Эксперты высказывают свое мнение и тогда встает другой вопрос: как обосновать адекватность экспертных оценок? Предлагаемый метод позволяет оценить непротиворечивость экспертных оценок и получить обоснованные интегральные индикаторы. Обе процедуры — $\alpha$-согласования и $\gamma^2$-согласования дают в численных экспериментах близкие результаты. Поэтому первая процедура может быть рекомендована в том случае, когда параметр согласования назначают сами эксперты. В том случае, когда требуется найти минимальное суммарное расстояние, предпочтательна процедура $\gamma^2$-согласования.

### 5.2.4. Монотонное согласование экспертных оценок

Предлагается процедура, где с оценками $\mathbf{y}_0, \mathbf{w}_0$ разрешены любые монотонные преобразования, т. е. введено отношение порядка на множестве элементов векторов $\mathbf{w}_0 = \{w_j : w_1 \leqslant ... \leqslant w_n\}_{j=1}^n$ и $\mathbf{y}_0 = \{y_i : y_1 \leqslant ... \leqslant y_m\}_{i=1}^m$, которое задает соответственно конусы $\mathcal{W} \in \mathbb{R}^n$ и $\mathcal{Q} \in \mathbb{R}^m$. При нахождении согласованных оценок вводятся монотонные корректирующие функции $T_\mathcal{Q} : \mathcal{Q} \longrightarrow \mathcal{Q}$ и $T_\mathcal{W} : \mathcal{W} \longrightarrow \mathcal{W}$, приближающие начальные экспертные оценки при сохранении отношения порядка.

Дана тройка $(\mathbf{y}_0, \mathbf{w}_0, \mathbf{X})$. Найдем такие векторы $\mathbf{y}_\tau = T_\mathcal{Q}(\mathbf{y}_0)$ и $\mathbf{y}_\tau = T_\mathcal{W}(\mathbf{w}_0)$, что выполняется условие минимума невязки

$$\mathbf{X}T_\mathcal{W}(\mathbf{w}_0) - T_\mathcal{Q}(\mathbf{y}_0) = \Delta. \tag{171}$$

Для $k = 0, ..., K$ укажем такие векторы

$$\begin{aligned} \mathbf{w}_{k+1} &= T_{\mathcal{W},k}(\mathbf{w}_k, \mathbf{X}^+\mathbf{y}_k), \\ \mathbf{y}_{k+1} &= T_{\mathcal{Q},k}(\mathbf{y}_k, \mathbf{X}\mathbf{w}_k), \end{aligned} \tag{172}$$

которые доставляют минимум функционалу $\|\Delta_k\|^2 = \|\mathbf{X}\mathbf{w}_k - \mathbf{y}_k\|^2$. Векторы $\mathbf{y}_\tau, \mathbf{y}_\tau$, находим в результате композиции $T_\mathcal{Q} = T_{\mathcal{Q},1} \circ ... \circ T_{\mathcal{Q},K}$ и $T_\mathcal{W} = T_{\mathcal{W},1} \circ ... \circ T_{\mathcal{W},K}$.

Нахождение корректирующей функции $T$. Рассмотрим два множества $\mathbf{x} = \{x_1, ..., x_m\}$ и $\mathbf{t} = \{t_1, ..., t_m : t_1 \leqslant ... \leqslant t_m\}$. Множество пар $\phi = \{(t_1, x_1), ..., (t_m, x_m)\}$ задают функцию $\phi$, и $x_i = \phi(t_i)$. Функция $\phi$, вообще говоря, немонотонна. Найдем такую монотонную функцию $f : t \longrightarrow x$, $f \in P_m$ которая аппроксимирует $\phi$,

$$f(t) = \arg\min_{f \in P_m} \sum_{i=1}^m \Big(f(t_i) - \phi(t_i)\Big)^2,$$

где $P_m$ — множество всех возрастающих полиномов степени $p \leqslant m$. Также найдем такую функцию $\varphi : t \longrightarrow x, \varphi \in \Theta$, которая интерполирует множество пар $\phi$:

$$\varphi(t) = \arg\min_{\varphi \in \Theta} \|\varphi(t) - \phi(t)\|,$$

где $\Theta$ — множество полиномиальных сплайнов с $m$ узлами степени $r$ дефекта 1.

Для приближения функции $\varphi$ функцией $f$ воспользуемся методом касательных Ньютона-Канторовича. Рассмотрим $f(t), \varphi(t)$ на отрезке $S = [a, b] \ni t$. Требуется найти гомеоморфизм $\vartheta : S \longrightarrow S, \vartheta(t) = t + \tau(t)$, такой, что

$$\vartheta = \arg\min_{\tau \in S} \|f(t) - \varphi(\vartheta(t))\|^2,$$

при значении $\tau = O(t)$. Для нахождения $\tau$ представим $\varphi(\vartheta(t))$ в виде $\varphi(\vartheta(t)) = \varphi(t) + \tau(t)\varphi'(t) + O(\tau^2(t))$.

**Теорема 15.** *Решением задачи оптимизации*

$$\tau_\epsilon(t) = \arg\min_{\tau \in S}\Big(\|f(t) - \varphi(t)\|^2 + \epsilon^2\|\tau(t)\|^2\Big)$$

*является выражение*

$$\tau_\epsilon(t) = \frac{\Big(f(t) - \varphi(t)\Big)\varphi'(t)}{\Big(\varphi(t)\Big)^2 + \epsilon^2}.$$

Зададим искомую функцию $T : \mathbf{x} \longrightarrow \mathbf{y}$ следующим образом. Подставляя в найденную функцию $\varphi(\vartheta(t))$ значения $t_i$ из $\phi$ получаем скорректированные оценки $y_i = \varphi(\vartheta(t_i))$, $i = 1, ..., m$. Параметр $\epsilon^2$, определяющий, насколько велика разность между значениями, которые принимает функция $\varphi$ в точках $t$ и $\vartheta(t)$, подбирается таким образом, чтобы функция $T(\vartheta(t))$ была монотонной.

### 5.2.5. Криволинейная регрессия для согласования экспертных оценок

Ниже предложен метод построения рангового интегрального индикатора на основе ранговой матрицы описаний, заданной экспертами. Предложен трехшаговый итеративный алгоритм оценки параметров и весов признаков. Рассмотрена задача выбора наиболее информативных признаков. Работа проиллюстрирована задачей определения статуса редких видов, включенных в Красную книгу РФ.

Задана множество $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, \quad i \in \mathcal{I} = \{1, \ldots, m\}$ пар. Каждая пара состоит из описания объекта $\mathbf{x}_i$ (таксона) и соответствующей ему метки класса $y_i$ (категория статуса таксона).

Описание объекта $\mathbf{x} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_n]^\intercal, \quad j \in \mathcal{J} = \{1, ..., n\}$ — это набор экспертных оценок признаков. Оценки объектов по признакам выставлены в ранговых шкалах. Каждый признак $\boldsymbol{\chi}_j$ имеет собственную ранговую шкалу $\mathbb{L}_j$, состоящую из $k_j$ упорядоченных элементов $\mathbb{L}_j = \{1 \prec 2 \prec \cdots \prec k_j\}$. Значение класса $y$ также принадлежит упорядоченному множеству $\mathbb{L}_0 = \{1 \prec 2 \prec \cdots \prec k_0\}$.

Рассмотрим постановку задачи многоклассовой классификации в ранговых шкалах, включающую криволинейную модель $f(\mathbf{w}, \mathbf{x}_i)$ и соответствующую ей вектор-функцию $\mathbf{f}(\mathbf{w}, \mathbf{X}) = [f(\mathbf{w}, \mathbf{x}_1), \ldots, f(\mathbf{w}, \mathbf{x}_m)]$ с матрицей описаний $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_i^\intercal, \ldots, \mathbf{x}_m^\intercal]^\intercal$ и зависисмой переменной $\mathbf{y} = [y_1, \ldots, y_i, \ldots, y_m]^\intercal$, где $\mathbf{w} = [w_1, \ldots, w_s]^\intercal$ — параметры модели. Эта модель должна доставлять минимум заданной функции ошибки $S(\mathbf{f}(\mathbf{w}, \mathbf{X}), \mathbf{y})$.

Криволинейная модель $f(\mathbf{w}, \mathbf{x}_i)$ имеет вид

$$f(\mathbf{w}, \mathbf{x}_i) = \xi\big(\mathbf{b}_0, h(\mathbf{w}, \mathbf{x}_i)\big), \tag{173}$$

$$h(\mathbf{w}, \mathbf{x}_i) = \sum_{j \in \mathcal{J}} u_j g(\mathbf{b}_j, x_{ij}). \tag{174}$$

где вектор параметров $\mathbf{w} = [\mathbf{b}_0; \mathbf{b}_1; \ldots; \mathbf{b}_n; \mathbf{u}] = [\mathbf{b}_0^\intercal, \mathbf{b}_1^\intercal, \ldots, \mathbf{b}_n^\intercal, \mathbf{u}^\intercal]^\intercal$ состоит из векторов $\mathbf{b}_j$ — параметров монотонной коррекции $j$-го признака $\chi_j$ и весовых коэффициентов признаков $\mathbf{u} = [u_1, \ldots, u_j, \ldots, u_n]^\intercal$. Функция $g$ монотонной коррекции задана следующим образом:

$$g(\mathbf{b}_j, \boldsymbol{\chi}) : \boldsymbol{\chi} \mapsto \mathbf{b}_j = \begin{cases} 1 \mapsto b_{j1}, \\ 2 \mapsto b_{j2}, \\ \ldots \\ k_j \mapsto b_{jk_j}. \end{cases}$$

При этом соблюдается условие монотонности параметров,

$$\mathrm{Ord}(\mathbf{b}_j) : 0 < b_{j1} < b_{j2} < \cdots < b_{jk_j} < 1 \quad \text{для} \quad j = 1, ..., n \quad \text{и} \tag{175}$$

$$\mathrm{Ord}(\mathbf{b}_0) : b_{01} < b_{02} < \cdots < b_{0k_0}.$$

Функция $\xi\big(\mathbf{b}_0, h(\mathbf{w}, \mathbf{x}_i)\big)$ определяет для числа $h(\mathbf{w}, \mathbf{x}_i)$ ближайшую по модулю компоненту вектора $\mathbf{b}_0$:

$$\xi\big(\mathbf{b}_0, h(\mathbf{w}, \mathbf{x}_i)\big) = \arg\min_{j \in \mathcal{J}} |b_{0j} - h(\mathbf{w}, \mathbf{x}_i)|.$$

Введя обозначение для матрицы скорректированных экспертных оценок

$$\mathbf{G} = [g_{ij}] = [g(\mathbf{b}_j, x_{ij})], \quad i \in \mathcal{I}, j \in \mathcal{J},$$

перепишем (173) и (174) в виде модели интегрального индикатора

$$f(\mathbf{w}, \mathbf{x}_i) = \xi(\mathbf{b}_0, [\mathbf{G}\mathbf{u}]_i). \tag{176}$$

Назначим функцией ошибки модели сумму квадратов регрессионных остатков,

$$S(\mathbf{w}) = \|\mathbf{f}(\hat{\mathbf{w}}, X) - \mathbf{y}\|_2^2 + \lambda \|\hat{\mathbf{u}}\|_2^2,$$

включающую регуляризующее слагаемое с фиксированным коэффициентом $\lambda$, где $\hat{\mathbf{w}}$ и $\hat{\mathbf{u}}$ — параметры, которые необходимо оценить.

**Оценивание параметров модели.** Оценивание параметров $\mathbf{w}$ модели $\mathbf{f}$ выполняется итеративно. Перед началом итераций значения векторов $\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_n$ назначены таким образом, что функция $g$ является тождественной, $g = \mathrm{id}$. Оценивание параметров выполняется в три этапа. Сначала при фиксированных значениях векторов $\hat{\mathbf{b}}_0, \ldots, \hat{\mathbf{b}}_n$ оцениваются весовые коэффициенты

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u} \in \mathbb{R}^n} S\big([\hat{\mathbf{b}}_0; \ldots; \hat{\mathbf{b}}_n; \mathbf{u}]^\intercal\big).$$

Затем при фиксированных значениях коэффициентов $\hat{\mathbf{u}}$ оцениваются параметры монотонной коррекции

$$[\mathbf{b}_1; ...; \mathbf{b}_n] = \underset{\mathrm{Ord}(\mathbf{b}_1),...,\mathrm{Ord}(\mathbf{b}_n)}{\arg\min} S\big([\hat{\mathbf{b}}_0; \ldots; \hat{\mathbf{b}}_n; \mathbf{u}]^\intercal\big)$$

с учетом требования монотонности (175) значений этих параметров. На последнем этапе оценивается вектор $\mathbf{b}_0$

$$\mathbf{b}_0 = \underset{\mathrm{Ord}(\mathbf{b}_0)}{\arg\min} S\big([\hat{\mathbf{b}}_0; \ldots; \hat{\mathbf{b}}_n; \mathbf{u}]^\intercal\big).$$

Итерации выполняются до стабилизации функции ошибки $S$.

Рассмотрим эти три этапа более подробно. За начальное приближение примем столбцы матрицы $\mathbf{G}$

$$\hat{\mathbf{G}} = [\mathbf{g}(\hat{\mathbf{b}}_1, \boldsymbol{\chi}_1), \ldots, \mathbf{g}(\hat{\mathbf{b}}_n, \boldsymbol{\chi}_n)] = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n],$$

поскольку, как было сказано выше, $g = \mathrm{id}$, и вектор $\hat{\mathbf{y}} = \mathbf{y}$. Таким образом, векторы $\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_n$ в начальном приближении в качестве элементов содержат элементы множеств $\mathbb{L}_0, \mathbb{L}_1, \ldots, \mathbb{L}_n$.

**Шаг 1.** Найдем $\hat{\mathbf{u}}$ при фиксированных $\hat{\mathbf{b}}_0, \ldots, \hat{\mathbf{b}}_n$:

$$\hat{\mathbf{u}} = \arg\min_{\mathbf{u}} \|\hat{\mathbf{y}} - \hat{\mathbf{G}}\mathbf{u}\| + \lambda\|\mathbf{u}\|.$$

Решение на шаге 1 имеет вид:

$$\hat{\mathbf{u}} = (\hat{\mathbf{G}}^T\hat{\mathbf{G}} + \lambda I)^{-1}\hat{\mathbf{G}}^T\hat{\mathbf{y}}.$$

**Шаг 2.** При фиксированных $\hat{\mathbf{b}}_0, \hat{\mathbf{u}}$ оценим скорректированную матрицу описаний

$$\mathbf{G} = [\mathbf{g}(\mathbf{b}_1, \boldsymbol{\chi}_1), \ldots, \mathbf{g}(\mathbf{b}_n, \boldsymbol{\chi}_n)] = [\mathbf{g}_1, \ldots, \mathbf{g}_n].$$

Для каждого $\mathbf{g}_j \in \mathbb{R}^m$ будем вычислять вектор $\hat{\mathbf{g}}_j$, являющийся монотонной коррекцией исходного вектора $\mathbf{g}_j$:

$$\begin{cases} [\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_n] = \arg\min \|\boldsymbol{\xi}(\mathbf{b}_0, \mathbf{G}\hat{\mathbf{u}}) - \hat{\mathbf{y}}\|_2^2, \\ \text{из} \quad g_{ij_1} \leqslant g_{ij_2} \quad \text{следует} \quad \hat{g}_{ij_1} \leqslant \hat{g}_{ij_2} \quad i \in \mathcal{I}, j_1, j_2 \in \mathcal{J}, \\ g_{ij} \in [0,1] \quad i \in \mathcal{I}, j \in \mathcal{J}, \text{ согласно (175).} \end{cases}$$

По векторам $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_n$ затем однозначно восстанавливаются векторы $\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_n$ как упорядоченные векторы, содержащие различные элементы $\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_n$. Для решения этой задачи используется алгоритм градиентного спуска, описанный в [54].

**Шаг 3.** Наконец, при фиксированных $\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_n, \hat{\mathbf{u}}$ оценим вектор $\mathbf{b}_0$ и $\hat{\mathbf{y}} = \boldsymbol{g}(\mathbf{b}_0, \mathbf{y})$:

$$\hat{\mathbf{b}}_0 = \arg\min_{\mathrm{Ord}(\mathbf{b}_0)} \|\boldsymbol{\xi}(\mathbf{b}_0, \hat{\mathbf{G}}\hat{\mathbf{u}}) - \boldsymbol{g}(\mathbf{b}_0, \mathbf{y})\|_2^2.$$

## 5.3. Согласование экспертных оценок в ранговых шкалах

В данной работе предполагается, что эксперты выставляют оценки качества объектов и важности показателей в ранговых шкалах. Предлагаемый метод базируется на идеях метода уточнения экспертных оценок, выставленных в линейных шкалах. Согласно этому методу, интегральный индикатор объектов можно оценить двумя путями: непосредственно через экспертную оценку $\mathbf{y}_0$ и через взвешенную сумму значений показателей объектов $\mathbf{y}_1 = \mathbf{X}\mathbf{w}_0$, где веса являются экспертными оценками показателей. В общем случае оценки $\mathbf{y}_0$ и $\mathbf{y}_1$ различны. Требуется построить интегральный индикатор, основанный на измеряемых данных и не противоречащий оценкам экспертов.

### 5.3.1. Постановка задачи

*Согласованными значениями* интегрального индикатора и весов показателей называются такие векторы $\mathbf{y}$ и $\mathbf{w}$, при которых выполняются условия

$$\begin{array}{rcl} \mathbf{y} & = & \mathbf{X}\mathbf{w}, \\ \mathbf{w} & = & \mathbf{X}^{+}\mathbf{y}, \end{array} \tag{177}$$

где $\mathbf{X}^{+}$ — линейное отображение, псевдообратное отображению $\mathbf{X}$, такое, что $\mathbf{X}\mathbf{X}^{+}\mathbf{X} = \mathbf{X}$, $\mathbf{X}^{+}\mathbf{X}\mathbf{X}^{+} = \mathbf{X}^{+}$ и $(\mathbf{X}\mathbf{X}^{+})^{\intercal} = \mathbf{X}\mathbf{X}^{+}$, $(\mathbf{X}^{+}\mathbf{X})^{\intercal} = \mathbf{X}^{+}\mathbf{X}$. Задачей предлагаемого метода является такое уточнение экспертных оценок, которое соответствовало бы условию (177).

Заданы экспертные оценки $\mathbf{y}_0, \mathbf{w}_0$, допускающие произвольные монотонные преобразования. Задана матрица описаний объектов $\mathbf{X} \in \mathbb{R}^{m\times n}$, удовлетворяющая условию (158). Без ограничения общности будем считать, что на наборах экспертных оценок введено отношение порядка такое, что

$$y_1 \geqslant y_2 \geqslant ... \geqslant y_m \geqslant 0 \;\; \text{и} \;\; w_1 \geqslant w_2 \geqslant ... \geqslant w_n \geqslant 0. \tag{178}$$

Для выполнения этого условия достаточно переставить элементы векторов $\mathbf{y}_0$, $\mathbf{w}_0$ и соответствующие им строки и столбцы матрицы $\mathbf{X}$ местами.

Условие (178) представимо в виде системы линейных неравенств (приведены только оценки интегральных индикаторов)

$$\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_{m-1} \\ y_m \end{pmatrix} \geqslant \mathbf{0}.$$

Обозначим двухдиагональную матрицу $\mathbf{J}$ и перепишем (178) в виде

$$\mathbf{J}_m\mathbf{y} \geqslant \mathbf{0} \;\; \text{и} \;\; \mathbf{J}_n\mathbf{w} \geqslant \mathbf{0}.$$

Число строк квадратной матрицы $\mathbf{J}$ равно числу неравенств в системе, а число элементов каждой строки равно числу элементов вектора ($\mathbf{y}$ или $\mathbf{w}$).

Обозначим конусы, заданные экспертными оценками в пространстве интегральных индикаторов и в пространстве весов показателей, соотвественно $\mathcal{Q}$ и $\mathcal{W}$:

$$\begin{array}{rcl} \mathcal{Q} & = & \{\mathbf{y}|\mathbf{J}_m\mathbf{y} \geqslant \mathbf{0}\}, \\ \mathcal{W} & = & \{\mathbf{w}|\mathbf{J}_n\mathbf{w} \geqslant \mathbf{0}\}. \end{array} \tag{179}$$

Нижний индекс 0, указывающий на то, что оценка поставлена экспертом, опущен, так как векторы $\mathbf{y}, \mathbf{w}$ рассматриваются как произвольные элементы множеств.

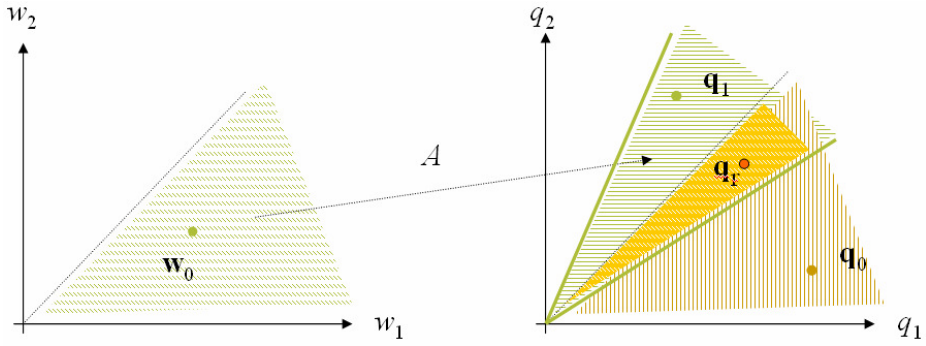



Рис. 60. Конусы в пространстве экспертных оценок показателей и интегральных индикаторов.

Линейное отображение $\mathbf{X}$ переводит конус $\mathcal{W} \ni \mathbf{w}_0$ экспертных оценок показателей (179) в вычисленный конус $\mathbf{X}\mathcal{W} = \mathcal{P} \ni \mathbf{w}_1$ (см. рис. 60):

$$\mathbf{X} : \mathcal{W} \to \mathcal{P},$$
$$\mathbf{X} : \mathbf{w}_0 \mapsto \mathbf{y}_1.$$

Рассмотрим следующие варианты:

1) конусы $\mathcal{P}$ и $\mathcal{Q}$ пересекаются, в этом случае экспертные оценки считаются согласованными и найдется такая пара $\mathbf{y}_p \in \mathcal{P} \cup \mathcal{Q}, \mathbf{w}_p = \mathbf{X}^+\mathbf{y}_p \in \mathcal{W}$, которая удовлетворяет условию согласованности (177);

2) пересечение конусов $\mathcal{P}$ и $\mathcal{Q}$ пусто, в этом случае требуется уточнение экспертных оценок.

Эти варианты рассмотрены разделах 4.3, 4.4.

### 5.3.2. Отображение и пересечение многогранных конусов

Для обоснования предложенного ниже алгоритма приведем некоторые свойства конусов. Множество точек $\mathcal{Q}$ в $\mathbb{R}^m$ называется *конусом*, если для любой точки $\mathbf{y} \in \mathcal{Q}$ точка $\lambda\mathbf{y}$ также принадлежит $\mathcal{Q}$. *Выпуклым многогранным конусом* называется пересечение конечного числа полупространств, граничные плоскости которых проходят через общую точку. Эта точка называется *вершиной конуса*.

Выпуклый многогранный конус с вершиной в начале координат — это область решений системы однородных неравенств:

$$\left\{ \begin{array}{ccccccccc} x_{11}w_1 & + & x_{12}w_2 & + & \ldots & + & x_{1n}w_n & \geqslant & 0, \\ x_{21}w_1 & + & x_{22}w_2 & + & \ldots & + & x_{2n}w_n & \geqslant & 0, \\ \ldots & & \ldots & & \ldots & & \ldots & & \ldots \\ x_{m1}w_1 & + & x_{m2}w_2 & + & \ldots & + & x_{mn}w_n & \geqslant & 0. \end{array} \right.$$

Согласно приведенному определению, система неравенств $\mathbf{Jw} \geqslant \mathbf{0}$ задает многогранный конус.

Непосредственно из этого следует утверждение: пересечение многогранных конусов с вершиной в начале координат является многогранным конусом. Действительно, рассмотрим два многогранных конуса. Выпуклый многогранный конус с вершиной в начале координат — это область решения некоторой системы однородных неравенств. Пусть первому конусу соответствует система неравенств $\mathbf{X}_1\mathbf{w} \geqslant \mathbf{0}$, а второму — $\mathbf{X}_1\mathbf{w} \geqslant \mathbf{0}$. Пересечение двух конусов — область решений системы, составленной из неравенств обеих систем, соответствующих конусам. Другими словами, пересечение двух данных конусов задается системой однородных неравенств с матрицей

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}.$$

Утверждение: множество $\mathcal{W}$ всех векторов $\mathbf{w} = \langle w_1, \ldots, w_n \rangle$, удовлетворяющих условиям $w_1 \geqslant w_2 \geqslant \ldots \geqslant w_n \geqslant 0$, является конусом. Действительно, если вектор $\mathbf{w}$ принадлежит множеству $\mathcal{W}$, то для любого $\lambda \geqslant 0$ справедливо неравенство $\lambda w_1 \geqslant \lambda w_2 \geqslant \ldots \geqslant \lambda w_n \geqslant 0$, поэтому вектор $\lambda\mathbf{w}$ также принадлежит множеству векторов $\mathcal{W}$.

Утверждение: геометрическое место точек, в которое отображение $\mathbf{X} : \mathcal{W} \to \mathcal{Q}$ переводит конус, является конусом. Действительно, для любого вектора $\mathbf{w}$, принадлежащего конусу $\mathcal{W}$, вектор $\lambda\mathbf{w}$ также ему принадлежит, а $\mathbf{y} = \mathbf{Xw}$. Поэтому, если вектор $\mathbf{y}$ принадлежит рассматриваемому геометрическому месту точек, то и вектор $\lambda\mathbf{y} = \lambda\mathbf{Xw} = \mathbf{X}(\lambda\mathbf{w})$ ему принадлежит.

Таким образом, если $\mathcal{W}$ – многогранный конус, то отображение $\mathbf{X}$ переводит его в многогранный конус $\mathcal{P} = \mathbf{X}\mathcal{W}$. Соответствующее псевдообратное отображение $\mathbf{X}^+$ переводит конус $\mathcal{W}$ в конус $\mathbf{X}^+\mathcal{W}$.

Утверждение: если конусы, задаваемые в пространстве интегральных индикаторов системами линейных неравенств $B_1\mathbf{y} \geqslant \mathbf{0}$ и $B_2\mathbf{y} \geqslant \mathbf{0}$, пересекаются, то их отображение в пространстве весов показателей тоже пересекаются. Действительно, рассмотрим отображение $\mathbf{X}\mathcal{W} \to \mathcal{Q}$. Так как по условию теоремы конусы пересекаются, то найдется вектор $\mathbf{y}$, такой что $(B_1\mathbf{X})\mathbf{w} \geqslant \mathbf{0}$ и $(B_2\mathbf{X})\mathbf{w} \geqslant \mathbf{0}$, то есть, конусы в пространстве $\mathcal{W}$ тоже пересекаются.

В контексте рассматриваемой задачи, если в пространстве интегральных индикаторов многогранные конусы, задаваемые неравенствами $\mathbf{J}_m\mathbf{y} \geqslant \mathbf{0}$ и $\mathbf{X}\mathbf{J}_n\mathbf{w} \geqslant \mathbf{0}$, пересекаются, то их псевдообратные отображения в пространство весов показателей $\mathbf{X}^+\mathbf{J}_m\mathbf{y} \geqslant \mathbf{0}$ и $\mathbf{J}_n\mathbf{w} \geqslant \mathbf{0}$, тоже пересекаются. Обозначим пересечения конусов в соответствующих пространствах как $\mathcal{W}_p = \mathcal{W} \cup \mathbf{X}^+\mathcal{Q}$ и $\mathcal{Q}_p = Q \cup \mathbf{X}\mathcal{W}$.

Если конус $\mathcal{Q}_p$ не пуст, то не пуст также и конус $\mathcal{W}_p$. В противном случае оба конуса пусты. Действительно, пусть конус $\mathcal{Q}_p$ не пуст, значит, существует вектор $\mathbf{y}_p$ такой, что принадлежит конусам $\mathcal{Q}$ и $\mathbf{X}\mathcal{W}$ одновременно. Покажем что конус $\mathcal{W}_p$ не пуст. Рассмотрим векторы $\mathbf{y}_p = \mathbf{X}\mathbf{w}_p \in \mathcal{Q}_p$ и $\mathbf{w}_p = \mathbf{X}^+\mathbf{y}_p\mathcal{W}_p$. Линейное отображение $\mathbf{X}^+$ переводит конус $\mathcal{Q}$ в конус $\mathbf{X}^+\mathcal{Q}$. Векторы $\mathbf{y} \in \mathbf{X}\mathcal{W}$, вектор $\mathbf{w}_p \in \mathcal{W}$ (линейное отображение $\mathbf{X} : \mathcal{W} \to \mathbf{X}\mathcal{W}$). Таким образом, вектор $\mathbf{w}_p$ принадлежит конусу $\mathcal{W}_p$ — пересечению конусов $\mathcal{W}$ и $\mathbf{X}^+\mathcal{Q}$.

Пусть теперь конус $\mathcal{Q}_p$ пуст. Покажем от противного, что конус $\mathcal{W}_p$ также пуст. Если это не так, то существует вектор $\mathbf{w}_p$, одновременно принадлежащий конусам $\mathcal{W}$ и $\mathbf{X}^+\mathcal{Q}$. Рассмотрим вектор $\mathbf{y}_p = \mathbf{X}\mathbf{w}_p$. Аналогичными рассуждениями приходим к выводу, что вектор $\mathbf{y}_p$

принадлежит конусам $\mathcal{Q}$ и $\mathbf{X}\mathcal{W}$, то есть конусу $\mathcal{Q}_p$. То есть, конус $\mathcal{Q}_p$ не пуст. Полученное противоречие показывает, что конус $\mathcal{W}_p$ пуст.

Доказанное утверждение эквивалентно следующему: для каждого вектора $\mathbf{w}_p$, принадлежащего конусу $\mathcal{W}_p$, найдется согласованный с ним вектор $\mathbf{y}_p \in \mathcal{Q}_p$, такой, что выполняются условия (177).

Для отыскания пресечения конусов $\mathcal{Q}_p$ опишем соответствующие множества системами линейных неравенств. Представим конус $\mathcal{Q}_p$, элементы которого удовлетворяют условию (178) в виде двухдиагональной матрицы $\mathcal{Q}_0$, в которой элементы на главной диагонали равны 1, а элементы на диагонали $(1,2),...,(n-1,n)$ равны $-1$. Представим отображение $(\mathbf{X}\mathcal{W})$ также в виде матрицы коэффициентов в пространстве $\mathbb{R}^{m\times m}$. Множество векторов $\mathcal{Q}_p \ni \mathbf{y}_p$ является решением объединенной системы линейных неравенств

$$\begin{cases} \mathcal{Q}\mathbf{y} & \geqslant & 0, \\ (\mathbf{X}\mathcal{W})\mathbf{y} & \geqslant & 0. \end{cases} \tag{180}$$

Полученное пересечение $\mathcal{Q}_p$ также является конусом (возможно, тривиальным), каждый элемент которого является интегральным индикатором, удовлетворяющим условию согласованности (177).

### 5.3.3. Уточнение оценок в случае непересекающихся конусов

В случае пустого пересечения конусов $\mathcal{Q}_p = \mathcal{Q} \cap \mathbf{X}\mathcal{W}$ и $\mathcal{W}_p = \mathcal{W} \cap \mathbf{X}^+\mathcal{Q}$ предлагается использовать модифицированный метод уточнения экспертных в линейных шкалах. В пространстве интегральных индикаторов рассмотрим лучи, заданные векторами $\mathbf{y} \in \mathcal{Q}$ и $\mathbf{p} \in \mathcal{P} = \mathbf{X}\mathcal{W}$. Найдем ближайшие друг к другу лучи на ребрах или гранях конусов $\mathcal{Q}, \mathcal{P}$,

$$\cos(\mathbf{y},\mathbf{p}) = \frac{\mathbf{y}^\intercal\mathbf{p}}{\|\mathbf{y}\|\|\mathbf{p}\|} \to \max.$$

и выполним процедуру уточнения (56) на точках, задающих эти лучи. Отыскиваемая пара $\mathbf{y}, \mathbf{p}$ должна выполнять следующие условия:

$$\begin{array}{ll} \text{maximize} & \mathbf{y}^\intercal\mathbf{p} \\ \text{subject to} & \mathbf{y}^\intercal\mathbf{y} = 1, \quad \mathbf{p}^\intercal\mathbf{p} = 1, \\ & \mathbf{J}_n\mathbf{y} \geqslant \mathbf{0} \quad \mathbf{X}\mathbf{J}_m\mathbf{p} \geqslant \mathbf{0}. \end{array}$$

Построим итерационный алгоритм, последовательно находящий приближения векторов $\mathbf{y}^{(2k)}, \mathbf{p}^{(2k+1)}$ на четном и нечетном шаге. Векторы $\mathbf{x} = \mathbf{y}^{(2k)}$ и $\mathbf{y} = \mathbf{p}^{(2k+1)}$ будем считать решениями двух последовательно решаемых оптимизационных задач, полагая произвольным вектор $\mathbf{p}^{(0)} \in \mathcal{P}$ на шаге $k = 0$.

| Задача $2k$ : | | Задача $2k+1$ : | |
|---|---|---|---|
| maximize | $\mathbf{x}^\intercal\mathbf{p}^{(2k)}$ | maximize | $\mathbf{y}^{T(2k+1)}\mathbf{y}$ |
| subject to | $\mathbf{x}^\intercal\mathbf{x} = 1,$ | subject to | $\mathbf{y}^\intercal\mathbf{y} = 1,$ |
| | $\mathbf{J}_n\mathbf{x} \geqslant 0.$ | | $\mathbf{X}\mathbf{J}_m\mathbf{y} \geqslant 0.$ |

При решении задач, на каждом шаге значение констант $\mathbf{p}^{(2k)}$ и $\mathbf{y}^{(2k+1)}$ принимается равным значениям соответствующих решений $\mathbf{x}$ и $\mathbf{y}$ предыдущего шага. Так как максимизируемые

функции и ограничения обеих задач являются выпуклыми, то решение будет найдено за счетное число шагов. Методы выпуклой оптимизации, используемые для получения численных решений, хорошо исследованы и описаны, например, в [54, 322].

Получив решения задачи — векторы $\hat{\mathbf{p}}$ и $\hat{\mathbf{y}}$, выполняем процедуру линейного уточнения оценок интегрального индикатора

$$\mathbf{y}_\alpha = (1-\alpha)\hat{\mathbf{p}} + \alpha\hat{\mathbf{y}},$$

при условии существования нетривиального решения $\mathbf{y}_\alpha$, то есть, $\hat{\mathbf{p}}^\intercal\hat{\mathbf{y}} \neq -1$. Как было показано ранее, вектор $\mathbf{y}_\alpha$ и соответствующий ему вектор $\mathbf{w}_\alpha = \mathbf{X}^+\mathbf{y}_\alpha$ удовлетворяют условию согласованности (177). Эти векторы задают в соответствующих пространствах конусы $\mathcal{W}$ и $\mathcal{Q}$, причем пересечение $\mathcal{W}_p = \mathbf{X}\mathcal{W} \cap \mathcal{Q}$ не пусто. Так же, как и в случае уточнения оценок у линейных шкалах, при значении параметра $\alpha \to 0$, предпочтение отдается экспертным оценкам качества объектов. При $\alpha \to 1$ предпочтение отдается экспертным оценкам важности показателей.

В следующем разделе показано, как по уточненным оценкам, выставленным в ранговых шкалах, можно получить оценки в линейных шкалах.

## 5.4. Устойчивость и регуляризация при выборе моделей экспертных оценок

### 5.4.1. Получение непротиворечивых экспертных оценок

Для проверки непротиворечивости выставленных экспертных оценок рекомендуется выполнить процедуру парного сравнения. На множестве объектов или на множестве показателей экспертом задается отношение частичного порядка $\rho(y_i, y_j)$, такое, что

$$\rho_{ij} = \begin{cases} +1, & \text{если } y_i \succ y_j, \\ -1, & \text{если } y_i \prec y_j, \\ 0, & \text{в случае отказа от оценки.} \end{cases}$$

Антисимметричная матрица $\mathcal{R} = \{\rho_{ij}\}_{i,j=1}^m$ задает направленный граф, в котором узлами являются объекты. Направление ребер задано элементами матрицы. В случае обнаружения петель в графе (например, петель вида $y_i \succ y_j \succ y_k \succ y_i$) требуется пересмотреть экспертные оценки с целью исключения петель. Экспертная оценка $y_i$ в ранговой шкале есть число ребер, исходящей из $i$-й вершины графа.

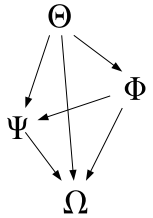



Рис. 61. Пример графа, построенного по матрице парных сравнений.

В качестве примера приведем сравнение четырех объектов: $\{\Theta, \Phi, \Omega, \Psi\}$. Пусть матрица парных экспертных предпочтений $\mathcal{R}$ задана как

| | $\Theta$ | $\Phi$ | $\Omega$ | $\Psi$ |
|---|---|---|---|---|
| $\Theta$ | 0 | $+1$ | $+1$ | $+1$ |
| $\Phi$ | $-1$ | 0 | $+1$ | $+1$ |
| $\Omega$ | $-1$ | $-1$ | 0 | $-1$ |
| $\Psi$ | $-1$ | $-1$ | $+1$ | 0 |

.

Тогда граф, соответствующий этой матрице, будет выглядеть как на рис. 61, а вектор экспертных оценок $\mathbf{y}_0 = \langle y_1, \ldots, y_4 \rangle^\intercal = \langle 3, 2, 0, 1 \rangle^\intercal$.

При попадании несравнимых объектов в один класс эквивалентности число неравенств в линейной системе (180) сокращается: исключается строка с номером $i$, где $i, i+1$ — номера линейно упорядоченных объектов, отнесенных экспертом в один класс. Процедура уточнения экспертных оценок при этом остается неизменной.

### 5.4.2. Интегральные индикаторы, устойчивые к возмущению матрицы описаний

Рассмотрим найденный конус $\mathcal{Q}_p$ и матрицу «объект-показатель» $\mathbf{X}$. Возмутим элементы этой матрицы, $\mathbf{X} = \mathbf{X} + \Delta$, принимая гипотезу нормального распределения матрицы $\Delta = \delta I$, $\delta \sim \mathcal{N}(0, \sigma^2)$. Образ линейного отображения $\mathbf{y} = (\mathbf{X} + \Delta)\mathbf{w}$ будет также иметь нормальное распределение. Согласно принятой гипотезе, будем считать устойчивым к малому возмущению матрицы $\mathbf{X}$ такой интегральный индикатор $\mathbf{y}_p$, который наиболее удален от всех граней конуса $\mathcal{Q}_p$ при условии нормировки $\|\mathbf{y}_p\| = 1$. Вектор $\mathbf{y}_p$ является центром сферы, вписанной в конус $\mathcal{Q}_p$ и называется точкой Чебшёва.

Расстояние от искомого вектора $\mathbf{y}_p$ до граней $\mathbf{b}$ конуса отыскивается как решение оптимизационной задачи

$$\mathbf{y}_p^* = \arg \max_{\mathbf{y}_p \in \mathcal{Q}_p} \{\|\mathbf{y}_p - \mathbf{b}\|^2, \text{ где } \mathbf{b} \in \mathbb{R}^m \setminus \mathcal{Q}_p \text{ и } \|\mathbf{y}_p\| \leqslant 1\}.$$

Рассмотрим систему из $L$ линейных неравенств (180), решение которой задает конус $\mathcal{Q}_p$. Обозначим $\mathbf{s}_\ell$ — вектор нормали, соответствующий строке с номером $\ell$ этой системы. Скалярное произведение $\mathbf{x}^\intercal \mathbf{s}_\ell = 0$ задает плоскость в пространстве интегральных индикаторов, проходящую через начало координат. Расстояние $d$ от вектора $\mathbf{y}_p$ до этой плоскости равно

$$d(\mathbf{y}_p, \mathbf{s}_\ell) = \frac{\mathbf{y}_p^\intercal \mathbf{s}_\ell}{\|\mathbf{s}_\ell\|}.$$

Эта задача является задачей выпуклой оптимизации и представима в виде

$$\begin{aligned} \text{maximize} \quad & \inf_{\ell=1,\ldots,L} \left(\mathbf{x}^\intercal \mathbf{s}_\ell \|\mathbf{s}_\ell\|^{-1}\right) \\ \text{subject to} \quad & \mathbf{x}^\intercal \mathbf{x} = 1, \\ & \mathbf{J}_n \mathbf{x} \geqslant \mathbf{0}, \\ & \mathbf{X}\mathbf{J}_m \mathbf{x} \geqslant \mathbf{0}. \end{aligned}$$

Методы выпуклой оптимизации, используемые для численного решения данной задачи описаны в [54, 76]. Результат решения — вектор $\hat{\mathbf{x}} = \mathbf{y}_p$ и вычисленный вектор весов показателей $\mathbf{w}_p = \mathbf{X}^+\mathbf{y}_p$ являются согласованными. Они получены с помощью экспертных оценок, выставленных в ранговых шкалах и могут быть использованы для построения интегральных индикаторов в линейных шкалах.

### 5.4.3. Регуляризация при согласовании экспертных оценок

При нахождении согласованных оценок требуется выбрать способ вычисления псевдообратного оператора $\mathbf{X}^+ : Q \longrightarrow W$. Предлагается следующее решение. Задано множество $\Omega = \{\omega_1, ..., \omega_k\}$, алгоритмов вычисления псевдообратного оператора $\mathbf{X}^+$. Из данного множества выбирается такой алгоритм $\omega$, что для полученного $\mathbf{X}^+ = \mathbf{X}^+(\omega)$ имеет место $\min_{\omega\in\Omega}(\frac{\varepsilon^2}{m-1} + \frac{\delta^2}{n-1})$, где $\varepsilon^2 = \|\hat{\mathbf{y}} - \mathbf{y}_0\|^2$, и $\delta^2 = \|\hat{\mathbf{w}} - \mathbf{w}_0\|^2$.

Для решения задачи предложены два способа нахождения псевдообратного оператора $\mathbf{X}^+$: регуляризация псевдообратного оператора методом Тихонова и обращение усеченного сингулярного разложения. В первом случае найден псевдообратный оператор $\mathbf{X}^+ = (\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})^{-1}$ со значением регуляризующего параметра $\gamma^2$, см. выражение (170).

Алгоритм обращения матрицы посредством усеченного сингулярного разложения состоит в следующем. Пусть матрица исходных данных $\mathbf{X}$ представлена в виде $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$. Тогда при нахождении обратной матрицы $\mathbf{X}^+ = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{U}^\intercal$ в силу ортогональности матриц $\mathbf{U}$ и $\mathbf{V}$: $\mathbf{U}^\intercal\mathbf{U} = \mathbf{V}\mathbf{V}^\intercal = I$ и в силу условия убывания диагональных элементов матрицы $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ псевдообратная матрица $\mathbf{X}^+$ будет более зависеть от тех элементов матрицы $\mathbf{\Lambda}$, которые имеют меньшие значения, чем от первых сингулярных чисел. Действительно, если по условию теоремы о сингулярном разложении матрица $\mathbf{X}$ имеет сингулярные числа $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_n$, то сингулярные числа матрицы $\mathbf{X}^+$ равны $\mathbf{\Lambda}^{-1} = \mathrm{diag}(\frac{1}{\lambda_1}, ..., \frac{1}{\lambda_n})$ и $\frac{1}{\lambda_1} \leqslant \frac{1}{\lambda_2} ... \leqslant \frac{1}{\lambda_n}$. Считая первые $r$ сингулярных чисел определяющими собственное пространство матрицы $\mathbf{X}$, используем при обращении матрицы $\mathbf{X}$ первые $r$ сингулярных чисел. Тогда обратная матрица $\mathbf{X}^+$ будет найдена как $\mathbf{X}^+ = \mathbf{V}\mathbf{\Lambda}_r^{-1}\mathbf{U}^\intercal$.

Для обоснования предложенных методов согласования докажем следующие теоремы. Лемма о непрерывности обратного отображения, впервые сформулированная А. Н. Тихоновым, приведена в обозначениях, принятых ранее в настоящей работе.

**Лемма 1** (А. Н. Тихонова)**.** *Пусть метрическое пространство $\mathbb{W}$ отображается на метрическое пространство $\mathbb{Q}$ и $Q$ — образ множества $W$, $W \subset \mathbb{W}$, при этом отображении. Если отображение $\mathbf{X} : \mathbb{W} \longrightarrow \mathbb{Q}$ непрерывно, взаимнооднозначно и множество $W$ компактно на $\mathbb{W}$, то обратное отображение $\mathbf{X}^+ : Q \longrightarrow W$ множества $Q$ на множество $W$ также непрерывно по метрике пространства $\mathbb{W}$.*

Тройка $(\mathbf{y}, \mathbf{w}, \mathbf{X})$ определена на следующих метрических пространствах. Вектор $\mathbf{y}$ является элементом $Q$, где область $Q$ является компактной в $\mathbb{Q}$: $Q \subset \mathbb{Q} \equiv \mathbb{R}^m$, так как область $Q$ замкнута и ограничена. Также вектор $\mathbf{w}$ является элементом $W$, где область $W$ является компактной в $\mathbb{W}$: $W \subset \mathbb{W} \equiv \mathbb{R}^n$, так как область $W$ замкнута и ограничена. Метрика задается нормами векторов $\|\mathbf{y}\|^2$ для компакта $Q$ и $\|\mathbf{w}\|^2$ для компакта $W$. Функционал $\rho_{\mathbb{Q}} = \rho_{\mathbb{Q}}(\mathbf{X}\mathbf{w}, \mathbf{y})$ определим как $\rho_{\mathbb{Q}} = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$.

**Следствие 2.** *Псевдообратный оператор* $\mathbf{X}^+$*, определенный как* $\mathbf{X}^+ = (\mathbf{X}^\intercal\mathbf{X} + \gamma^2\mathbf{I})^{-1}$*, является непрерывным по метрике пространства* $\mathbb{W}$*.*

**Теорема 16.** *Оператор* $\mathbf{X}^+ = \mathbf{U}^\intercal\mathbf{\Lambda}_r\mathbf{V}$*, полученный методом обращения усеченного сингулярного разложения, является непрерывным в* $r$*-мерном подпространстве.*

*Доказательство.* Отметим, что оператор $\mathbf{X}$ является обратным для оператора $\mathbf{X}^+ : \mathbb{Q} \longrightarrow \mathbb{W}$. Оператор $\mathbf{X}^+$ определен в пространстве $\mathbb{R}^r$, так как согласно теореме о сингулярном разложении матрицы $\mathbf{U}$ и $\mathbf{V}$ являются ортогональными, а матрица $\mathbf{\Lambda}$ является диагональной. Матрица $\mathbf{\Lambda}_r$ получается из матрицы $\mathbf{\Lambda}$ путем замены части диагонали, начиная с элемента с номером $r+1$, нулевыми значениями. Прообраз $\mathbf{X}(G)$ всякого открытого в $\mathbb{W}$ множества $G$ открыт в $\mathbb{Q}$ в силу того, что $\mathbf{X}$ — линейный оператор. Также прообраз $\mathbf{X}(F)$ всякого замкнутого в $\mathbb{W}$ множества $F$ замкнут в $\mathbb{Q}$. Следовательно, оператор $\mathbf{X}^+$ непрерывен в $r$-мерном подпространстве. □

Так как оператор $\mathbf{X}$ в уравнении $\mathbf{X}\mathbf{w} = \mathbf{y}$ вполне непрерывный, то построение устойчивого к малым изменениям правой части $\mathbf{y}$ приближенного решения этого уравнения по формуле $\mathbf{y} = \mathbf{X}^+\mathbf{w}$ возможно в тех случаях, когда решение ищется на компакте $W \subset \mathbb{W}$ и правая часть уравнения принадлежит множеству $\mathbf{X}W$.

Покажем, что согласованные векторы $\hat{\mathbf{y}}, \hat{\mathbf{w}}$, получаемые с помощью процедур согласования, являются единственными.

**Утверждение 1.** *Для данного параметра* $\alpha \in (0,1)$ *и псевдообратного оператора* $\mathbf{X}^+$*, определенного как* $\mathbf{X}^+ = \mathbf{U}\mathbf{\Lambda}_r\mathbf{V}^\intercal$*, задача* $\alpha$*-согласования (163) имеет единственное решение* $(\mathbf{y}_\alpha, \mathbf{w}_\alpha, \mathbf{X})$*.*

**Утверждение 2.** *Для данного параметра* $\gamma^2 \in (0, \infty)$ *задача* $\gamma^2$*-согласования (170) имеет единственное решение* $(\mathbf{y}_\gamma, \mathbf{w}_\gamma, \mathbf{X})$*.*

Задача нахождения тройки $(\hat{\mathbf{y}}, \hat{\mathbf{w}}, \mathbf{X})$ называется корректно поставленной на паре метрических пространств $(\mathbb{Q}, \mathbb{W})$, если удовлетворяются условия:

1) для всякого элемента $\hat{\mathbf{y}} \in \mathbb{Q}$ существует решение $\hat{\mathbf{y}} \in \mathbb{Q}$;

2) решение определяется однозначно;

3) задача устойчива на пространствах $\mathbb{Q}, \mathbb{W}$.

Таким образом, мы получили решения задач (163) и (170), корректные по Адамару.

### 5.4.4. Устойчивые интегральные индикаторы с выбором опорного множества описаний объектов

Нижеприведенный метод разделяет исходное множество описаний объектов на два подмножества — опорное, и множество выбросов. При этом используется критерий вероятности принадлежности описаний объекта одному из двух подмножеств. По опорному множеству, с помощью метода главных компонент, вычисляются веса. Эти веса используются для получения интегральных индикаторов всей выборки.

Для получения интегральных индикаторов, устойчивых к выбросам, в рамках линейной модели ранее было предложено использовать регуляризацию. А. М. Шурыгин в работе [360] рассмотрел два способа регуляризации ковариационной матрицы $\mathbf{\Sigma}$. Первый способ — регуляризация посредством ридж-регрессии, $\mathbf{\Sigma}_{r\beta} = \mathbf{\Sigma} + \beta\mathbf{I}$, где $\beta$ — регуляризующий множитель. Второй способ — диагональная регуляризация $\mathbf{\Sigma}_{d\nu} = (1-\nu)\mathbf{\Sigma} + \nu\mathrm{diag}(\mathbf{\Sigma})$, где $\nu \in [0,1]$ — регуляризующий множитель. Было показано, что второй способ дает лучшую устойчивость к выбросам.

Использование регуляризации приводит к потере информативности. Поставим задачу так, чтобы сохранить значение критерия наибольшей информативности на опорном множестве описаний.

Задано множество описаний объектов, $S_0 = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$. Обозначим $\mathcal{S} = \{S_1, ..., S_l\}$ — множество всех подмножеств $S_0$, в котором число элементов $l = 2^m$. Алгоритм, вычисляющий наиболее информативный линейный предиктор, использует множество $S_\xi$, отыскивает веса $\mathbf{w}_\xi = \mathbf{w}(S_\xi) \in \mathbb{R}^n$ и возвращает интегральный индикатор $\mathbf{y}_\xi = \mathbf{X}\mathbf{w}_\xi \in \mathbb{R}^m$. Обозначим $\bar{S}_\xi$ дополнение $S_\xi$ до $S_0$. Исключим из рассмотрения тривиальные пары $(S_\xi, \bar{S}_\xi)$, в которых $\#S_\xi = 1$ и $\bar{S}_\xi = \varnothing$. Будем считать, что значения показателей объектов являются независимыми случайными величинами и принята гипотеза Гауссовского распределения этих величин.

Пусть $p_\xi = P(\mathbf{x}_i \in S_\xi)$ обозначает вероятность принадлежности некоторого объекта из $S_0$ множеству $S_\xi$, и $\bar{p}_\xi$ — вероятность того, что этот объект принадлежит дополнению до $S_0$. Найдем в $\mathcal{S}$ такое опорное множество $S_\xi$, для которого отношение $f_\xi = p_\xi/\bar{p}_\xi$ максимально.

Рассмотрим суммарные дисперсии $\sigma_\xi$ и $\bar{\sigma}_\xi$ проекций $\mathbf{p}_i$ элементов $\mathbf{x}_i$ множеств $S_\xi$ и $\bar{S}_\xi$ на первые главные компоненты, определяемые матрицей $S_\xi$. Обозначим $n_\xi, \bar{n}_\xi, n_0$ — число элементов во множествах $S_\xi, \bar{S}_\xi, S_0$ соответственно. Суммарная дисперсия проекций $\mathbf{p}_i$ элементов множеств $S_\xi$ и $\bar{S}_\xi$ всей выборки $\sigma^2(S_0)$ равна сумме дисперсий каждой выборки, взвешенных вероятностями принадлежности вектора $\mathbf{x}_i$ с проекцией $\mathbf{p}_i$ множествам $S_\xi, \bar{S}_\xi$,

$$\sigma^2(S_0) = {p_\xi}^2\sigma^2(S_\xi) + {p_\xi}^2\sigma^2(\bar{S}_\xi) = \frac{{p_\xi}^2{\sigma_\xi}^2}{n_\xi} + \frac{\bar{p}_\xi^2\bar{\sigma}_\xi^2}{\bar{n}_\xi}. \tag{181}$$

Для получения выражения отношения вероятностей $f_\xi$ минимизируем дисперсию $\sigma^2(S_0)$. Так как выражение (181) должно удовлетворять равенству $n_\xi + \bar{n}_\xi = n_0$, при дифференцировании используем метод множителей Лагранжа, обозначив множитель $\lambda$. Тогда

$$L = \sigma^2(S_0) + \lambda(n_\xi + \bar{n}_\xi - n_0) = \frac{{p_\xi}^2{\sigma_\xi}^2}{n_\xi} + \frac{\bar{p}_\xi^2\bar{\sigma}_\xi^2}{\bar{n}_\xi} + \lambda(n_\xi + \bar{n}_\xi - n_0).$$

Приравняв частные производные по $\lambda$ и по $n_\xi$ к нулю, получаем

$$\frac{\partial L}{\partial n_\xi} = -\frac{{p_\xi}^2{\sigma_\xi}^2}{{n_\xi}^2} + \lambda = 0, \quad \frac{\partial L}{\partial \lambda} = n_\xi + \bar{n}_\xi - n_0 = 0,$$

откуда получаем $p_\xi\sigma_\xi = n_\xi\sqrt{\lambda}$. Из двух последних выражений $n_0\sqrt{\lambda} = (p_\xi\sigma_\xi + \bar{p}_\xi\bar{\sigma}_\xi)$ и $p_\xi = n_\xi(p_\xi\sigma_\xi + \bar{p}_\xi\bar{\sigma}_\xi)(n_0\sigma_\xi)^{-1}$. Продифференцировав лагранжиан $L$ по $\bar{n}_\xi$, получим аналогичное отношение для вероятности $\bar{p}_\xi$. Искомое отношение вероятностей равно

$$\frac{p_\xi}{\bar{p}_\xi} = \frac{n_\xi\bar{\sigma}_\xi}{\bar{n}_\xi\sigma_\xi}. \tag{182}$$

Таким образом, вероятность принадлежности описания объекта одному из множеств прямо пропорциональна мощности этого множества и обратно пропорциональна среднеквадратичному отклонению. Искомый интегральный индикатор $\mathbf{y}_\xi = \mathbf{X}\mathbf{w}_\xi$ доставляется таким множеством $S_\xi$, для которого отношение $f_\xi = \frac{n_\xi \bar{\sigma}_\xi}{\bar{n}_\xi \sigma_\xi}$ максимально.

В качестве примера приведены результаты сравнительного анализа регионов России по уровню загрязнения ртутью основных продуктов питания. Каждому региону поставлен в соответствие интегральный индикатор, указывающий на загрязненность продуктов. Рассматриваются три показателя загрязненности: мясные продукты, молочные продукты и хлебобулочные изделия. Используются данные 29 регионов. Данные нормированы следующим образом. В каждом регионе для каждого из трех показателей был проведен ряд стандартизованных измерений. Элемент $x_{ij}$ матрицы описаний — величина загрязнения $j$-го продукта в $i$-м регионе. Его значение есть отношение квантиля уровня 0.9 распределения содержания ртути в серии измерений к величине предельно допустимой концентрации ртути в данном продукте.

Найдем опорное множество $S_\xi$ с целью вычисления весов показателей $\mathbf{w}_\xi$ для получения интегральных индикаторов, устойчивых к выбросам. Алгоритм состоит из трех шагов: назначения начального опорного множества, отыскания опорного множества и вычисления интегрального индикатора.

1. Отыскивается центр исходного множества. Для этого находится вектор-среднее по всем компонентам векторов $\mathbf{x}_i$, вошедших в выборку $S_0$, и изымается вектор, наиболее удаленный в евклидовой метрике. Это действие производится итеративно, до получения последнего вектора, который и является центром. Для сокращения времени работы алгоритма, две трети описаний объектов, наименее удаленных от центра, заносятся в ядро опорного множества.

2. Исходное множества $S_0$ разбивается на множества $S_\xi$ и $\bar{S}_\xi$ таких, что $S_\xi$ включает ядро опорного множества в качестве собственного подмножества, а $\bar{S}_\xi$ являются объектами-выбросами. Для каждого разбиения вычисляется целевая функция $f_\xi = \frac{n_\xi \bar{\sigma}_\xi}{\bar{n}_\xi \sigma_\xi}$, где $n_\xi, \bar{n}_\xi$ — мощности множеств $S_\xi, \bar{S}_\xi$; и $\sigma_\xi, \bar{\sigma}_\xi$ — суммарная дисперсия проекций объектов множеств $S_\xi, \bar{S}_\xi$ на собственные векторы ковариационной матрицы, определяемой множествами $S_\xi, \bar{S}_\xi$. Из множества полученных функций $f_\xi$ выбираем функцию, на которой достигается максимум.

3. Объекты выбранного опорного множества $S_\xi$ задают матрицу "объект-показатель" $\mathbf{X}_\xi$. Для нее вычисляется ковариационная матрица $\mathbf{\Sigma} = \mathbf{X}_\xi^\intercal \mathbf{X}_\xi$. Первый собственный вектор матрицы $\mathbf{\Sigma}$ определяет веса $\mathbf{w}_\xi$ показателей исходного множества методом главных компонент [154]. Интегральный индикатор объектов, вычисленный с помощью предложенного алгоритма, есть $\mathbf{y}_\xi = \mathbf{X}\mathbf{w}_\xi$.

Множество исходных данных — описаний регионов — содержит три выброса по второму показателю (молочные продукты) в трех регионах: республика Карелия, г.Санкт-Петербург, Московская область. Данные Карелии, кроме того, содержат выброс по всем трем показателям. Эти три региона не входят в опорное множество объектов.

В таблице 1 показано распределение весов показателей, полученных для трех алгоритмов построения интегральных индикаторов. Первый алгоритм — применение метода главных компонент к исходным данным без использования регуляризации. Второй алгоритм — метод

Таблица 13. Веса показателей для алгоритма без регуляризации, с регуляризацией и с опорным множеством.

| **w** | Без регуляризации | С регуляризацией | С опорным множеством |
|---|---|---|---|
| $w_1$ | 0.0204 | 0.2264 | 0.4693 |
| $w_2$ | 0.9983 | 0.7687 | 0.7706 |
| $w_3$ | 0.0548 | 0.5982 | 0.4312 |

главных компонент с регуляризацией. Третий алгоритм — метод главных компонент для опорного множества описаний объектов. При использовании первого алгоритма выбросы по второму показателю приводят к неадекватному увеличению вклада этого показателя в интегральный индикатор. Предложенный метод доставляет более адекватные значения весов показателей, как показано в последнем столбце таблицы.

Для иллюстрации результатов работы алгоритмов введен критерий устойчивости

$$\varphi = \arg\min_{\Phi} \|\mathbf{w}_{\mathbf{X}} - \mathbf{w}_{\mathbf{X}^*}\|_2,$$

где множество $\Phi$ определено как

$$\Phi = \{\mathbf{x}^* | \ \|\mathbf{x}^*\|_2 = \max \|\mathbf{x}_i\|_2, i = 1, ..., m\}.$$

Вектор $\mathbf{w}_{\mathbf{X}}$ получен с помощью метода главных компонент для исходной матрицы $\mathbf{X}$. Вектор $\mathbf{w}_{\mathbf{X}^*}$ вычисляется с помощью метода главных компонент для матрицы $\mathbf{X}$ с присоединенным вектором-столбцом $\mathbf{x}^*$, который рассматривается как выброс. Значение критерия устойчивости $\varphi$ вычисляется для трех алгоритмов: без использования регуляризации, с диагональной регуляризацией и с предложенным алгоритмом выбора опорного множества, $\varphi = 0.4727$, $\varphi = 0.0962$ $\varphi = 0.0$, соответственно.

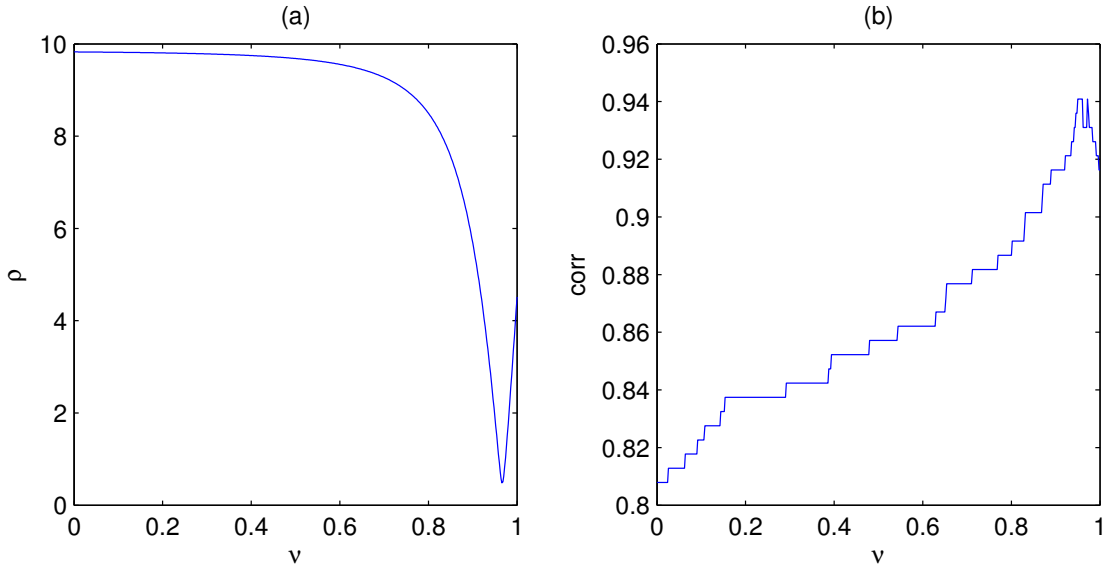



Рис. 62. Расстояние между регуляризованным и устойчивым интегральным индикатором.

Алгоритм, использующий диагональную регуляризацию, позволяет получить адекватный индикатор, но тем не менее влияние объектов-выбросов на индикатор полностью не

исключено. На рисунке 1(a) показана зависимость евклидова расстояния $\rho = \|\mathbf{y}_2 - \mathbf{y}_3\|$ от регуляризующего параметра. Вектор $\mathbf{y}_2$ — индикатор, полученный с помощью диагональной регуляризации, вектор $\mathbf{y}_3$ — индикатор, полученный с помощью алгоритма выбора опорного множества описаний объектов. При значении $\nu = 0.9660$ расстояние $\rho$ достигает минимума. На рисунке 1(b) показана зависимость ранговой корреляции между индикаторами $\mathbf{y}_2$ и $\mathbf{y}_3$. Максимальное значение коэффициента ранговой корреляция, вычисленного по формуле Кендалла, равно 0.94. Коэффициент ранговой корреляции вычисляется по формуле Кендалла.

$$\text{corr} = 1 - 6D^2\big(m(m^2-1)\big)^{-1},$$

где $\mathfrak{D}$ — число перестановок между теми значениями пар интегральных индикаторов объектов, которые имеют различный порядок следования, $m$ — число объектов. Коэффициент ранговой корреляции используется для сравнения в связи с тем, что он инвариантен относительно монотонных преобразований интегральных индикаторов и учитывает только порядок их значений, игнорируя при этом величину выбросов.

Таблица 14. Значения интегрального индикатора без регуляризации и построенного на основе опорного множества.

| Регион РФ | $\mathbf{y}_1$ | $r(\mathbf{y}_1)$ | $\mathbf{y}_3$ | $r(\mathbf{y}_3)$ |
|---|---|---|---|---|
| Архангельская область | 0.5367 | 19 | 0.8356 | 23 |
| Хабаровский край | 0.7986 | 21 | 0.6165 | 19 |
| ... | ... | ... | ... | ... |
| Владимирская область | 0.0324 | 12 | 0.3577 | 14 |
| Краснодарский край | 0.0449 | 16 | 0.1578 | 10 |

Алгоритм, не использующий регуляризацию, вычисляет интегральный индикатор, который существенно зависит от наличия в выборке объектов-выбросов. Коэффициент ранговой корреляции между интегральным индикатором, полученным посредством такого алгоритма, и между интегральным индикатором, полученным с помощью опорного множества, равен 0.82. Это означает, что у 37 пар, из всех возможных пар элементов двух индикаторов, порядок следования объектов отличается. В таблице 2 приведены примеры таких пар. В столбцах $\mathbf{y}_1$ и $\mathbf{y}_3$ приведены значения интегральных индикаторов указанных регионов. В столбцах $r(\mathbf{y}_1)$ и $r(\mathbf{y}_3)$ приведены ранговые номера регионов.

### 5.4.5. Построение коллаборативного интегрального индикатора

Совместный интегральный индикатор вычисляется по спискам публикаций за последние годы, находящимся в открытом доступе, с использованием алгоритма коллаборативной фильтрации. В качестве функционала качества используется функция близости интегральных индикаторов авторов и журналов, в которых они публикуют свои работы.

Построим интегральный индикатор качества научных публикаций. Рассматриваемый индикатор базируется на существующих методиках подсчета импакт-фактора (IF) [8] и индекса Хирша [133] и предназначен для более точной оценки эффективности научной работы.

В настоящее время понятие «качество журнала», помимо импакт-фактора, измеряется при помощи рейтингов, составляемых государственными структурами. В частности, в России издания делятся на «журналы из списка ВАК» [143] и прочие. Эффективность научной деятельности исследователя оценивается при помощи различных индексов [335], наиболее часто используемый из которых — индекс Хирша [133]. Упомянутые способы оценки качества журналов и успешности научной работы базируются на подсчете числа цитирования публикаций и имеют ряд недостатков [6].

Предлагается связать качество публикаций автора с качеством журнала, в котором он печатает свою работу и построить интегральный индикатор исходя из следующих принципов. Каждому автору можно поставить в соответствие список журналов, в которых он публиковал или хотел бы публиковать свои работы по некоторой тематике. Каждому журналу можно поставить в соответствие список авторов, опубликованных в журнале. Поэтому

1) более высокое значение индикатора имеет тот автор, который публикует свои работы в журналах с более высоким индикатором;

2) более высокое значение индикатора имеет тот журнал, в котором публикуют свои работы авторы с более высоким индикатором.

Для построения модели используются списки публикаций за последние годы [7], находящиеся в свободном доступе. Составляется матрица «журналы-авторы». Предполагается, что эта матрица разрежена, то есть каждый автор публикуется в небольшом, по сравнению с общим количеством, множестве журналов и каждый журнал печатает работы небольшой группы авторов. Для определения значения индикатора проводится кластеризация авторов и журналов с помощью алгоритма k-Means, построение ко-кластеров с использованием алгоритма коллаборативной фильтрации [221]. Затем ненулевые элементы внутри каждого ко-кластера концентрируются вблизи диагонали с помощью алгоритма редукции матриц Cuthill-McKee [74, 194]. Также по размерам полученных ко-кластеров оценивается интегрированность журналов и авторов в мировую науку. Чем больше размер кластера, тем большее значение индикатора получают входящие в него журналы и авторы.

Дана матрица $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_m^\intercal]^\intercal = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_n]$ «журналы-авторы», заполненная нулями и единицами $\mathbf{X} \in \{0,1\}^{m \times n}$. Строки $\mathbf{x}_i$ матрицы соответствуют авторам, столбцы $\boldsymbol{\chi}_j$ — журналам. Единица на пересечении строки $i \in \mathcal{I} = \{1, \ldots, m\}$ и столбца $j \in \mathcal{J} = \{1, \ldots, n\}$ означает, что $i$-й автор опубликовал работу в $j$-м журнале.

Требуется задать отношение линейного $\varphi$ порядка на множестве авторов $\mathcal{I}$:

$$\varphi \colon \mathcal{I} \to \{0,1\}^{m \times m}$$

и отношение линейного порядка $\psi$ на множестве журналов $\mathcal{J}$:

$$\psi \colon \mathcal{J} \to \{0,1\}^{n \times n}.$$

Для решения этой задачи предлагается посредством перестановки строк и столбцов матрицы получить ленточную матрицу, в которой элементы были бы как можно ближе к диагонали. Обозначим $\varphi \colon \mathcal{I} \to \hat{\mathcal{I}}$ и $\psi \colon \mathcal{J} \to \hat{\mathcal{J}}$ искомые перестановки строк и столбцов матрицы $\mathbf{X}$ соответственно.

Введем функционал качества диагонализации матрицы:

$$Q(\varphi, \psi) = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} x_{ij} |\varphi(i) - \psi(j)|, \tag{183}$$

Искомый алгоритм $(\hat{\varphi}, \hat{\psi})$ ранжирования авторов и журналов определяется решением задачи дискретной оптимизации

$$(\hat{\varphi}, \hat{\psi}) = \arg\min_{\varphi, \psi} Q(\varphi, \psi). \tag{184}$$

В предположении что автор, публикуют работы по некоторой определенной тематике, предлагается выделить кластеры «журналы-авторы», а затем вычислить интегральный индикатор внутри каждого кластера. Размеры получаемых ко-кластеров интерпретируются как «степень вовлеченности в мировое научное сообщество» входящих в него авторов и изданий. Предлагаемый алгоритм построения совместного интегрального индикатора включает четыре основных этапа. На первых трех (кластеризация авторов, кластеризация журналов, ко-кластеризация) находятся совместные кластеры «журналы-авторы». На последнем этапе (редукция ко-кластеров) путем оптимизации функционала качества (183) решается задача ранжирования (184).

*Кластеризация авторов* проводится алгоритмом $k$-Means. Число кластеров $K$ считаем фиксированным. Задаем начальное приближение положений центров кластеров $\mathbf{u}_q$, $q \in \{1, \ldots, K\}$. Затем для каждого элемента $\mathbf{x}_i$ находим ближайший к нему центр $\mathbf{u}_q$ и относим его к кластеру с номером $y_i = q$:

$$y_i = \arg\min_{q \in \{1,\ldots,K\}} \rho(\mathbf{x}_i, \mathbf{u}_q).$$

Осуществляем пересчет положений центров, помещая их в центр масс соответствующих кластеров:

$$\mathbf{u}_q = \frac{\sum_{i \in \mathcal{I}} [y_i = q] \mathbf{x}_i}{\sum_{i \in \mathcal{I}} [y_i = q]},$$

где индикатрная функция $[y_i = q]$ принимает значение 1, если $y_i = q$, и 0, если $y_i \neq q$. Алгоритм останавливается, когда кластеризация $y_i$ элементов $\mathbf{x}_j$ стабилизируется.

Поскольку в рассматриваемой задаче число авторов значительно превышает число журналов, то в качестве признаков для объектов-авторов используется наличие или отсутствие публикаций в конкретных журналах. В работе используется метрика суммы модулей разностей компонент векторов:

$$\rho(\mathbf{x}_i, \mathbf{x}_l) = \sum_{j \in \mathcal{J}} |x_{ij} - x_{lj}|,$$

В качестве исходных положений центров кластеров $\mathbf{u}_q, q \in \{1, \ldots, K\}$ используются объекты выборки (авторы) $\mathbf{x}_i$, публиковавшиеся в самых популярных журналах среди активно публикующихся авторов, и авторы, опубликовавшие статьи в изданиях, популярных среди тех, кто имеет статьи лишь в одном журнале. *Популярностью* $P(\mathcal{A}, j)$ журнала $j$ для группы авторов $\mathcal{A}$ будем считать число авторов из рассматриваемой группы, опубликовавших работы в данном журнале:

$$P(\mathcal{A}, j) = \sum_{k \in \mathcal{A}} [x_{kj} = 1], \quad \mathcal{A} \subseteq \mathcal{I},$$

где $x_{kj}$ — элемент входной матрицы $\mathbf{X}$. Зададим пороги для искомых журналов:

$$P(\mathcal{A}_{\text{act}}, j) \geq \alpha_{\text{act}}, \quad P(\mathcal{A}_{\text{once}}, j) \geq \alpha_{\text{once}}, \tag{185}$$

где $\mathcal{A}_{\text{act}}$ и $\mathcal{A}_{\text{once}}$ — группы авторов, активно публикующихся и напечатавших лишь одну статью, соответственно. Таким образом, в число начальных центров кластеров $\mathbf{u}_q, q \in \{1, \ldots, K\}$ попадут те строки $\mathbf{x}_i$ исходной матрицы $\mathbf{X}$, соответствующие авторам, имеющим единственную статью $i \in \mathcal{A}_{\text{once}}$) в одном из популярных журналов среди мало публикующихся авторов $x_{ij} = 1 \Rightarrow P(\mathcal{A}_{\text{once}}, j) \geq \alpha_{\text{once}}$, и авторам, имеющим статьи во всех популярных журналах $i \in \mathcal{A}_{\text{act}}$ среди активно публикующихся авторов $x_{ij} = 1$ для любого $j$, такого что $P(\mathcal{A}_{\text{act}}, j) \geq \alpha_{\text{act}}$:

$$\mathcal{A}_{\text{start}} = \{\mathbf{u}_q\}_{q=1}^{K} = \begin{cases} \mathbf{x}_i \colon (i \in \mathcal{A}_{\text{once}}) \bigcap \big(x_{ij} = 1 \Rightarrow P(\mathcal{A}_{\text{once}}, j) \geq \alpha_{\text{once}}\big), \quad \text{либо} \\ \mathbf{x}_i \colon (i \in \mathcal{A}_{\text{act}}) \bigcap \big(x_{ij} = 1 \text{ для любого } j, \text{ такого что } P(\mathcal{A}_{\text{act}}, j) \geq \alpha_{\text{act}}\big). \end{cases} \tag{186}$$

*Кластеризация журналов* также проводится алгоритмом k-Means. В качестве признаков для журналов используется их типичность для каждого полученного кластера авторов. *Типичность* $T(q, j)$ журнала $j$ для кластера авторов с индексом $q$ — это доля авторов из данного кластера, опубликовавших статьи в данном журнале:

$$T(q, j) = \frac{\sum_{i \in \mathcal{I}} [x_{ij} = 1][y_i = q]}{\sum_{i \in \mathcal{I}} [y_i = q]},$$

где $x_{ij}$ — элемент входной матрицы $\mathbf{X}$, $y_i$ — номер кластера, приписанный автору на предыдущем этапе алгоритма. Сформируем $(K \times n)$-матрицу $\mathbf{Y}$, в которой столбцы соответствуют журналам, а строки — кластерам авторов.

Исходные положения центров кластеров $\mathbf{v}_p, p \in \{1, \ldots, P\}$ определяются с помощью разделения журналов на три группы $\mathcal{B}_{\text{big}}, \mathcal{B}_{\text{av}}, \mathcal{B}_{\text{small}}$ по величине их суммарной типичности

$$T_j = \sum_{q=1}^{K} T(q, j) \tag{187}$$

по всем кластерам. Из каждой группы отбираются журналы, имеющие наибольшие значения тех признаков, которые соответствуют наиболее типичным для данной группы журналов кластерам. *Типичность* $T(q, \mathcal{B})$ кластера авторов $q$ для группы журналов $\mathcal{B}$ находится как сумма типичностей соответсвующих журналов для данного кластера авторов:

$$T(q, \mathcal{B}) = \sum_{j \in \mathcal{B}} T(q, j), \quad \mathcal{B} \subseteq \mathcal{J}.$$

Для каждой группы журналов зададим пороги для определения наиболее типичных кластеров авторов:

$$T(q, \mathcal{B}_{\text{big}}) \geq \beta_{\text{big}}, \quad T(q, \mathcal{B}_{\text{av}}) \geq \beta_{\text{av}}, \quad T(q, \mathcal{B}_{\text{small}}) \geq \beta_{\text{small}}. \tag{188}$$

В число начальных положений центров кластеров $\mathbf{v}_p, p \in \{1,\ldots,P\}$ попадут столбцы $\mathbf{y}_j$ матрицы $\mathbf{Y}$, соответствующие журналам $j \in \mathcal{B}_k$ из каждой из трех групп $\mathcal{B}_{\text{big}}, \mathcal{B}_{\text{av}}, \mathcal{B}_{\text{small}}$, для которых выполнено условие $f(\mathbf{q}_k) > \beta$:

$$\mathcal{B}_{\text{start}} = \{v_p\}_{p=1}^{P} = \{\mathbf{y}_j : (j \in \mathcal{B}_k) \bigcap (f(\mathbf{q}_k) > \beta)\}, \tag{189}$$

где $\mathbf{q}_j$ — множество кластеров авторов, удовлетворяющих условию $T_{q,\mathcal{B}_k} \geq \beta_k$, $k \in \{\text{big,av,small}\}$. Функция $f(\mathbf{q}_k)$ в выражении (189) определяется как

$$f(\mathbf{q}_k) > \beta \Leftrightarrow \begin{cases} \prod_{q=1}^{K} y_{qj} > 0, & k = \text{“big”} \text{ и } j \in \mathcal{B}_{\text{big}} \text{ или } k = av \text{ и } j \in \mathcal{B}_{\text{av}}; \\ \sum_{q=1}^{K} y_{qj} > \beta, & k = \text{“small”} \text{ и } j \in \mathcal{B}_{\text{small}}. \end{cases}$$

*Формирование ко-кластеров* $c \in \{1,\ldots,C\}$ проходит путем отнесения кластера авторов $q$ к наиболее типичному для него кластеру журналов $p$:

$$c = q \cup \arg\max_{p} T(q,p), \tag{190}$$

где $T(q,p)$ — типичность кластера авторов с меткой $q$ для кластера журналов с меткой $p$. Кластер журналов с меткой $\tilde{p}$, оставшийся без авторов, присоединяется к тому кластеру $\hat{p}$, к которому отнесся наиболее типичный для него кластер авторов с меткой $\tilde{q}$:

$$\tilde{q} = \arg\max_{q} T(q,\tilde{p});$$

$$\hat{p} = \arg\max_{p} T(\tilde{q},p), \quad \hat{p} = \hat{p} \cup \tilde{p}. \tag{191}$$

Затем проводится повторное формирование кластеров по формуле (190).

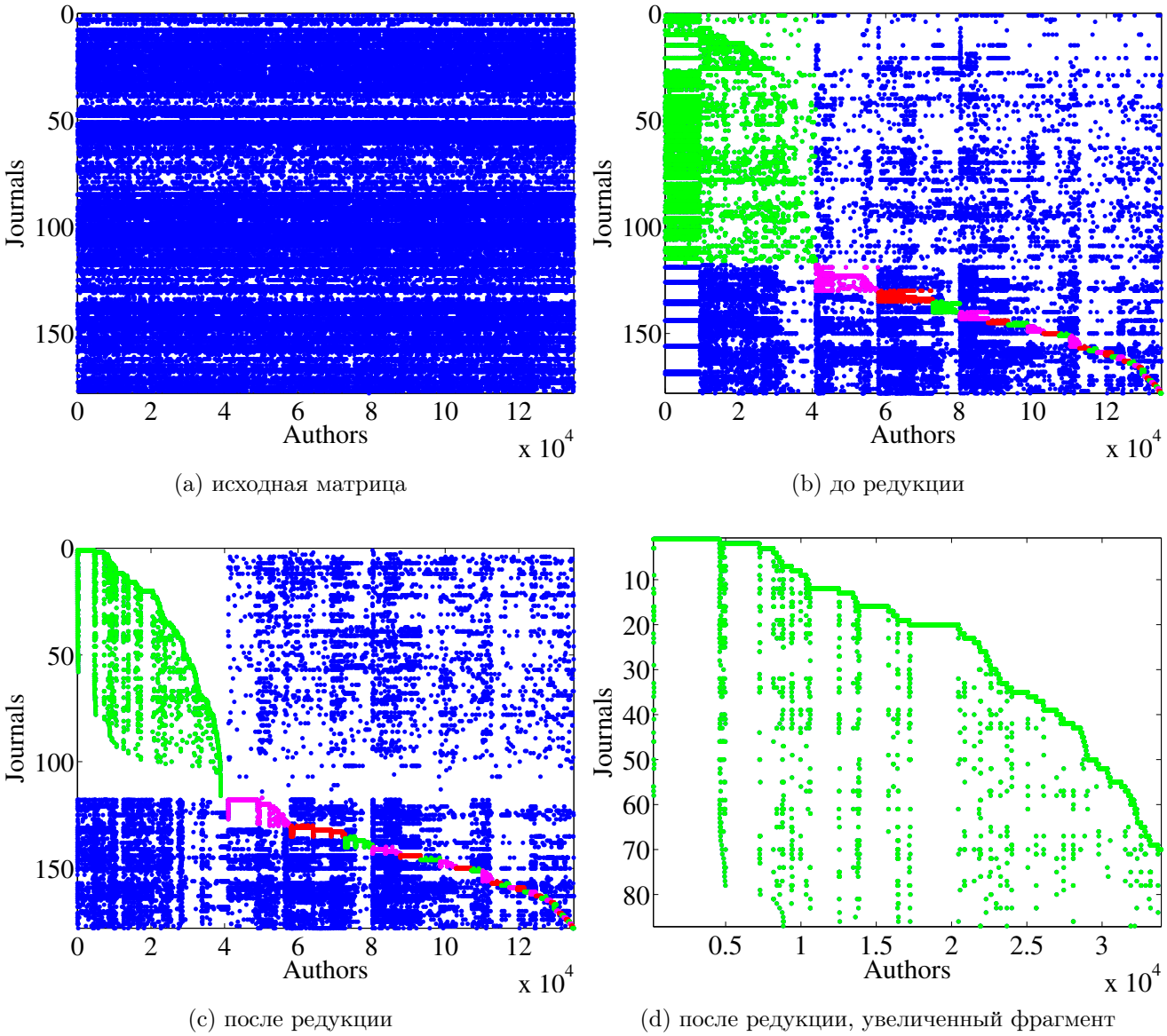
*Редукция ко-кластеров* проводится с помощью модифицированного алгоритма Cuthill-McKee. Введем матрицу $\mathbf{Z}_c$, которая является подматрицей входной матрицы $\mathbf{X}$, содержащей строки и столбцы (авторов и журналы), принадлежащие ко-кластеру с индексом $c$. Алгоритм работает с квадратной симметричной матрицей $\mathbf{W}$, интерпретируемой как матрица инцидентности соответствующего ей графа, которая строится в виде:

$$W = \begin{pmatrix} \mathbf{0} & \mathbf{Z}_c^{\mathsf{T}} \\ \mathbf{Z}_c & \mathbf{0} \end{pmatrix}, \quad c \in \{1,\ldots,C\},$$

где $\mathbf{0}$ — нулевая матрица необходимого размера. Алгоритм составляет перестановку $R$ вершин графа, обеспечивающую приведение матрицы к ленточной структуре. Последовательность шагов:

1) выбрать вершину $v$ графа, соответствующего матрице индидентности $\mathbf{W}$ , имеющую наибольшую степень (количество ребер, смежных с этой вершиной) и поместить ее в $R$: $R = \{v\}$,

2) для каждого элемента $u \in R$ найти все смежные вершины $a$, удалить из $a$ вершины, уже находящиеся в $R$, отсортировать $a$ по убыванию степени вершин, присоединить $a$ к $R$: $R = R \cup a$,




(a) исходная матрица

(b) до редукции

(c) после редукции

(d) после редукции, увеличенный фрагмент

Рис. 63. Полученные ко-кластеры в задаче построения интегрального индикатора.

3) если $A = \emptyset$, то следующая вершина выбирается из необработанных по наибольшему значению степени.

После проведения перестановки строк и столбцов в полученной матрице оставляют строки, соответствующие авторам и столбцы, соответствующие журналам.

*Редукция ко-кластеров с целью построения интегрального индикатора.* Модифицированным алгоритмом Cuthill-McKee проведена редукция каждого ко-кластера с целью сконцентрировать все ненулевые элементы матриц вблизи диагонали. Результат представлен на рис. 63. Представлена выборка, полученная из базы данных [7] и содержащая 134 966 авторов и 178 журналов. Для удобства изображения матрицы транспонированы. Синие точки соответсвуют ненулевым элементам исходной матрицы. Зелеными, фиолетовыми и красными точками обозначены ненулевые элементы подматриц, являющихся ко-кластерами. Ко-кластеры отсортированы в порядке убывания числа ненулевых элементов, что соответствует убыванию интегрированности авторов и журналов в мировую науку.

## 5.5. Порядковая классификация объектов по частично упорядоченным множествам

Решается задача ранжирования объектов [261, 182, 67] для определения статуса угрожаемых видов животных, входящих в список Красной книги IUCN. Приняты следующие предположения о составе и свойствах признаков:

1) состав признаков считается исчерпывающим для получения адекватной модели;

2) на значениях признаков задано отношение полного порядка;

3) выполняется правило «the bigger the better», то есть большему (благоприятному) значению признака соответствует больший (благоприятный) статус вида;

4) допускаются различные экспертные оценки одного и того же вида;

5) каждый из признаков принимает на выборке все допустимые значения и только их.

Признаки, использованные для описания объектов, принимают значения из множеств, на которых задано отношение порядка. Отношение частичного порядка является одним из видов бинарных отношений, свойства которых рассматриваются в [240]. Объекты, описанные в ранговых шкалах, не являются точками в некотором линейном пространстве, они представляют собой объекты нечисловой природы. Подходы к обработке нечисловой информации описаны, например, в [11]. В рассматриваемой прикладной задаче имеется экспертная информация о важности признаков относительно друг друга, то есть над множеством признаков тоже задано бинарное отношение предпочтения. Количество признаков сопоставимо с количеством объектов, доступных для обучения алгоритма. Задачи с избыточным числом признаков рассматриваются в работах [222, 217].

Задача монотонной классификации объектов нечисловой природы с учетом предпочтений освещается в [79, 100, 126, 147, 146, 282]. Она решается методом попарных сравнений. Задача монотонной классификации [182, 67] часто возникает в сфере информационного поиска. Для их решения используют ранговую регрессию [72], модифицированный алгоритм SVM [284] и модифицированный бустинг [97].

Прогнозирование состояния вида выполняется в два этапа: построение модели и классификация. Для построения модели используется алгоритм многоклассовой монотонной Парето-классификации. Обоснование принципа Парето представлено работе [216]. Предполагается, что для более устойчивой работы алгоритма целесообразно минимизировать количество объектов, входящих в Парето-фронт. Подходы к сужению множества Парето описаны в [216, 217]. В данной работе предлагается сузить множество Парето учитывая экспертное предпочтение важности признаков [228] при определении отношения доминирования объектов. Построенный алгоритм является альтернативой алгоритму решающего дерева, алгоритму обобщенной линейной регрессии и алгоритму на основе копул [170].

### 5.5.1. Матрица отношения порядка

Дано множество пар

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, \quad i \in \mathcal{I} = \{1, \ldots, m\},$$

состоящее из объектов $\mathbf{x}_i$ и меток классов $y_i$. Все объекты

$$\mathbf{x} = [\chi_1, \ldots, \chi_j, \ldots, \chi_d]^\top, \quad j \in \mathcal{J} = \{1, \ldots, d\},$$

описаны в порядковой шкале. Это означает, что каждый признак $\chi_j$ принимает значение из множества $\mathbb{L}_j = \{l_1, \ldots l_{k_j}\}$, на элементах которого задано отношение линейного порядка,

$$\chi_j \in \mathbb{L}_j = \{l_1, \ldots l_{k_j}\}, \text{ где } l_1 \prec \cdots \prec l_{k_j}.$$

Метки классов $y$ принимают значения из множества $\mathbb{Y} = \{l_1, \ldots, l_Y\}$, на элементах которого также задано отношение порядка $l_1 \prec \ldots \prec l_Y$.

Требуется построить монотонную функцию

$$\varphi \colon \mathbf{x} \mapsto \hat{y}, \tag{192}$$

определенную на всем множестве $\mathbb{X} = \mathbb{L}_1 \times \cdots \times \mathbb{L}_d$ и принимающую значения из множества $\mathbb{Y}$. Эта функция должна доставлять минимум функции ошибки

$$S(\varphi) = \sum_{i \in \mathcal{I}} r(y_i, \hat{y}_i), \tag{193}$$

где $\hat{y}_i = \varphi(\mathbf{x}_i)$; а функция

$$r(\cdot, \cdot) \tag{194}$$

задает расстояние между метками упорядоченного множества и будет определена ниже.

**Функция расстояния между элементами множества с линейным порядком.** Определим функцию расстояния (194) между элементами упорядоченного множества. Для этого запишем отношение порядка между элементами некоторого упорядоченного множества $\mathbb{Z} = \{l_1, \ldots, l_z\}$, $l_1 \prec \ldots \prec l_z$, с помощью бинарной матрицы 15. Если в матрице на пересечении строки $i$ и столбца $j$ стоит 1, то элементы множества $l_i$ и $l_j$ связаны отношением порядка $l_i \succ l_j$. Таким образом заданная матрица является нижнетреугольной с нулевой диагональю.

Таблица 15. Матрица отношения порядка.

| Метки | $l_1$ | $l_2$ | ... | $l_{z-1}$ | $l_z$ |
|---|---|---|---|---|---|
| $l_1$ | 0 | 0 | ... | 0 | 0 |
| $l_2$ | 1 | 0 | ... | 0 | 0 |
| ... | ... | ... | ... | ... | ... |
| $l_{z-1}$ | 1 | 1 | ... | 0 | 0 |
| $l_z$ | 1 | 1 | ... | 1 | 0 |

Поставим в соответствие элементу множества $l_i \in \mathbb{Z}$ $i$-тую бинарную строку $str_i$ из табл. 15. Тогда расстояние (194) между элементами $l_i$ и $l_j$ будет задаваться расстоянием Хэмминга между бинарными векторами

$$r(l_i, l_j) = R_{\text{Ham}}(str_i, str_j), \tag{195}$$

где $R_{\text{Ham}}(str_i, str_j)$ — количество несовпадающих разрядов в строках $str_i$ и $str_j$. Функция расстояния (195) будет использоваться для определения расстояния между метками классов из множества $\mathbb{Y}$ и метками значений признаков из множеств $\mathbb{L}_j$, $j = 1, \ldots, d$.

### 5.5.2. Парето-классификация для случая двух классов

Рассмотрим частный случай поставленной задачи, где $\mathbb{Y} = \{l_1, l_2\} = \{, 1\}$, $0 \prec 1$, то есть выборка $\mathfrak{D}$ содержит объекты только двух классов с метками 0 и 1. Искомую монотонную функцию, минимизирующую (193), обозначим

$$f\colon \mathbf{x} \mapsto \hat{y}. \tag{196}$$

Решим задачу нахождения функции $f(\mathbf{x})$ с помощью разделимой выборки $\hat{\mathfrak{D}} = \{(\mathbf{x}_i, y_i)\}$, $i \in \hat{\mathcal{I}} \subseteq \mathcal{I}$. Предполагается, что каждому из классов соответствует выпуклая оболочка POF, заданная отношением доминирования «$\succ$» объектов, и эти оболочки не пересекаются. Искомая функция $f$ будет сначала определена на множестве объектов разделимой выборки $\hat{\mathfrak{D}}$, то есть такой выборки, на которой функция $f$ не будет допускать ошибок, а затем доопределена на всем множестве $\mathbb{X}$.

**Отношение доминирования без учета важности признаков** Введем на объектах каждого из классов отношение доминирования. Разобьем множество индексов $\hat{\mathcal{I}}$ объектов разделимой выборки $\hat{\mathfrak{D}}$ на два подмножества $\hat{\mathcal{I}} = \mathcal{N} \bigsqcup \mathcal{P}$ так, что $y_n = 0$, $n \in \mathcal{N}$, а $y_p = 1$, $p \in \mathcal{P}$. Введем на множествах $\{\mathbf{x}_n : n \in \mathcal{N}\}$ и $\{\mathbf{x}_p : p \in \mathcal{P}\}$ отношения доминирования $\succ_n$ и $\succ_p$. Объект $\mathbf{x}_n = [x_{n1}, \ldots, x_{nd}]^\top$ $n$-доминирует объект $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]^\top$, если значения всех его признаков не менее предпочтительны, чем значения признаков $\mathbf{x}_i$:

$$\mathbf{x}_n \succ_n \mathbf{x}_i, \quad \text{если} \quad x_{nj} \succeq x_{ij} \quad \text{для всех} \quad j = 1, \ldots, d.$$

Объект $\mathbf{x}_p = [x_{p1}, \ldots, x_{pd}]^\intercal$ $p$-доминирует объект $\mathbf{x}_k = [x_{k1}, \ldots, x_{kd}]$, если значения всех его признаков не более предпочтительны, чем значения признаков $\mathbf{x}_k$:

$$\mathbf{x}_p \succ_p \mathbf{x}_k, \quad \text{если} \quad x_{pj} \preceq x_{kj} \quad \text{для всех} \quad j = 1, \ldots, d.$$

Будем считать, что объект не доминирует сам себя ни в одном из смыслов:

$$\mathbf{x} \not\succ_n \mathbf{x}, \quad \mathbf{x} \not\succ_p \mathbf{x}.$$

На рис. 64 приведен пример доминирования для случая двух признаков. По осям отложены значения признаков: для первого признака из множества $\mathbb{L}_1$, для второго — из множества $\mathbb{L}_2$, желтым цветом показаны область $n$-доминирования объекта $\mathbf{x}_n$ и область $p$-доминирования объекта $\mathbf{x}_p$. Объекты, попадающие в области, закрашенные желтым цветом, доминируются в соответствующем смысле рассмотренными объектами.

**Отношение доминирования с учетом важности признаков** Также введем на объектах каждого из классов отношение доминирования с учетом важности признаков $\succ_{\tilde{n}}$ и $\succ_{\tilde{p}}$. Пусть признак $\chi_r$ с индексом $r$ предпочтительнее (важнее), чем признак $\chi_r$ с индексом $t$:

$$r \succ t, \text{ где } r, t \in \mathcal{J}.$$

Объект $\mathbf{x}_n = [x_{n1}, \ldots, x_{nr}, \ldots, x_{nt}, \ldots, x_{nd}]^\intercal$ $\tilde{n}$-доминирует объект $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]^\intercal$ (будем обозначать $\mathbf{x}_n \succ_{\tilde{n}} \mathbf{x}_i$), если выполнено одно из двух условий

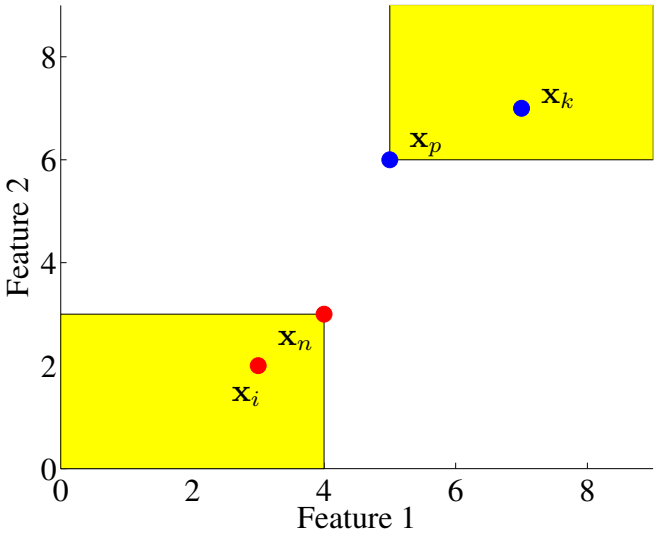



Рис. 64. Доминирование без учета важности признаков.

1) $\mathbf{x}_n$ $n$-доминирует $\mathbf{x}_i$ без учета важности признаков $\mathbf{x}_n \succ_n \mathbf{x}_i$, или

2) $x_{nr} \succ x_{nt}$ и $\mathbf{x}_n^{tr}$ доминирует $\mathbf{x}_i$ без учета важности признаков $\mathbf{x}_n^{tr} \succ_n \mathbf{x}_i$, где $\mathbf{x}_n^{tr} = [x_{n1}, \ldots, x_{nt}, \ldots, x_{nr}, \ldots, x_{nd}]^\intercal$, то есть соответствует объекту $\mathbf{x}_n$ с переставленными значениями признаков $r$ и $t$ и доминируется этим объектом.

Объект $\mathbf{x}_p = [x_{p1}, \ldots, x_{pr}, \ldots, x_{pt}, \ldots, x_{pd}]^\intercal$ $\tilde{p}$-доминирует объект $\mathbf{x}_k = [x_{k1}, \ldots, x_{kd}]^\intercal$ (будем обозначать $\mathbf{x}_p \succ_{\tilde{p}} \mathbf{x}_k$), если выполнено одно из двух условий

1) $\mathbf{x}_p$ $p$-доминирует $\mathbf{x}_k$ без учета важности признаков $\mathbf{x}_p \succ_p \mathbf{x}_k$, или

2) $x_{pr} \prec x_{pt}$ и $\mathbf{x}_p^{tr}$ доминирует $\mathbf{x}_k$ без учета важности признаков $\mathbf{x}_p^{tr} \succ_p \mathbf{x}_k$, где $\mathbf{x}_p^{tr} = [x_{p1}, \ldots, x_{pt}, \ldots, x_{pr}, \ldots, x_{pd}]^\intercal$, то есть соответствует объекту $\mathbf{x}_p$ с переставленными значениями признаков $r$ и $t$ и доминируется этим объектом.

Будем считать, что объект не доминирует сам себя ни в одном из смыслов:

$$\mathbf{x} \not\succ_{\tilde{n}} \mathbf{x}, \quad \mathbf{x} \not\succ_{\tilde{p}} \mathbf{x}.$$

На рис. 65 приведен пример доминирования для случая двух признаков, первый из которых важнее второго. По осям отложены значения признаков, для первого признака из множества $\mathbb{L}_1$, для второго — из множества $\mathbb{L}_2$. Обозначены объекты $\mathbf{x}_n$ и $\mathbf{x}_p$, зелеными точками отмечены $\mathbf{x}_n^{21}$ и $\mathbf{x}_p^{21}$. которые доминируются $\mathbf{x}_n$ и $\mathbf{x}_p$ и расширяют области их доминирования.

Возможные формы областей доминирования с учетом важности признаков приведены в табл. 16. Расширенные области доминирования в форме «ступеньки» соответствуют объектам, у которых значение более важного признака более предпочтительно в случае $\tilde{n}$-доминирования и менее предпочтительно в случае $\tilde{p}$-доминирования. Для всех остальных объектов форма области доминирования не отличается от случая без учета важности признаков и представляет собой прямоугольник.

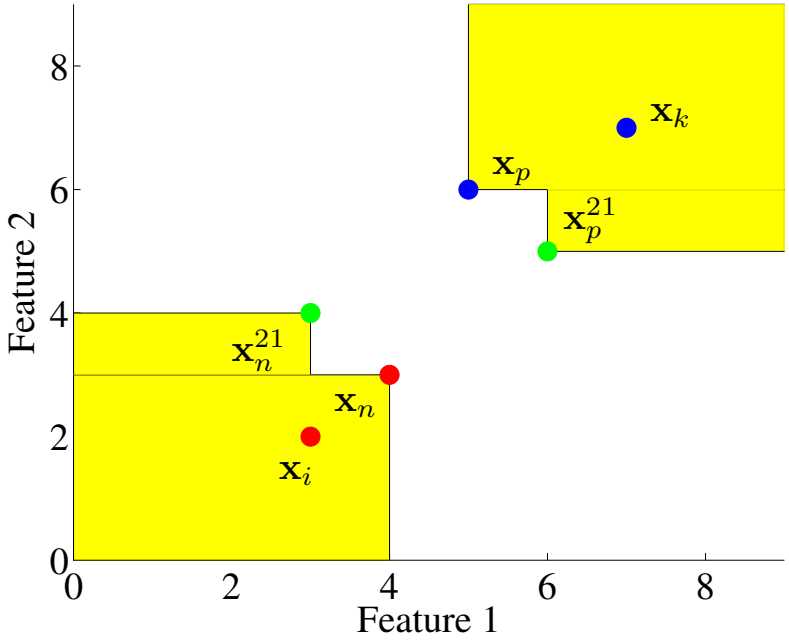
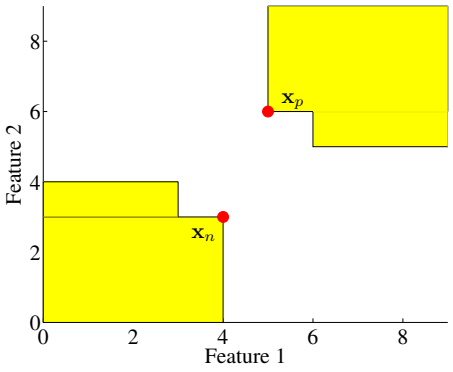
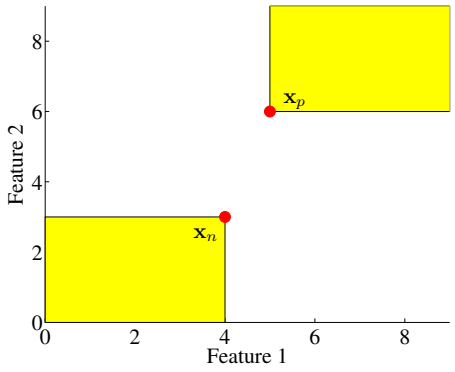
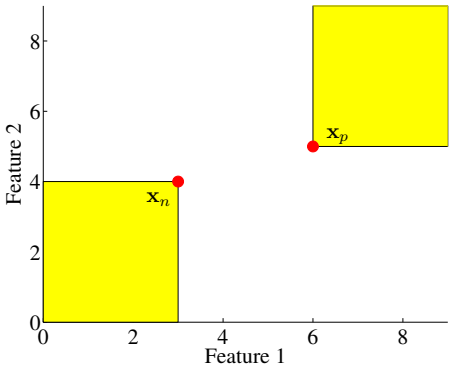
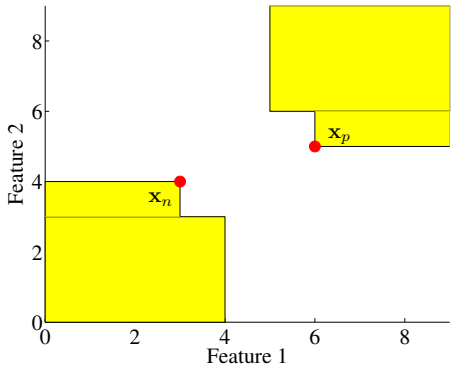



Рис. 65. Расширение областей доминирования при учете важности признаков.

Таблица 16. Формы областей доминирования при введении важности признаков.

| | Признак 1 важнее, чем признак 2 | Признак 2 важнее, чем признак 1 |
|---|---|---|
| $x_{n1} \succ x_{n2}$, $x_{p1} \prec x_{p2}$ |  |  |
| $x_{n1} \prec x_{n2}$, $x_{p1} \succ x_{p2}$ |  |  |

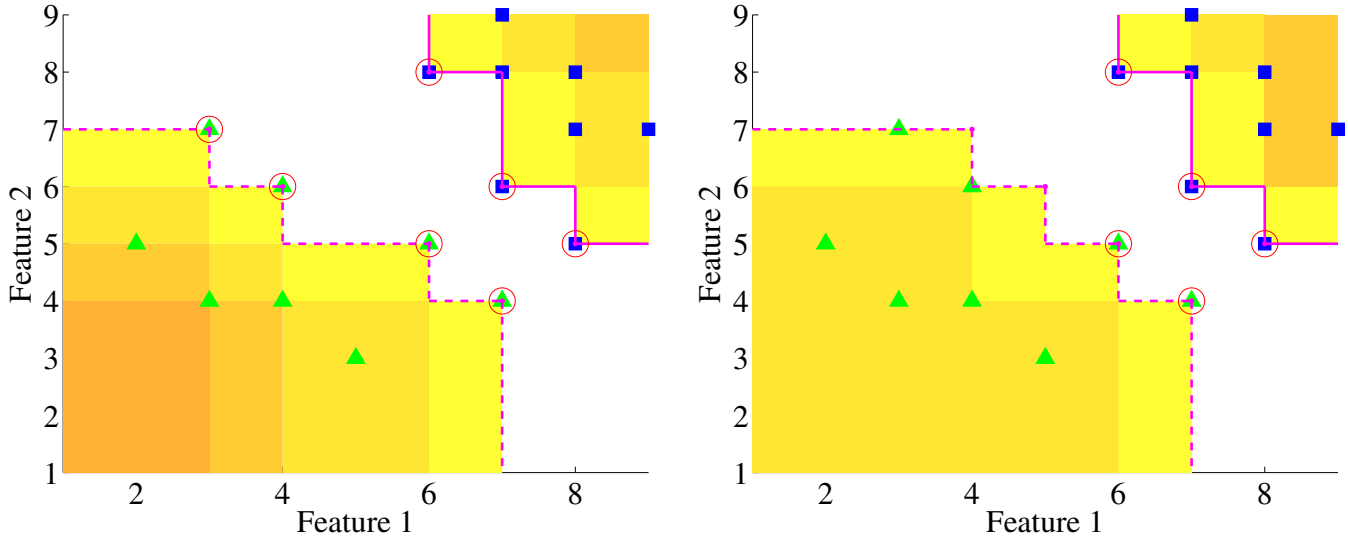


### 5.5.3. Построение набора Парето-оптимальных фронтов

Определим Парето-оптимальные фронты — множества, которые будут задавать границы классов разделимой выборки.

**Определение 17.** *Парето-оптимальный фронт $POF_n$ — множество объектов $\mathbf{x}_n, n \in \mathcal{N}$, для каждого элемента которого $\mathbf{x}_n \in POF_n$ не существует ни одного объекта $\mathbf{x}$, такого, что $\mathbf{x} \succ_n \mathbf{x}_n$ ($\mathbf{x} \succ_{\tilde{n}} \mathbf{x}_n$ для отношения доминирования с учетом важности признаков).*

**Определение 18.** *Парето-оптимальный фронт $POF_p$ — множество объектов $\mathbf{x}_p, p \in \mathcal{P}$, для каждого элемента которого $\mathbf{x}_p \in POF_p$ не существует объекта $\mathbf{x}$, такого, что $\mathbf{x} \succ_p \mathbf{x}_p$ ($\mathbf{x} \succ_{\tilde{p}} \mathbf{x}_p$ для отношения доминирования с учетом важности признаков).*




(a) Без учета важности признаков

(b) С учетом важности признаков

Рис. 66. Парето-оптимальные фронты.

На рис. 66 показаны примеры Парето-оптимальных фронтов для двухклассовой разделимой выборки, объекты которой описаны двумя признаками, принимающими значения из множеств $\mathbb{L}_1$ и $\mathbb{L}_2$ соответственно. Зелеными треугольниками и синими квадратами обозначены объекты разных классов. Объекты, вошедшие во фронты, обозначены красными кружками. Граница класса, задаваемая $n$-фронтом, обозначена пунктирной линией, $p$-фронтом — сплошной. Объединения областей доминирования объектов из фронтов закрашены желтым цветом, чем темнее оттенок области, тем большее количество объектов ее доминирует. На рис. 66(a) изображены Парето-оптимальные фронты, соответствующие отношению доминирования без учета важности признаков. На рис. 66(b) изображены Парето-оптимальные фронты, соответствующие отношению доминирования с учетом важности признаков (первый признак важнее второго). При введении в модель важности признаков количество объектов во фронте и форма объединения областей их доминирования может не измениться, как для $\mathrm{POF}_p$ в этом примере. Но в $\mathrm{POF}_n$ для рассматриваемого примера количество объектов уменьшилось, а объединение областей их доминирования, напротив, увеличилось, что объясняется расширением областей доминирования при учете экспертной информации о важности признаков.

Далее во всех рассуждениях и выкладках будут использоваться отношения доминирования с учетом важности признаков.

### 5.5.4. Классификация для случая двух классов

Построенные Парето-оптимальные фронты и границы классов, им соответствующие, будут использованы для определения монотонного классификатора (196).

Функция $f\colon \mathbf{x} \mapsto \hat{y}$ (197) ставит в соответствие произвольному объекту $\mathbf{x} \in \mathbb{X}$ метку класса «0», если найдется объект $\mathbf{x}_n \in \mathrm{POF}_n$, $\tilde{n}$-доминирующий $\mathbf{x}$, и метку класса «1», если найдется объект $\mathbf{x}_p \in \mathrm{POF}_p$, $\tilde{p}$-доминирующий $\mathbf{x}$.

$$f(\mathbf{x}) = \begin{cases} \text{«0»}, & \text{если найдется } \mathbf{x}_n \in \mathrm{POF}_n\colon \mathbf{x}_n \succ_{\tilde{n}} \mathbf{x}; \\ \text{«1»}, & \text{если найдется } \mathbf{x}_p \in \mathrm{POF}_p\colon \mathbf{x}_p \succ_{\tilde{p}} \mathbf{x}. \end{cases} \tag{197}$$

Если таких элементов не найдется, то функция $f$ доопределяется на множестве $\mathbb{X}$ согласно правилу ближайшего множества POF:

$$f(\mathbf{x}) = f\left(\underset{\mathbf{x}' \in \overline{\mathrm{POF}}_n \cup \overline{\mathrm{POF}}_p}{\arg\min} \left(\rho(\mathbf{x}, \mathbf{x}')\right)\right),$$

где множества $\overline{\mathrm{POF}}_n, \overline{\mathrm{POF}}_p$ включают Парето-оптимальные фронты и точки границы областей их доминирования и однозначно заданы построенными Парето-оптимальными фронтами. Функция $\rho$ задана с помощью функции (194), примененной к меткам значений признаков:

$$\rho(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{d} r(x_j, x'_j). \tag{198}$$

Таким образом, если не находится элементов из Парето-оптимальных фронтов, доминирующих объект $\mathbf{x}$, то $\mathbf{x}$ относится к тому классу, к Парето-оптимальному фронту которого он ближе.

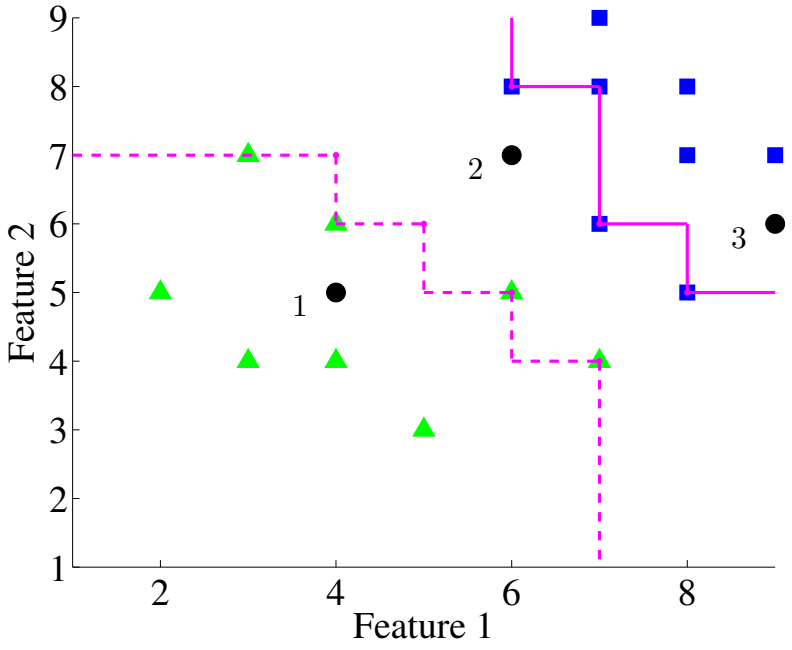



Рис. 67. Пример двухклассовой классификации методом Парето-оптимальных фронтов.

Таблица 17. Пример двухклассового классификатора.

| № | Объект $\mathbf{x}$ | $f(\mathbf{x})$ |
|---|---|---|
| 1 | (4,5) | 0 |
| 2 | (6,7) | 1 |
| 3 | (9,6) | 1 |

На рис. 67 изображена синтетическая выборка, содержащая объекты двух классов. Объекты первого класса обозначены зелеными треугольниками, второго — синими квадратами. Объекты описаны двумя признаками, значения которых отложены по осям и принадлежат множествам $\mathbb{L}_1$ и $\mathbb{L}_2$ соответственно. Классифицируемые объекты на графике отмечены черными кружками. Результат работы классификатора $f$ для этих объектов приведен в табл. 17, содержащей три столбца. В первом столбце номера классифицируемых объектов, во втором столбце координаты этих объектов, в третьем — результаты работы классификатора. Метка «0» во втором столбце означает, что объект отнесен к первому классу, который обозначен зелеными треугольниками, метка «1» — ко второму классу, изображенному синими квадратами.

### 5.5.5. Приведение выборки к разделимой

Рассмотрим процедуру нахождения множества $\hat{\mathcal{I}}$, на котором функция $f\colon \mathbf{x} \mapsto \hat{y}$ монотонна. Разобьем множество индексов $\mathcal{I}$ объектов выборки $\mathfrak{D}$ на два подмножества $\mathcal{I} = \mathcal{N} \bigsqcup \mathcal{P}$ так, что $y_n =$ «0», $n \in \mathcal{N}$, а $y_p =$ «1», $p \in \mathcal{P}$. Рассмотрим мощность $\mu$ доминируемого объектом $\mathbf{x}_i$ множества объектов другого класса:

$$\mu(\mathbf{x}_i) = \begin{cases} \#\{\mathbf{x}_j \mid \mathbf{x}_i \succ_n \mathbf{x}_j, j \in \mathcal{P}\}, & \text{если } i \in \mathcal{N}, \\ \#\{\mathbf{x}_j \mid \mathbf{x}_i \succ_p \mathbf{x}_j, j \in \mathcal{N}\}, & \text{если } i \in \mathcal{P}, \end{cases}$$

где знак # означает число элементов множества. Для нахождения множества $\hat{\mathcal{I}}$ проведем процедуру последовательного удаления объектов из выборки $\mathfrak{D}$:

1) $\hat{\mathcal{I}} = \mathcal{I}$, $\hat{\mathcal{P}} = \mathcal{P}$, $\hat{\mathcal{N}} = \mathcal{N}$,

2) пока в выборке $\mathfrak{D}$ есть объекты $\mathbf{x}_i$ с индексом $i \in \hat{\mathcal{I}}$ такие, что $\mu(\mathbf{x}_i) > 0$ повторять пункты 3–6,

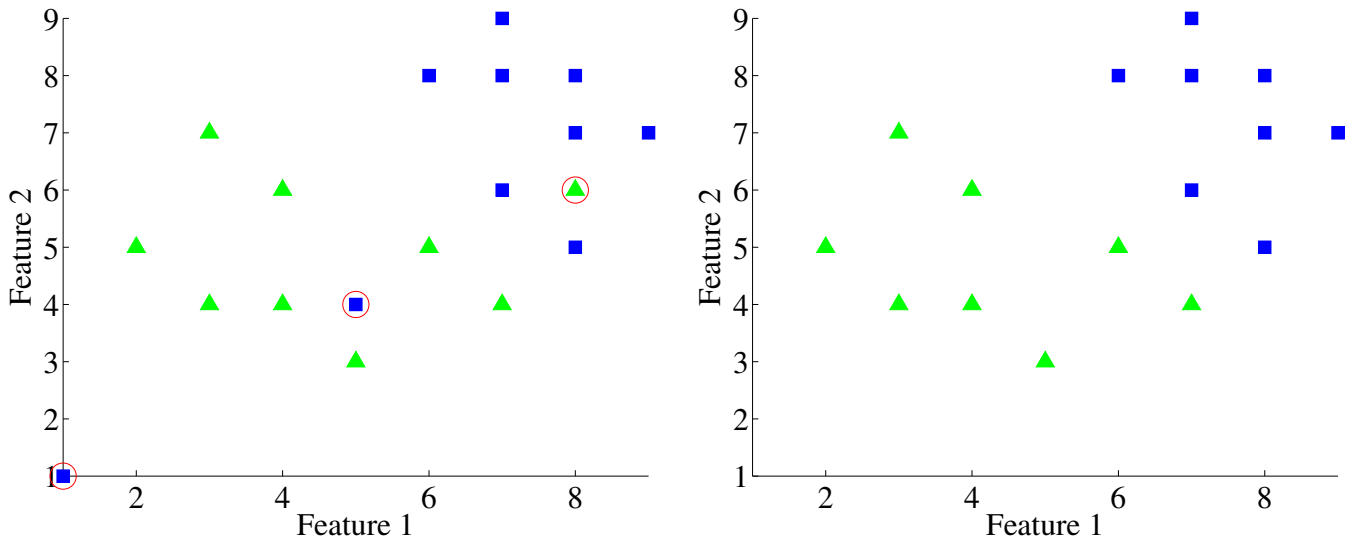
3) $\hat{i} = \underset{i \in \hat{\mathcal{I}} = \hat{\mathcal{N}} \bigsqcup \hat{\mathcal{P}}}{\arg\max} \mu(\mathbf{x}_i)$,

4) $\hat{\mathcal{I}} = \hat{\mathcal{I}} \backslash \{\hat{i}\}$,

5) Если $\hat{i} \in \hat{\mathcal{P}}$, то $\hat{\mathcal{P}} = \hat{\mathcal{P}} \backslash \{\hat{i}\}$,

6) Если $\hat{i} \in \hat{\mathcal{N}}$, то $\hat{\mathcal{N}} = \hat{\mathcal{N}} \backslash \{\hat{i}\}$.

На рис. 68 изображена синтетическая выборка, объекты которой описываются двумя признаками. Выборка включает два класса, которые обозначены разными маркерами (зелеными




(a) С дефектными объектами (b) Без дефектных объектов

Рис. 68. Иллюстрация исключения дефектных объектов из выборки.

треугольниками и синими квадратами). На рис. 68(a) изображена неразделимая выборка с дефектными объектами (1;1), (5;4) и (8;6), доминирующими объекты чужого класса. Эти объекты выделены красными окружностями. На рис. 68(b) показана разделимая выборка, полученная после применения описанной выше процедуры.

### 5.5.6. Монотонная классификация

**Построение монотонного классификатора.** Рассмотрим общий случай задачи,

$$\mathbb{Y} = \{l_1, \ldots, l_u, l_{u+1}, \ldots, l_Y\}, \quad l_1 \prec \ldots \prec l_u \prec l_{u+1} \prec \ldots \prec l_Y.$$

Обозначим $\{1, \ldots, u, u+1, \ldots, Y\}$ индексы меток классов. Для каждой смежной пары классов $u, u+1$ построим монотонный двухклассовый классификатор

$$f_{u,u+1}\colon \mathbf{x} \mapsto \hat{y} \in \{\text{«0»}, \text{«1»}\},$$

$\mathbf{x} \in \mathbb{X}$. Для построения каждого из классификаторов будем делить выборку на два класса с метками «0» и «1», относя к классу «0» все объекты из классов с метками, не более предпочтительными, чем $l_u$, и к классу «1» — с метками, более предпочтительными, чем $l_u$. При этом множество индексов объектов $\hat{\mathcal{I}}$ разделимой выборки $\hat{\mathfrak{D}}$ разбивается на два непересекающихся подмножества

$$\hat{\mathcal{I}} = \mathcal{N}_u \bigsqcup \mathcal{P}_{u+1}, \text{ где } n \in \mathcal{N}_u, \text{ если } y_n \preceq l_u, \text{ и } p \in \mathcal{P}_{u+1}, \text{ если } y_p \succeq l_{u+1}.$$

Монотонный классификатор

$$\varphi(\mathbf{x}) = \varphi(f_{1,2}, \ldots, f_{Y-1,Y})(\mathbf{x}), \quad \varphi\colon \mathbb{X} \to \mathbb{Y},$$

задан следующим образом:

$$\varphi(\mathbf{x}) = \begin{cases} \min\limits_{l_u \in \mathbb{Y}} \{l_u \mid f_{u,u+1}(\mathbf{x}) = \text{«0»}\}, & \text{если } \{l_u \mid f_{u,u+1}(\mathbf{x}) = \text{«0»}\} \neq \emptyset; \\ l_Y, & \text{если } \{l_u \mid f_{u,u+1}(\mathbf{x}) = \text{«0»}\} = \emptyset. \end{cases} \tag{199}$$

Таблица 18. Иллюстрация монотонного классификатора.

| $1,\ 2$ | ... | $u-1,\ u$ | $u,\ u+1$ | ... | $Y-1,\ Y$ |
|---|---|---|---|---|---|
| «1» | ... | «1» | «0» | ... | «0» |

В табл. 18 показана иллюстрация формулы 199. Результатом монотонной классификации является метка класса, на котором двухклассовый классификатор впервые дает ответ «0», если все ответы двухклассовых классификаторов равны «1», то результатом многоклассовой классификации будет метка последнего класса $l_y$.

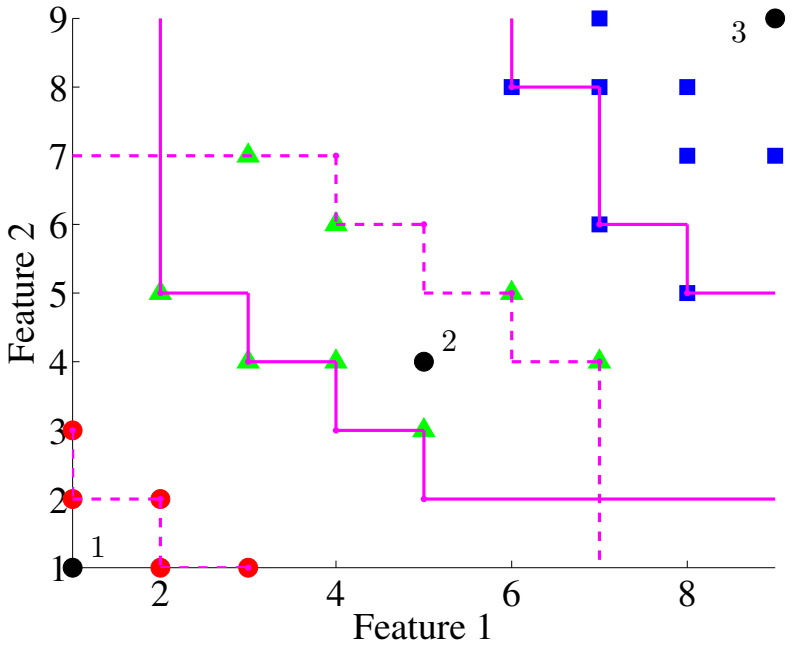
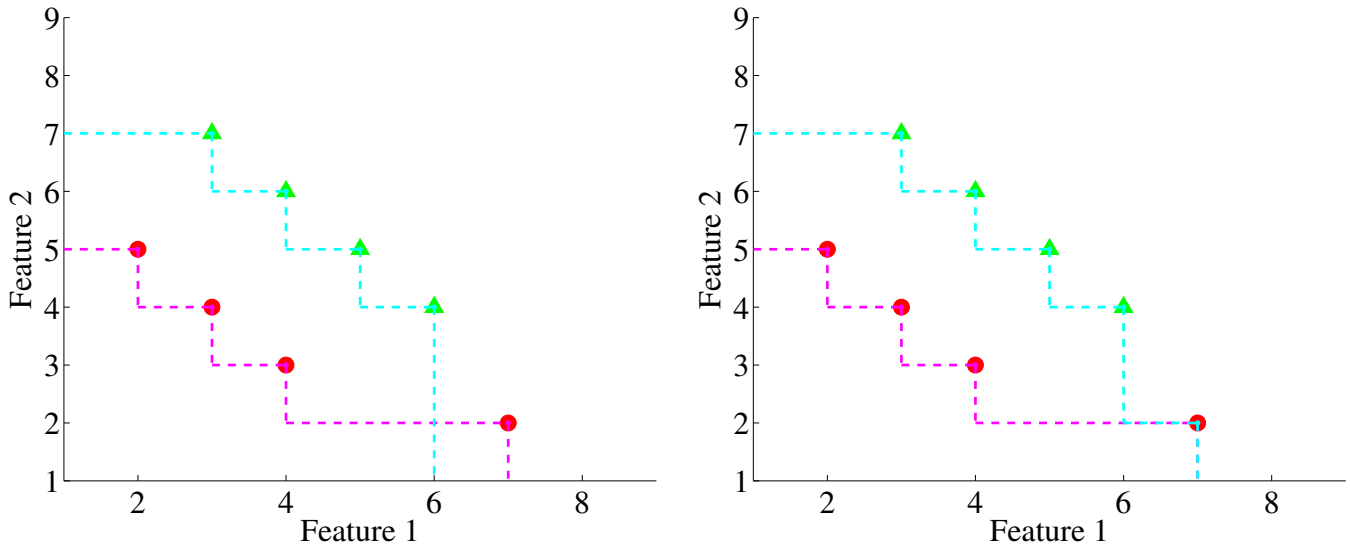



Рис. 69. Парето-фронты, первый признак важнее второго.

На рис. 69 изображена синтетическая выборка, содержащая объекты трех классов. Значения двух признаков, которыми описываются объекты, отложены по осям графика, объекты, принадлежащие разным классам, отмечены красными кружками, зелеными треугольниками и синими квадратами. Показаны построенные с учетом важности признаков фронты. Границы классов, соответствующие $n$-фронтам обозначены пунктирной линией, $p$-фронтам — сплошной. Классифицируемые объекты отмечены черными кружками. Пример результата работы набора функций $f_{1,2}$, $f_{2,3}$, входящих в классификатор $\varphi$, для выборки на рис. 69, выглядит как табл. 19. Первый столбец таблицы содержит номера классифицируемых объектов, второй — их координаты, третий и четвертый столбцы — результаты работы двухклассовых классификаторов для смежных первого и второго, второго и третьего классов соответственно на представленных объектах. Метка «0» во втором столбце означает, что объект был отнесен к первому классу классификатором $f_{1,2}$, метка «1» — ко второму классу этим же классификатором. Метка «0» в третьем столбце означает, что объект был отнесен ко второму классу классификатором $f_{2,3}$, метка «1» — к третьему классу классификатором $f_{2,3}$. Последний столбец содержит результаты монотонной классификации объектов. Значения в этом столбце соответствуют номеру класса, к которому в итоге был отнесен объект.

Таблица 19. Пример монотонного классификатора.

| № | Объект $\mathbf{x}$ | $f_{12}(\mathbf{x})$ | $f_{23}(\mathbf{x})$ | $\varphi(\mathbf{x})$ |
|---|---|---|---|---|
| 1 | (1,1) | «0» | «0» | «1» |
| 2 | (5,4) | «1» | «0» | «2» |
| 3 | (9,9) | «1» | «1» | «3» |




(a) Без доопределения (b) С доопределением

Рис. 70. Общий объект для двух $n$-фронтов.

**Доопределение Парето-оптимальных фронтов при монотонной классификации.** При построении фронтов между классами с метками $l_u$ и $l_{u+1}$ используются объекты классов с метками $l_1, \ldots, l_u$ для построения $n$-фронта для класса $l_u$ и объекты классов с метками $l_{u+1}, \ldots, l_Y$ для построения $p$-фронта для класса $l_{u+1}$. Поэтому одни и те же объекты могут попадать во фронты для разных классов, доопределяя их, и фронт одного класса может содержать объекты нескольких классов. На рис. 70 приведен фрагмент синтетической выборки, содержащей объекты трех классов. На графике изображены только объекты первого (красные кружки) и второго (зеленые треугольники) классов, иллюстрирующие ситуацию, когда объект с координатами (7;2) из первого класса попадает в $n$-фронты первого и второго классов.

Таким образом, $n$-фронт доопределяется объектами, принадлежащими классам с метками, не превосходящими метку класса, для которого этот фронт строится; $p$-фронт доопределяется аналогичным образом объектами классов с метками, превосходящими метку класса,для которого строится фронт.

**Допустимые классификаторы.**

**Определение 19.** *Классификатор $\varphi$ (199) будем называть* допустимым, *если для всех входящих в него функций $f_{u,u+1}$ соблюдается условие транзитивности:*

$$\begin{cases} \text{если} \quad f_{u,u+1}(\mathbf{x}) = \text{«0»}, \quad \text{то} \quad f_{(u+s)(u+1+s)}(\mathbf{x}) = \text{«0»} \quad \text{для всех} \quad s\colon (u+1+s) \leqslant Y, \\ \text{если} \quad f_{u,u+1}(\mathbf{x}) = \text{«1»}, \quad \text{то} \quad f_{(u-s)(u+1-s)}(\mathbf{x}) = \text{«1»} \quad \text{для всех} \quad s\colon (u-s) \geqslant 1. \end{cases} \tag{200}$$

**Определение 20.** *Парето-оптимальные фронты $POF_n(u)$ и $POF_p(u+1)$, $u = 1, \ldots, Y-1$ называются непересекающимися $POF_n(u) \cap POF_p(u+1) = \emptyset$, если границы их областей доминирования $\overline{POF_n(u)}$ и $\overline{POF_p(u+1)}$ не имеют общих точек.*

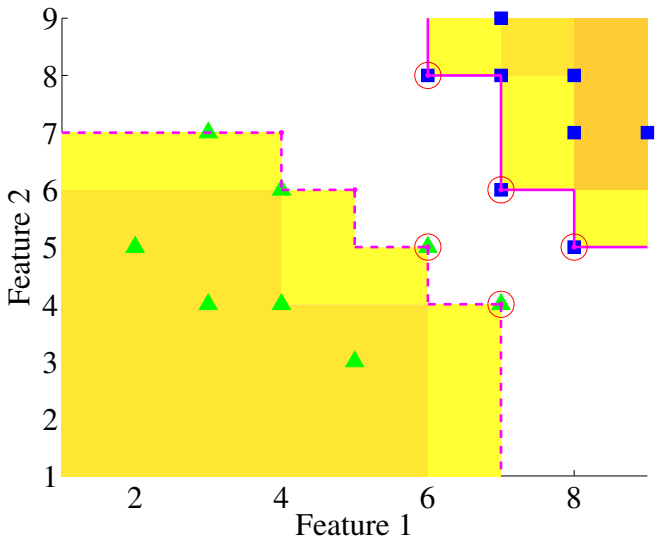



Рис. 71. Пример непересекающихся Парето-оптимальных фронтов.

На рис. 71 изображена синтетическая выборка объектов двух классов (зеленые треугольники и синие квадраты), описанных двумя признаками. Построены непересекающиеся фронты, с учетом важности признаков (первый признак важнее второго).

**Теорема 17.** *Непересечения Парето-оптимальных фронтов $POF_n(u) \cap POF_p(u+1) = \emptyset$, $u = 1, \ldots, Y-1$ достаточно для выполнения отношения транзитивности (200) для любого классифицируемого объекта.*

Поскольку в работе для построения Парето-оптимальных фронтов используются разделимые выборки, построенные фронты не пересекаются. Поэтому построенный монотонный классификатор (199) допустим и для любого классифицируемого объекта выполняется условие транзитивности (200).

# 6. Анализ прикладных задач

Поставленные в первой главе задачи регрессионного анализа и их решения играют важную роль в ряде прикладных областей. Принятый способ постановки задач, в терминах $\arg\max$, позволяет разбить работы по решению прикладных задач на несколько независимых частей. Постановка прикладной задачи как задачи регрессионного анализа включает следующие шаги.

1. Строится регрессионная выборка, определяются общие цели моделирования.
2. Назначается функция ошибки и ограничения на регрессионную модель. Функция ошибки может быть назначена исходя из гипотезы порождения данных, либо исходя из прикладных соображений, например, из требований к минимизации риска, максимизации прибыли, из стандартов физико-химических измерений и прочих.
3. Назначается класс регрессионных моделей, из которых будет выбрана модель оптимальной структурной или статистической сложности.
4. Задача выбора модели ставится как оптимизационная задача с ограничениями. Выбираются алгоритмы оптимизации для ее решения.
5. Исходя из гипотезы порождения данных или исходя из прикладных соображений выполняется ряд тестов, которые оценивают качество и свойства выбранной модели.

В этом разделе приводится анализ постановки регрессионных задач с прикладной точки зрения. Предложено несколько новых постановок авторегрессионных задач, включающих выбор моделей и построение их смесей. При этом учитывается технология планирования прикладных проектов, включающая текстовое описание проекта, предназначенное для фиксации целей, методов и результатов проекта. План проекта включает:

1) цель проекта, основная цель исследований, ожидаемые результаты,

2) обоснование проекта и область применения результатов проекта,

3) описание данных, включающее форматы и структуры данных,

4) описание критериев качества моделирования данных или целевых функций,

5) требования к проекту и условия успешного завершения проекта,

6) возможные риски и сложности, связанные с выполнением проекта,

7) краткое перечисление методов, предлагаемых для решения задачи.

## 6.1. Анализ постановок прикладных задач с использованием порождающих методов

Задано *множество прецедентов* — наборов результатов измерений

$$\mathfrak{S} = \{\mathfrak{s}_1, ..., \mathfrak{s}_m\}.$$

Элементом $\mathfrak{s}_i$ множества $\mathfrak{S}$ может являться, например, временной ряд некоторой длины, видеоряд или анкета скорингового клиента. Задано множество меток классов, или переменных отклика $\mathbf{y} = \{y_1, ..., y_m\}$

Помимо множества $\mathfrak{S}$, задан *словарь* — множество $V = V(\mathfrak{S})$, представляющее собой набор знаний о множестве прецедентов, необходимый для порождения моделей. Словарь может быть получен в результате анализа структуры прецедентов.

Задано экспертное множество порождающих функций $G = \{g_1, ..., g_n\}$, где каждая функция $g_j$ отображает объект анализа $\mathfrak{s}_i$ в элемент $(i, j)$ матрицы плана $\mathbf{X}$:

$$g_j : \quad (\mathbf{b}_j, \mathfrak{s}_i, V) \mapsto x_{ij} \in \mathbb{R}^1,$$

где $\mathbf{b}_j$ — набор параметров порождающей функции $g_j$.

Задана модель $\mathbf{f}$ и функция ошибки $S(\mathbf{w}|\mathbf{f}, \mathbf{X}, \mathbf{y})$. Требуется решить задачу поиска оптимальных параметров $\hat{\mathbf{w}}$ и оптимального поднабора признаков $\mathcal{A}$:

$$(\hat{\mathbf{w}}, \hat{\mathcal{A}}) = \underset{\substack{\mathbf{w} \in \mathbb{R}^n, \\ \mathcal{A} \subseteq \mathcal{J} = \{1, ..., n\}}}{\arg\min} S(\mathbf{w}_{\mathcal{A}}|\mathbf{f}, \mathbf{X}_{\mathcal{A}}, \mathbf{y}).$$

### 6.1.1. Прогнозирование квазипериодических временных рядов

Рассмотрим постановку задачи авторегрессионного прогнозирования временного ряда, как одну из наиболее показательных при создании многоуровневой прогностической модели, в которой требуется одновременно выбрать объекты и признаки для каждой модели.

В качестве примера приведем прогноз цен и объемов потребления электроэнергии с использованием набора временных рядов. Даны исторические ряды цен, регистрируемых каждый час и объемов потребления электроэнергии. Дополнительные временные ряды продолжительность светового дня, температура воздуха, влажность, сила ветра, производственный календарь. Требуется спрогнозировать потребление по часам на следующий день. На рис. 72 синей ломаной показан квазипериодический временной ряд, значения которого нужно спрогнозировать. Временной ряд имеет периоды: год, неделя, сутки, а также содержит апериодические сегменты, например, праздничные дни. На рис. 73 показана годовая периодичность временного ряда. Почасовые значения ряда, сгруппированные по дням, находятся строках матрицы. Одна строка матрицы описывает один день недели. Всего показано 170 строк, что примерно соответствует трем календарным годам.

Рассмотрим формальную постановку задачи. Дан временной ряд

$$\mathbf{s} = \{s_1, \ldots, s_\tau, \ldots, s_{T-1}\}, \qquad s \in \mathbb{R}.$$

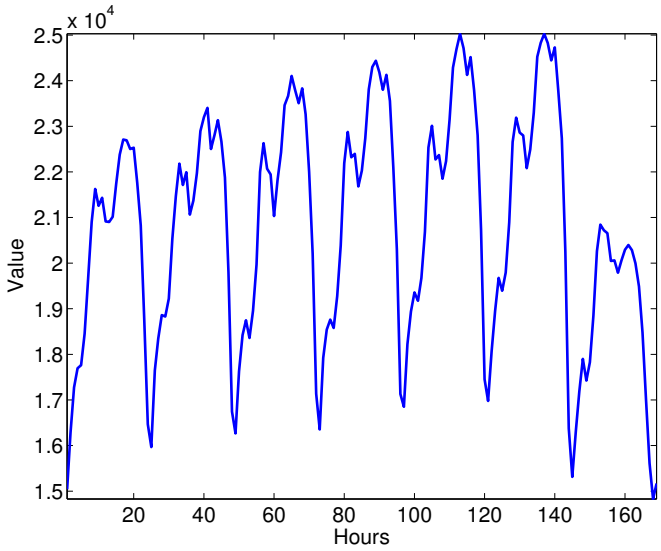



Рис. 72. Исходный временной ряд цен на электроэнергию, почасовые значения.

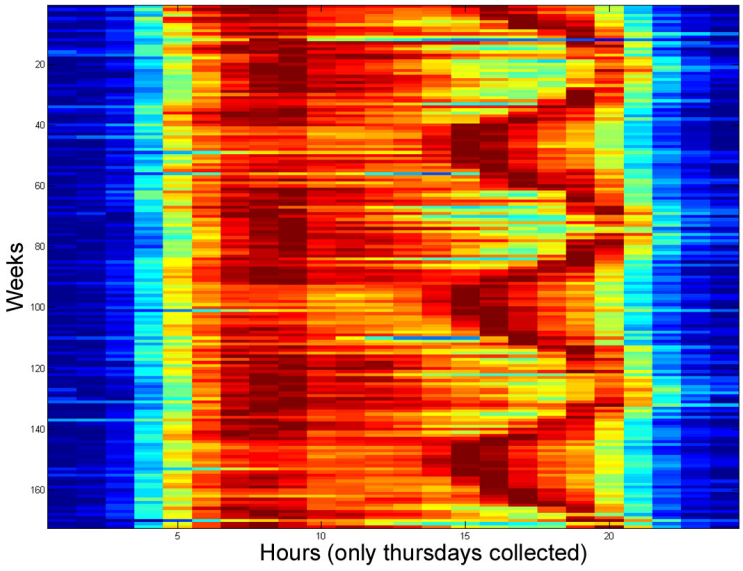



Рис. 73. Годичные периоды временного ряда, почасовые значения, сгруппированные по неделям.

Предполагается, что

(201)

1) отсчеты времени сделаны через равные промежутки; следовательно, значения временного ряда можно без ограничения общности проиндексировать натуральными числами, которые будем считать тождественными отсчетам времени $\tau$, сам ряд $\mathbf{s}$ будем считать вектором,

2) ряд имеет единственную периодическую составляющую и период $\kappa$ известен,

3) ряд не имеет пропущенных значений,

4) длина ряда кратна периоду, в противном случае из начала ряда следует удалить необходимое число элементов.

Требуется

1) спрогнозировать следующее значение $s_T$ временного ряда в момент времени $T$,

2) спрогнозировать значения $s_T, \ldots, s_{T+\kappa-1}$ временного ряда на следующем периоде $T, \ldots, T+\kappa-1$.

**Прогноз следующего значения временного ряда.** Предлагается спрогнозировать следующее значение временного ряда с помощью линейной регрессии. Для этого построим авторегрессионную матрицу $\mathbf{X}^*$ следующим образом. Элемент матрицы в строке с номером $i$ и столбце с номером $j$ тождественно равен элементу временного ряда с индексом

$$\tau = (i-1)\kappa + j \quad \text{при} \quad \mathrm{mod}\frac{T}{\kappa} = 0.$$

Эта матрица, в которой $m$ строк и $\kappa$ столбцов при длине ряда $T = m\kappa$, имеет вид

$$\mathbf{X}^* = \left[\begin{array}{ccc|c} s_1 & \ldots & s_{\kappa-1} & s_\kappa \\ \ldots & \ldots & \ldots & \ldots \\ s_{(m-2)\kappa+1} & \ldots & s_{(m-1)\kappa-1} & s_{(m-1)\kappa} \\ \hline s_{T-\kappa+1} & \ldots & s_{T-1} & s_T \end{array}\right],$$

Строка с номером $i$ содержит один период, а столбец с номером $j$ — некоторую фазу периода. Другими словами, столбец матрицы содержит элементы ряда $\mathbf{s}$ с индексами, разность которых кратна периоду $\kappa$. На рис. 74 показан пример авторегрессионной матрицы. Красный цвет соответствует бóльшим значениям временного ряда, синий — меньшим.

Представим $\mathbf{X}^*$ в виде матрицы, состоящей из соединенных векторов

$$\mathbf{X}^* = \left[\begin{array}{ccc|c} s_1 & \ldots & s_{\kappa-1} & s_\kappa \\ \ldots & \ldots & \ldots & \ldots \\ s_{(m-2)\kappa+1} & \ldots & s_{(m-1)\kappa-1} & s_{(m-1)\kappa} \\ \hline s_{T-\kappa+1} & \ldots & s_{T-1} & s_T \end{array}\right], \tag{202}$$

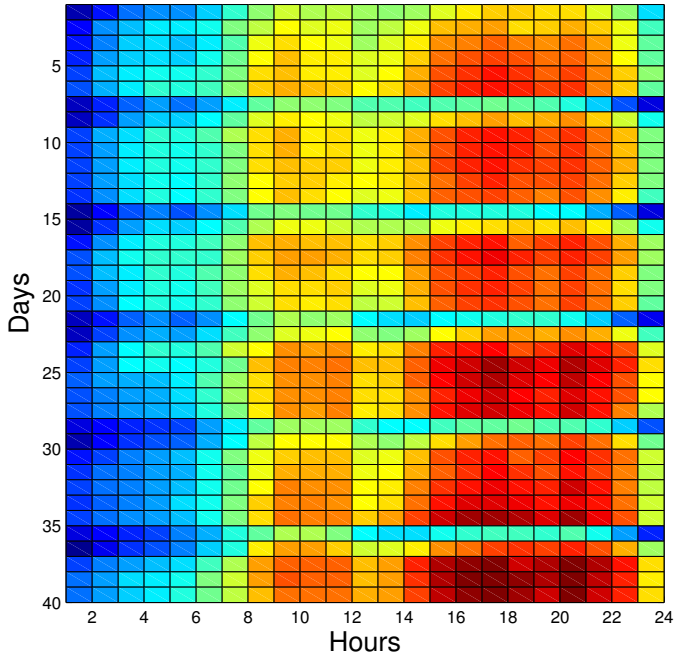



Рис. 74. Авторегрессионная матрица для временного ряда цен на электроэнергию.

или кратко,

$$\mathbf{X}^* = \left[\begin{array}{c|c} \underset{(m-1)\times n}{\mathbf{X}} & \underset{(m-1)\times 1}{\mathbf{y}} \\ \hline \underset{1\times n}{\mathbf{x}_m} & \underset{1\times 1}{s_T} \end{array}\right].$$

Здесь $\mathbf{X}$ — матрица плана с числом столбцов $n = \kappa - 1$, а $\mathbf{y}$ — последний столбец матрицы $\mathbf{X}^*$. Принимая линейную модель зависимости $\mathbf{y} = \mathbf{X}\mathbf{w}$, после оценки наиболее вероятного вектора параметров $\hat{\mathbf{w}}$ получаем прогнозируемое значение

$$s_T = \mathbf{x}_m^{\mathsf{T}}\hat{\mathbf{w}}.$$

Для случая нескольких рядов выполняется та же операция построения регрессионной матрицы. Пусть даны $\ell$ рядов $\mathbf{s}_1, \ldots \mathbf{s}_\ell$. Для каждого ряда строится матрица, и получается набор матриц $\mathbf{X}_1^*, \ldots \mathbf{X}_\ell^*$. Матрицы соединяются, и от полученной матрицы отделяются столбец значений зависимой переменной $\mathbf{y}$ и последняя строка $\mathbf{x}_m$.

При использовании набора порождающих функций $G = \{g_1, \ldots, g_r\}$, например, $g_1 = \sqrt{x}$, $g_2 = \operatorname{arcsinh}(x)$, $g_3 = x\sqrt{x}$, $g_4 = \mathrm{id}(x)$, матрица $\mathbf{X}^*$ будет иметь вид

$$\mathbf{X}^* = \left(\begin{array}{cccccc|c} g_r \circ s_1 & \ldots & g_1 \circ s_1 & \ldots & g_r \circ s_\kappa & \ldots & g_1 \circ s_\kappa \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ g_r \circ s_{(m-1)\kappa+1} & \ldots & g_1 \circ s_{(m-1)\kappa+1} & \ldots & g_r \circ s_{(m-1)\kappa} & \ldots & g_1 \circ s_{(m-1)\kappa} \\ \hline g_r \circ s_{T-\kappa+1} & \ldots & g_1 \circ s_{T-\kappa+1} & \ldots & g_r \circ s_T & \ldots & g_1 \circ s_T \end{array}\right).$$

Два вышеприведенных варианта построения матрицы $\mathbf{X}^*$ на практике приводят к тому, что число столбцов матрицы может значительно превышать число ее строк. Пусть, например, имеются значения потребления электроэнергии за три года по часам. Тогда матрица $\mathbf{X}^*$ имеет размер

$$156 \times 168, \text{ то есть } \quad 52 \text{ недели} \cdot 3 \text{ года} \times 24 \text{ часа} \cdot 7 \text{ дней};$$

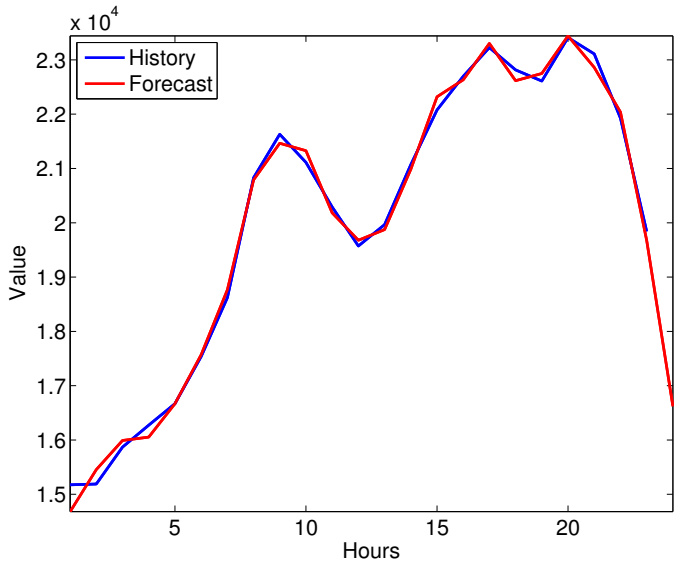



Рис. 75. Поточечный прогноз временного ряда на сутки вперед с использованием предыстории на каждом шаге.

Для случая шести вышеперечисленных временных рядов матрица $\mathbf{X}^*$ имеет размер $156 \times 1008$. При использовании четырех упомянутых порождающих функций матрица $\mathbf{X}$ имеет размер $156 \times 4032$. Таким образом, матрицу плана $\mathbf{X}$ можно считать плохо обусловленной, а ее столбцы — мультикоррелирующими. Для прогноза требуется решить задачу выбора столбцов матрицы плана $\mathbf{X}$. Таким образом, даны

1) матрица плана $\mathbf{X} = [\mathbf{x}_1^\mathsf{T}, \ldots, \mathbf{x}_i^\mathsf{T}, \ldots, \mathbf{x}_{m-1}^\mathsf{T}]$, иначе $\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_n]^\mathsf{T}$, $\mathcal{J} = \{1, \ldots, n\}$,

2) вектор значений зависимой переменной $\mathbf{y}$, который вместе с матрицей плана является регрессионной выборкой $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$,

3) класс моделей $\{\mathbf{f}_\mathcal{A} = \mathbf{X}_\mathcal{A}\mathbf{w}_\mathcal{A} | \mathcal{A} \subseteq \mathcal{J}\}$,

4) гипотеза порождения данных $\mathbf{y} \sim \mathcal{N}(\mathbf{f}, \mathbf{B})$ и функция ошибки $S(\mathbf{w}_\mathcal{A}|\mathbf{f}_\mathcal{A}, \mathfrak{D})$.

Требуется найти модель $\mathbf{f}_{\hat{\mathcal{A}}}$, другими словами, множество индексов $\hat{\mathcal{A}}$ столбцов матрицы плана $\mathbf{X}$, такое, что

$$\hat{\mathcal{A}} = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S(\hat{\mathbf{w}}_\mathcal{A}|\mathbf{f}_\mathcal{A}, \mathfrak{D}_\mathcal{T}). \tag{203}$$

Оптимизационная задача решается на подвыборке $\mathfrak{D}_\mathcal{T} \subset \mathfrak{D}$ либо на всей выборке $\mathfrak{D}_\mathcal{T} \equiv \mathfrak{D}$ в зависимости от вида функции ошибки $S$.

При постановке задачи считаем, что оценка $\hat{\mathbf{w}}_\mathcal{A}$ параметра модели была получена ранее согласно гипотезе порождения данных.

Следует отметить, однако, что качество прогноза на практике вычисляется с использованием средней абсолютной функции ошибки (англ. mean absolute percentage error)

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{f_i - y_i}{y_i}\right|, \tag{204}$$

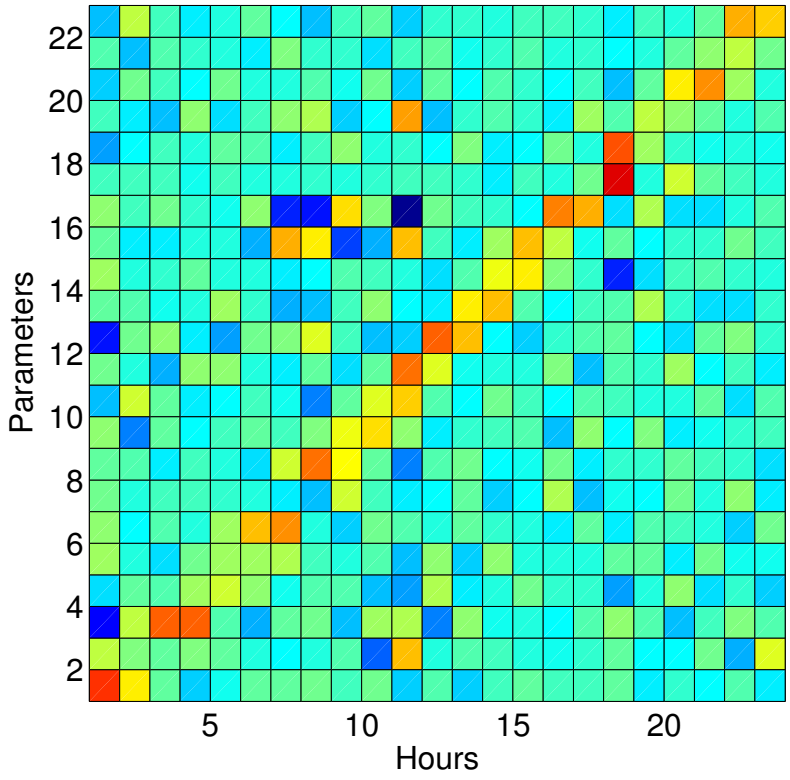



Рис. 76. Наборы параметров при прогнозировании периода временного ряда.

а не среднеквадратичной функции ошибки (англ. mean squared error)

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(f_i - y_i)^2.$$

Для разрешения задачи (204) с функцией ошибки MAPE используется метод наименьших модулей [351, 330, 350]. Решается задача (203). Затем для оценки параметров, доставляющих минимум функции ошибки (204), решается задача линейного программирования для случая наименьших модулей. На рисунке 75 показан суточный прогноз, сделанный по последовательным отсчетам с использованием предыстории на каждом шаге.

**Прогноз следующего периода временного ряда.** При прогнозе значений временного ряда на несколько отсчетов вперед матрицу $\mathbf{X}^*$, приведенную в (202) необходимо перестроить. Так как временной ряд имеет $T$ значений, а прогнозируется значение с индексом $T + h$, то возникают $h$ неизвестных значений, $h \in \{0, \ldots, \kappa\}$. При разбиении матрицы в качестве вектора значений зависимой переменной используется столбец матрицы $\mathbf{X}$ с индексом $j = (T + h) \bmod \kappa$. Так как $T$ делится нацело на $\kappa$, то $(T + h) \bmod \kappa = h \bmod \kappa$. На рис. 76 цветом показаны значения параметров модели при последовательном прогнозировании. Столбцы графика соответствуют элементам вектора параметров при прогнозе одного значения. На рис. 77 показан прогноз на неделю вперед.

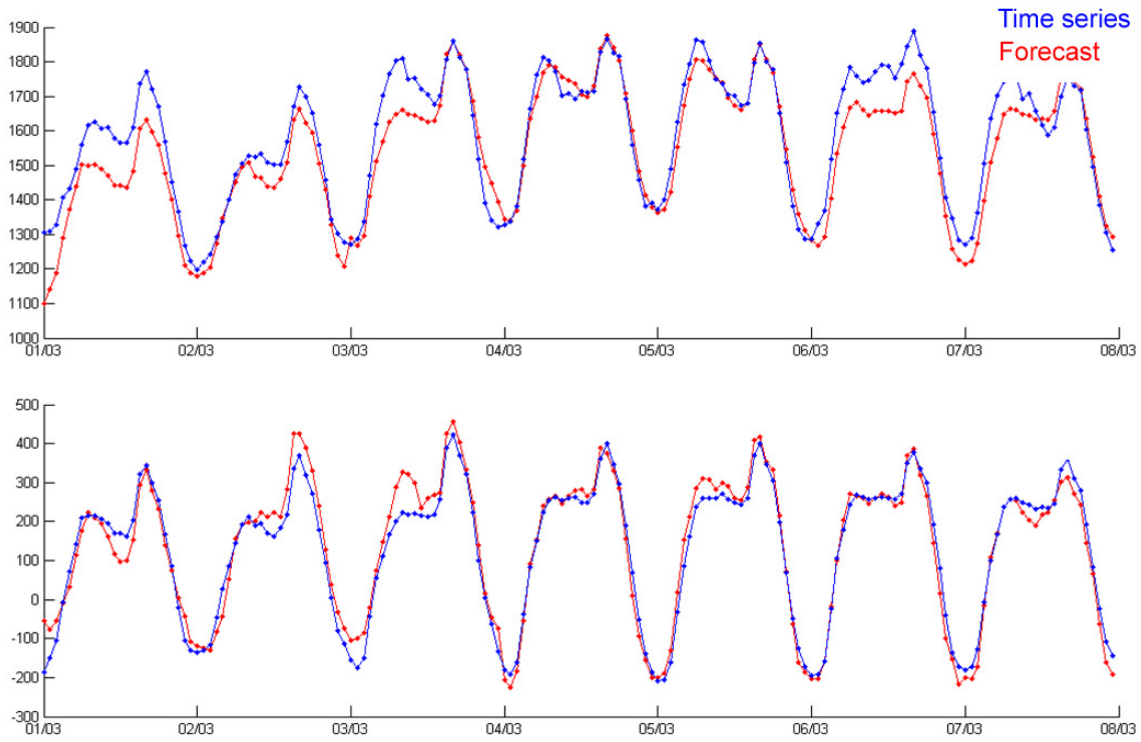



Рис. 77. Прогноз временного ряда на неделю вперед.

### 6.1.2. Векторная авторегрессия и сглаживание

Векторная авторегрессия [249], как метод краткосрочного прогнозирования набора временных рядов, была предложена в качестве альтернативы методу, использующему для получения прогноза систему одновременных линейных уравнений. Векторная авторегрессия является одним из основных методов краткосрочного прогноза макроэкономических показателей [186, 35, 105].

Заданы $n$ временных рядов $\mathbf{s}_1, \ldots, \mathbf{s}_n$. Как и ранее, $\mathbf{s}_j = [x_{1j}, \ldots, x_{(T-1)j}]$, $j = 1, \ldots, n$ — столбцы матрицы $\mathbf{X}$. Рассмотрим матрицу $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \ldots & x_{mn} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\intercal \\ \vdots \\ \mathbf{x}_m^\intercal \end{bmatrix}, \quad \text{где } m = T - 1.$$

Предполагается, что значение прогнозируемого вектора $\mathbf{x}_T = [x_{T1}, \ldots, x_{Tn}]^\intercal$ в момент времени $t = T$ линейно зависит от $H$ предыдущих значений временных рядов,

$$\mathbf{f}_t = \sum_{\tau=1}^{H} W_\tau \mathbf{x}_{t-\tau} + \boldsymbol{\mu}_t,$$

где $H$ называется также глубиной лагирования. Исключим из рассмотрения вектор-слагаемое $\boldsymbol{\mu}_t$ путем добавления временного ряда, состоящего из единиц. При этом число строк матрицы $W$ и число столбцов матрицы $\mathbf{X}$ увеличится на единицу. Сохраним для простоты изложения ранее введенные обозначения размерности матриц и векторов.

Пусть назначена функция ошибки $S$ — евклидова норма вектора невязок при ретроспективном прогнозировании $j$-го временного ряда,

$$S_j = \sum_{\tau=t_1}^{t_2} (f_{\tau j} - x_{\tau j})^2.$$

Требуется найти матрицы параметров $W_\tau$ при $\tau = 1, \ldots, H$, которые бы доставляли минимум функции ошибки на ретроспективном прогнозе в моменты времени $t_1, \ldots, t_2$ для каждого из временных рядов. Так как значения элементов $f_{T1}, \ldots, f_{Tn}$ прогнозируемого вектора $\mathbf{f}_T$ не зависят друг от друга, а зависят только от предыдущих векторов $\mathbf{x}_{T-1}, \ldots, \mathbf{x}_{T-H}$, то задача распадается на ряд задач оценки векторов-строк матриц $W$. Так как принята линейная модель, эту оценку можно получить методом наименьших квадратов. В случае нелинейной модели векторной авторегрессии используется соответствующий алгоритм оценки параметров.

Представим модель прогнозирования в виде

$$\mathbf{f}_t = \sum_{\tau=1}^{H} W_\tau \mathbf{x}_{t-\tau-1} = W_1 \mathbf{x}_{t-1} + \cdots + W_H \mathbf{x}_{t-H-1}.$$

Соединим строки с номером $j$ матриц $W_\tau$, $\tau = 1, \ldots, H$ и обозначим полученный вектор $\boldsymbol{\zeta}_j$. Этот вектор будет играть роль вектора параметров при прогнозировании одного элемента $f_{Tj}$ вектора $\mathbf{f}_T$. Соединим векторы $\mathbf{x}_{t-\tau}$, $\tau = 1, \ldots, H$, транспонируем, и обозначим полученный вектор $\boldsymbol{\varkappa}_t$. Тогда прогнозируемое значение равно скалярному произведению

$$f_{Tj} = \boldsymbol{\zeta}_j^\intercal \boldsymbol{\varkappa}_T.$$

Задача нахождения векторов $\boldsymbol{\zeta}_j$, $j = 1, \ldots, n$, из которых состоят матрицы $W_\tau$ имеет следующий вид. Рассмотрим $j$-й локальный временной ряд $[x_{t_1 j}, \ldots, x_{t_2 j}]^\intercal = \mathbf{y}_j$, состоящий из отсчетов за интервал времени $(t_1, t_2)$. Этот ряд приближается последовательностью

$$[f_{t_1 j}, \ldots, f_{t_2 j}]^\intercal = [\boldsymbol{\zeta}_j^\intercal \boldsymbol{\varkappa}_{t_1}, \ldots, \boldsymbol{\zeta}_j^\intercal \boldsymbol{\varkappa}_{t_2}]^\intercal = \boldsymbol{\varphi}_j,$$

Для решения задачи требуется минимизировать евклидову норму вектора

$$S_j = \|\mathbf{y}_j - \boldsymbol{\varphi}_j\|^2 = \|\mathbf{y}_j - \begin{bmatrix} \boldsymbol{\varkappa}_{t_1}^\intercal \\ \ldots \\ \boldsymbol{\varkappa}_{t_2}^\intercal \end{bmatrix} \boldsymbol{\zeta}_j\|^2.$$

**Авторегрессия и скользящее среднее.** Частным случаем вышеприведенной задачи является задача прогнозирования методом авторегрессии и скользящего среднего. Метод широко распространен и используется для прогнозирования временных рядов как в экспериментальной физике [226], так и в экономике [207].

Задан временной ряд $\mathbf{x} = [x_1, \ldots, x_t, \ldots, x_m]^\intercal$. Предположим, что этот временной ряд содержит две аддитивные составляющие: периодическую составляющую и тренд. Первая составляющая также называется авторегрессионной моделью и имеет вид

$$x_t = \sum_{\tau=1}^{H_{\mathrm{AR}}} \varphi_\tau x_{t-\tau} + c + \varepsilon_t,$$

где $\varphi_1, \ldots, \varphi_{H_{\mathrm{AR}}}$ — параметры модели, $c$ — константа и $\varepsilon_t$ — случайная переменная, реализация которой может быть использована второй моделью.

Вторая составляющая называется моделью скользящего среднего и имеет вид

$$x_t = \sum_{\tau=1}^{H_{\mathrm{MA}}} \theta_\tau \varepsilon_{t-\tau} + \mu + \varepsilon_t,$$

где $\theta_1, \ldots, \theta_{H_{\mathrm{MA}}}$ — параметры модели, $\mu$ —константа, и $\varepsilon_t$ — случайная переменная.

Суммируя две модели, получаем

$$x_t = \sum_{\tau=1}^{H_{\mathrm{AR}}} \varphi_\tau x_{t-\tau} + \sum_{\tau=1}^{H_{\mathrm{MA}}} \theta_\tau \varepsilon_{t-\tau} + c + \varepsilon_t.$$

Предполагается, что регрессионные остатки — реализации случайной величины $\varepsilon_t$, распределены нормально с матожиданием $c$,

$$[\varepsilon_1, \ldots, \varepsilon_m]^\intercal = \boldsymbol{\varepsilon} \sim \mathcal{N}(c\mathbf{1}, \sigma^2 \mathbf{I}_m).$$

При такой гипотезе порождения данных функция ошибки, как и предыдущем параграфе, будет иметь вид суммы квадратов регрессионных остатков:

$$S = \sum_{t=t_1}^{t_2} (f_t - x_t)^2 = \sum_{t=t_1}^{t_2} \varepsilon_t^2.$$

Функция вычисляется при ретроспективном прогнозировании на интервале времени $(t_1, t_2)$. Глубины лагирования двух моделей $H_{\mathrm{AR}}$, $H_{\mathrm{MA}}$ называются структурными параметрами.

Так как вектор параметров $[\varphi_1, \ldots, \varphi_{H_{\mathrm{AR}}}, \theta_1 \ldots \theta_{H_{\mathrm{MA}}}]^\intercal = \mathbf{w}$ входит в модель линейно, то в терминах задачи порождения и выбора моделей, модель авторегрессии и скользящего среднего будет иметь следующий вид. Построим строку $\mathbf{x}_t$ с индексом $t$ матрицы $\mathbf{X}$ с учетом глубины лагирования:

$$\mathbf{x}_t = [x_t, \ldots, x_{t-H_{\mathrm{AR}}}, \varepsilon_t, \ldots, \varepsilon_{t-H_{\mathrm{MA}}}] = [x_t, \ldots, x_{t-H_{\mathrm{AR}}}, \hat{f}_t - x_t, \ldots, \hat{f}_{t-H_{\mathrm{MA}}} - x_{t-H_{\mathrm{MA}}}].$$

Будем считать значения $\hat{f}_t$ в предыдущем выражении фиксированными. Тогда прогнозируемое значение вычисляется как скалярное произведение

$$f_{t+1} = \mathbf{w}^\intercal \mathbf{x}_t,$$

а оценка вектора параметров $\hat{\mathbf{w}}$ является решением задачи

$$\hat{\mathbf{w}} = \arg\min_{\hat{\mathbf{w}} \in \mathcal{W}} \|[x_{t_1}, \ldots, x_{t_2}]^\intercal - [\hat{f}_{t_1}, \ldots, \hat{f}_{t_2}]^\intercal\|^2$$

ретроспективного прогноза на интервале времени $(t_1, t_2)$. Так как прогнозируемые значения временного ряда $\hat{f}$ зависят, в свою очередь, от оценок параметров $\hat{\mathbf{w}}$, $\hat{f} = f(\hat{\mathbf{w}}, \mathbf{x})$, то оценка параметров выполняется итеративно.

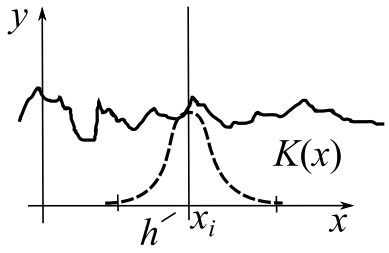



Рис. 78. Использование функции плотности при непараметрическом прогнозировании.

**Параметрическое сглаживание временных рядов.** Представим задачу о вычислении скользящего среднего следующим образом. Пусть при решении задачи восстановления регрессии $\mathsf{E}(y|\mathbf{x})$ значения зависимой переменной $y$ определяются по регрессионной выборке $\mathfrak{D} = \{(\mathbf{x}_i, y_i)|i = 1, \ldots, m\}$ не только вектором $\mathbf{x}$, но и его окрестностью. В качестве примера приведем две регрессионные модели: модель Парзена-Розенблатта [223, 236]

$$f(x) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{x_i - x}{h}\right)$$

и модель Надрая-Ватсона[88, 354]

$$f(x,y) = \frac{\sum_{i=1}^{m} y_i w_i(x)}{\sum_{i=1}^{m} w_i(x)} = \frac{\sum_{i=1}^{m} y_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{m} K\left(\frac{x_i - x}{h}\right)}.$$

Ядро $K$ соответствующее предполагаемой плотности распределения зависимой переменной $y$ при заданном $x$, является неотрицательной симметричной интегрируемой функцией, играющей роль взвешивающей функции: при удалении от точки $x_i$ ее значение уменьшается. Функция ошибки при восстановлении регрессии имеет вид

$$S = \sum_{i=1}^{m} \beta_i(x)(f_i - y_i)^2 \to \min,$$

где весовые коэффициенты $\beta_i$ определены значениями ядра $K$:

$$\beta_i(x) = K\left(\frac{(x_i - x)}{h}\right),$$

.

Параметр $h$ — ширина окна сглаживания, см. рис.78. В предположении о нормальном распределении регрессионных остатков с дисперсией $\sigma^2$, при гауссовом ядре

$$K = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - x)^2\right)$$

оптимальная ширина окна сглаживания [354] равна

$$h_{\text{opt}} = 1.059\frac{\sigma}{\sqrt[5]{m}}.$$

### 6.1.3. Построение криволинейных моделей

Рассмотрим одну из наиболее распространенных задач — задачу выбора мономов полиномиальной регрессионной модели. Задана выборка $\mathfrak{D} = \{(\boldsymbol{\xi}_i, y_i)\}$, $i = 1, \ldots, m$, $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_u, \ldots, \xi_U]^\intercal \in \mathbb{R}^U$. Задано множество $G = \{g_v\}$, $v = 1, \ldots, V$ порождающих функций, не содержащих параметры в качестве аргументов. Например $G = \{\xi^{-1}, \xi^0, \sqrt{\xi}, \mathrm{id}(\xi), \ln(\xi), \tanh(\xi)\}$. Как видно из примера, в множество порождающих функций входит как сама независимая переменная $\xi = \mathrm{id}(\xi)$, так и ее функции $g_v(\xi)$.

Для построения полиномиальных криволинейных моделей выполним следующие два шага. Во-первых, построим декартово произведение $G \times \boldsymbol{\xi}$ набора непорождаемых независимых переменных $\boldsymbol{\xi}$ и порождающих функций $G$ и обозначим $a_\iota$ — всевозможные суперпозиции $g_v(\xi_u)$, поставленные в соответствие элементам этого декартова произведения. Во-вторых, построим все произведения элементов $a_\iota$ степени, не превосходящей заданное число $P$. Другими словами,

$$a_\iota = g_v(\xi_u), \quad \text{где индекс } \iota = (v-1)U + u$$

и

$$x_j = \prod \underbrace{a_{\iota_1} \ldots a_{\iota_p}}_{p \text{ раз}}, \quad \text{где } \iota \in \{1, \ldots, UV\}, \quad p \in \{1, \ldots, P\}.$$

Вышеприведенные шаги также можно представить в виде диаграммы

$$\xi_u \xrightarrow{g_v} g_v(\xi_u) \equiv a_\iota \xrightarrow{\Pi^p} x_j.$$

Индекс $j$ монома $x_j$ принадлежит множеству индексов $\mathcal{J}$. Так как число сочетаний с повторениями из $UV$ по $p$ элементов декартова произведения $G \times \boldsymbol{\xi}$ равно

$$\binom{UV+p-1}{p} = (-1)^p \binom{-UV}{p} = \frac{(UV+p-1)!}{p!\,(UV-1)!},$$

то количество элементов множества $\mathcal{J} \ni j$ равно

$$|\mathcal{J}| = \sum_{p=1}^{P} \frac{(UV+p-1)!}{p!\,(UV-1)!}.$$

Полиномиальные криволинейные модели $f$ являются линейными моделями относительно своих параметров $f = f(\mathbf{w}, \mathbf{x})$,

$$f_{\mathcal{A}}(\mathbf{w}_{\mathcal{A}}, \mathbf{x}) = \sum_{j \in \mathcal{A}} w_j x_j, \tag{205}$$

где $\mathcal{A} \subseteq \mathcal{J}$ является набором индексов свободных переменных $x_j$, $\mathbf{w}_{\mathcal{A}}$ — вектор параметров с числом элементов $|\mathcal{A}|$. Таким образом, задан класс регрессионных моделей $\mathfrak{F} = \{f\}$ — параметрических функций.

В обозначениях всевозможных суперпозиций $a_\iota = g_v(\xi_u)$ модель (205) представима в виде полинома Колмогорова-Габора

$$f(\mathbf{w}, \mathbf{x}) = \sum_{\iota=1}^{UV} w_\iota a_\iota + \sum_{\iota=1}^{UV} \sum_{\kappa=1}^{UV} w_{\iota\kappa} a_\iota a_\kappa + \sum_{\iota=1}^{UV} \sum_{\kappa=1}^{UV} \sum_{\tau=1}^{UV} w_{\iota\kappa\tau} a_\iota a_\kappa a_\tau + \cdots.$$

Например, пусть задана выборка — множество $\mathfrak{D} = \{\boldsymbol{\xi}, \mathbf{y}\}$

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \dots \\ \xi_m \end{bmatrix}, \quad \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix}.$$

Задана регрессионная модель — квадратичный полином

$$f = w_3\xi^2 + w_2\xi^1 + w_1\xi^0 = \sum_{j=1}^{3} w_j \xi^{j-1}.$$

Эта модель является линейной относительно параметров. Для нахождения оптимального значения вектора параметров $\mathbf{w} = [w_1, w_2, w_3]^\intercal$ выполняется следующее переобозначение:

$$x_{i1} = \xi_i^0, \quad x_{i2} = \xi_i^1, \quad x_{i3} = \xi_i^2.$$

Тогда матрица $\mathbf{X}$ значений свободной переменной $x_{ij}$ будет иметь вид

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} \end{bmatrix}.$$

Параметры $\mathbf{w}$ находятся из решения задачи $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$ при $\|\boldsymbol{\varepsilon}\|^2 \to \min$.

**Восстановление форм геометрических фигур.** Рассмотрим задачу восстановления линейной регрессии в случае, когда регрессионная модель не присутствует явно в постановке прикладной задачи. Для решения прикладной задачи необходимо ее переформулировать.

В качестве примера рассмотрим задачу контроля состояния трубопроводов. Пусть заданы координаты точек окружности (сечения трубы) — множество точек $\{(x, y)\}$, измеренных с некоторой погрешностью. Требуется найти центр $(c_1, c_2)$ и радиус $r$ окружности.

Запишем регрессионную модель — координаты окружности относительно центра и радиуса и выделим линейно входящие компоненты:

$$\begin{aligned} &(x - c_1)^2 + (y - c_2)^2 = r^2, \\ &2xc_1 + 2yc_2 + (r^2 - c_1^2 - c_2^2) = x^2 + y^2, \\ &c_3 = (r^2 - c_1^2 - c_2^2). \end{aligned}$$

Тогда матрица плана $\mathbf{X}$, параметры $\mathbf{w}$ и зависимая переменная $\mathbf{y}$ линейной модели $\mathbf{X}\mathbf{w} = \mathbf{y} + \boldsymbol{\varepsilon}$ будет иметь вид

$$\begin{bmatrix} 2x_1 & 2y_1 & 1 \\ 2x_2 & 2y_2 & 1 \\ \vdots & \vdots & \vdots \\ 2x_m & 2y_m & 1 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} x_1^2 + y_1^2 \\ x_2^2 + y_2^2 \\ \vdots \\ x_m^2 + y_m^2 \end{bmatrix} + \boldsymbol{\varepsilon}.$$

Параметры $\mathbf{w} = [c_1, c_2, c_3]^\intercal$ находятся, как и ранее, из решения задачи $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$ при $\|\boldsymbol{\varepsilon}\|^2 \to \min$. Аналогичным путем ставятся задачи нахождения параметров эллипсоида, параллелограмма и других геометрических фигур по измерениям координат точек, находящихся на их границах.

Все три вышеприведенных примера в данном разделе иллюстрируют одну из наиболее актуальных в современном регрессионном анализе задач — задачу нахождения такого множества индексов

$$\hat{\mathcal{A}} = \arg\min_{\mathcal{A}\subseteq\mathcal{J}} S(\hat{\mathbf{w}}|\mathfrak{D}_\mathcal{T}, f),$$

которое бы имело оптимальную мощность. Оптимизационная задача решается на подвыборке $\mathfrak{D}_\mathcal{T} \subset \mathfrak{D}$ либо на всей выборке $\mathfrak{D}_\mathcal{T} \equiv \mathfrak{D}$ в зависимости от вида функции ошибки $S$.

### 6.1.4. Порождение нелинейных моделей для оценки волатильности случайных процессов

Задача оценки дисперсии стационарного случайного процесса в финансовой математике называется задачей оценки волатильности [358]. Она рассматривается вместе с задачей прогнозирования волатильности при вычислении справедливой стоимости биржевых опционов [75]. Справедливая цена опциона (теоретически обоснованная минимальная цена, при которой продавец может выполнить свои обязательства по договору) вычисляется с помощью модели Блэка–Шоулза [145].

Эта модель включает оценку волатильности цены базового инструмента. Волатильность — это величина, равная стандартному отклонению стоимости базового инструмента, вычисленная на основе текущей стоимости финансового инструмента, в предположении, что рыночная стоимость финансового инструмента отражает ожидаемые риски. В предположениях, на которых основана модель Блэка–Шоулза, волатильность не зависит ни от цены исполнения опциона, ни от времени до его исполнения. Однако на практике волатильность зависит от этих двух величин, что и является основанием для поиска этой зависимости.

Заданы сетка цен исполнения опциона $\mathcal{K} = \{K_s\}$, время до исполнения опциона, выраженное в годах $\mathcal{T} = \{t_\tau\}$. В каждый момент времени $t_\tau$ для цены исполнения $K_s$ известна историческая цена опциона $C^{\text{hist}}(K, t)$. Заданы безрисковая ставка $B$ и цена базового инструмента $P_t$ в каждый момент времени. Известна предполагаемые значения волатильности $\sigma^{\text{imp}} = \sigma^{\text{imp}}(K, t)$, высчиленная по формуле Блэка-Шоулза как аргумент минимума разности между исторической и справедливой ценой опциона:

$$\sigma^{\text{imp}} = \arg\min_{\sigma\in\mathbb{R}_+} \big(C^{\text{hist}} - C(\sigma, K, t, P, B)\big), \tag{206}$$

где $C^{\text{hist}}$ — историческая цена опциона, $C$ — справедливая цена опциона, вычисленная по формуле Блэка–Шоулза, $P$ — цена базового инструмента, $B$ — банковская процентная ставка, $K$ — цена исполнения опциона, и $t$ — время до исполнения опциона. Предлагается рассматривать задачу оценки волатильности как задачу восстановления регрессии в предположении о нормальном распределении волатильности. При решении задачи $\mathsf{E}(y|\mathbf{x}) = \mathsf{E}(\sigma|[t, K]^\intercal)$ используются исторические временные ряды $C^{\text{hist}}_{tK}$, $P_t$, набор цен исполнения $\{K\}$ и константа $B$.

По формуле (206) вычисляется предполагаемая волатильность $\sigma^{\text{imp}}$. Требуется найти модель улыбки волатильности

$$\sigma^{\text{imp}} = \sigma^{\text{imp}}(K, t).$$

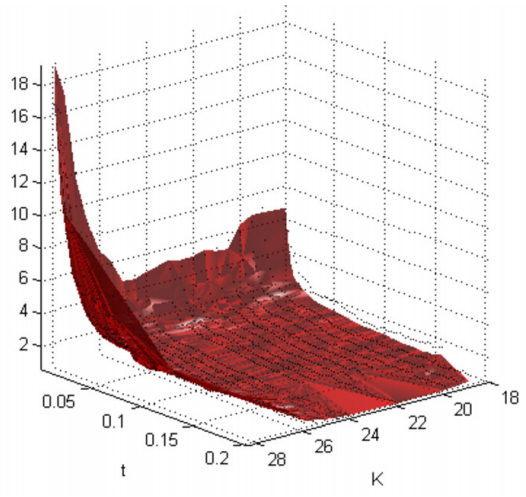



Рис. 79. Модель зависимости предполагаемой волатильности от цены и времени до исполнения опциона.

Для этого подготовим регрессионную выборку. Индексы $s$ и $\tau$ задают значения цены исполнения и времени до исполнения опциона. Декартово произведение множеств индексов $\{s\} \times \{\tau\}$ задает множество пар $(K_s, t_\tau)$ и, соответственно, декартово произведения множеств $\mathcal{K} \times \mathcal{T}$. Присвоим каждому элементу этого произведения номер $i \in \{1, \ldots, N\}$ и представим элементы произведения в виде векторов-столбцов вместе с соответствующими значениями волатильности $\sigma$:

$$\mathbf{y} = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_i \\ \vdots \\ \sigma_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} K_1 & t_1 \\ \vdots & \vdots \\ K_i & t_i \\ \vdots & \vdots \\ K_m & t_m \end{bmatrix}.$$

Регрессионная выборка $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\} = \{([t_i, K_i]^\intercal, \sigma_i^{\text{imp}})\}$, $i \in \mathcal{I} = \{1, \ldots, m\}$ строится с помощью исходных данных — исторических цен опциона $C_{\tau\kappa}$ и базового инструмента $P_\tau$ следующим образом. Даны отсчеты времени $t \in \{t_\tau\} = \mathcal{T}$, набор цен исполнения опциона $K \in \{K_\kappa\} = \mathcal{K}$. Для каждого значения $K \in \mathcal{K}$ и $t \in \mathcal{T}$ вычисляется значение предполагаемой волатильности как аргумент минимума (206)

$$\sigma_{\tau\kappa} = \arg\min_{\sigma \in \mathbb{R}_+} \left(C_{\tau\kappa}^{\text{hist}} - C(\sigma, P_\tau, B, K_\kappa, t_\tau)\right).$$

Для построения выборки двойной индекс переменной $\sigma^{\text{imp}}$ заменяется на одинарный

$$\sigma_{\tau\kappa}^{\text{imp}} \mapsto \sigma_i, \quad i = \tau + (\kappa - 1)|\mathcal{T}|,$$

то есть вектор $\mathbf{x}_i$, поставлен в соответствие элементу декартова произведения:

$$\mathbf{x}_i = [t_\tau, K_\kappa]^\intercal \in \mathcal{T} \times \mathcal{K}.$$

Требуется восстановить регрессию

$$\sigma_i^{\text{imp}} = f(\mathbf{w}, [t_i, K_i]^{\mathsf{T}}) + \varepsilon_i, \quad \text{в принятых ранее обозначениях,} \quad y_i = f(\mathbf{w}, \mathbf{x}_i) + \varepsilon_i,$$

то есть выбрать модель из заданного семейства моделей $f \in \mathfrak{F}$ и оценить ее параметры $\mathbf{w}$, исходя из предположения о нормальном распределении зависимой переменной $y \sim \mathcal{N}(f, \sigma^2 I)$.

На рис. 79 показана одна из нелинейных моделей оценки и прогноза волатильности. По оси абсцисс отложена цена исполнения опциона $K$ (доллар США), по оси ординат отложено время до исполнения $t$ (доли года). Точками на рисунке показаны исходные данные. По оси аппликат отложены подразумеваемая волатильность $\sigma^{\text{imp}}$ и восстановленная волатильность $f(\mathbf{w}, \mathbf{x})$. Полученная модель является адекватной и удовлетворительно приближает исторические данные.

### 6.1.5. Использование параметров модели в качестве независимых переменных

Рассмотрим задачу прогнозирования, в которой в качестве независимых переменных используются параметры вспомогательной модели. Требуется спрогнозировать концентрацию кислорода в выпускном коллекторе двигателя внутреннего сгорания. Каждый элемент регрессионной выборки соответствует одному рабочему циклу двигателя (измерения выполнены на одном цилиндре) и включает температуру масла в двигателе, потребление топлива, крутящий момент, мощность, угол поворота во время первого и второго впрыскивания топлива, длительность замкнутого состояния контактов прерывателя, дымность, концентрацию $NO_x$, CO, $CO_2$, $HC_{\text{prop}}$ в выхлопе и температуру выхлопа. Помимо этого измеряется давление в камере сгорания; измерение производится 7200 раз в течение рабочего цикла (два полных оборота коленчатого вала). Полагая, что коленчатый вал вращается равномерно, будем считать измерения давления временным рядом, который соответствует одному рабочему циклу. На рис. 80 точками показано изменение давления (ось аппликат) в зависимости от номера цикла (ось абсцисс) и от угла поворота (ось ординат).

Рис. 80. Зависимость давления в камере сгорания от угла поворота коленчатого вала и номера рабочего цикла

Обозначим величины, измеряемые в течение $i$-го рабочего цикла однократно $x_{i1}, \ldots, x_{i13}$, а значения давления — $s_{i1}, \ldots, s_{i\tau}, \ldots, s_{7200}$. Соединяя два этих вектора, и обозначая прогнозируемую величину $y_i$, концентрацию кислорода, получаем элемент $(\mathbf{x}_i, y_i)$ регрессионной выборки $\mathfrak{D}$. Эксперимент состоит из измерений сотен циклов, поэтому необходимо сокращать число признаков. Предлагается два подхода:

1) выбрать фиксированное число отсчетов временного ряда, которые снижают значение функции ошибки при ретроспективном прогнозе,

2) приблизить временной ряд вспомогательной моделью — параметрическим семейством функций и использовать параметры, оцененные в $i$-м временном ряде в качестве значений признаков.

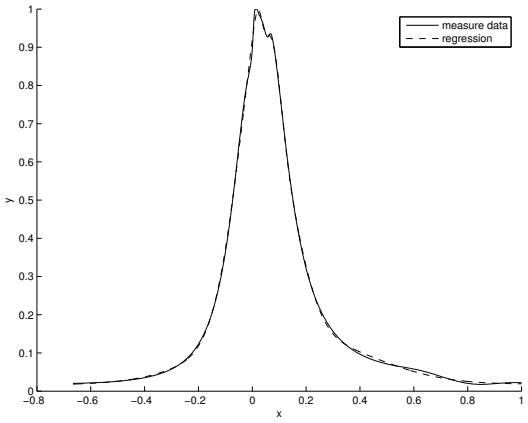



Рис. 81. Зависимость давления в камере сгорания от угла поворота коленчатого вала.

Опуская ранее обсуждавшийся первый подход, рассмотрим задачу приближения набора временных рядов. Без ограничения общности будем считать индекс $\tau$ значения $s_{i\tau}$ временного ряда $\mathbf{s}_i$ независимой переменной. Тогда задача восстановления регрессии будет иметь вид

$$\mathbf{s}_i = \boldsymbol{\phi}(\mathbf{u}, \boldsymbol{\tau}) + \boldsymbol{\varepsilon}_i,$$

где $\mathbf{u}$ — вектор параметров. Считая вектор $\boldsymbol{\varepsilon}_i$ нормально распределенным, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\hat{\boldsymbol{\phi}}, \sigma^2 I)$, используем для оценки вектора параметров $\hat{\mathbf{u}}_i$ функцию ошибки

$$\hat{\mathbf{u}}_i = \arg\min_{\mathbf{u}\in\mathcal{W}} S_i(\mathbf{u}), \qquad S_i(\mathbf{u}) = \|\boldsymbol{\phi}(\mathbf{u}, \boldsymbol{\tau}) - \mathbf{s}_i\|^2.$$

Так как эксперимент содержит серию временных рядов $\mathbf{s}_i$, $i = 1, \ldots, m$ поставим задачу выбора оптимальной модели $\boldsymbol{\phi} \in \mathfrak{F}$ так, чтобы функции регрессии, соответствующие этой модели, доставляли минимум среднему значению функции ошибки

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi}\in\mathfrak{F}} S^*(\boldsymbol{\phi}), \qquad S^*(\boldsymbol{\phi}) = \frac{1}{m}\sum_{i=1}^{m} S_i(\hat{\mathbf{u}}_i) = \frac{1}{m}\sum_{i=1}^{m} \|\boldsymbol{\phi}(\hat{\mathbf{u}}_i, \boldsymbol{\tau}) - \mathbf{s}_i\|^2.$$

Семейство регрессионных моделей $\mathfrak{F}$ при этом может быть задано экспертно как множество существенно нелинейных моделей ограниченной структурной сложности.

После выбора оптимальной модели $\boldsymbol{\phi}$ и нахождения оценок ее параметров $\hat{\mathbf{u}}_i$ на каждом временном ряде $\mathbf{s}_i$ регрессионная выборка имеет вид

$$\mathfrak{D} = \{([\mathbf{x}_i, \hat{\mathbf{u}}_i]^\mathsf{T}, y_i)\},\ i = 1, \ldots, m.$$

Эта регрессионная выборка используется при решении основной задачи прогнозирования концентрации кислорода и выбора модели

$$\mathbf{y}_i = f(\mathbf{w}, [\mathbf{x}_i, \hat{\mathbf{u}}_i]^\mathsf{T}) + \epsilon_i.$$

На рис. 81 показана зависимость давления в камере сгорания от угла поворота коленчатого вала. По оси абсцисс отложен угол в градусах, а по оси ординат — давление в МПа. Нулевому углу соответствует верхняя мертвая точка. Начало временного ряда соответствует углу в $-360$ градусов, конец — углу в $+359.9$ градусов. Всего один полный цикл насчитывает 7200 отсчетов. Лабораторный эксперимент включает измерения давления 122 полных

циклов. Сплошной кривой показаны исходные данные, пунктиром показаны значения модели № 2. По оси абсцисс отложено значение свободной переменной, по оси ординат — значение зависимой переменной. Временной ряд, приближенный данной кривой, содержит четыре тысячи отсчетов. Для верификации полученных моделей использовалось 118 временных рядов.

Экспертами задано множество порождающих функций $G$ из которых порождаются регрессионные модели. Список функций приведен в таблице 20.

Выбор моделей произведен из более чем тысячи порожденных моделей. В таблице 21 приведены три модели, которые доставили наименьшие ошибки $S$ при заданной структурной сложности. Дополнительно качество моделей оценивалось по ошибкам $\rho_1, \rho_2$ и числу элементов вектора параметров $\mathbf{w}$. Ошибка $\rho_1$ — среднеквадратичная относительная ошибка

$$\rho_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - f(\mathbf{x}_i)}{\max(y_i)}\right)^2},$$

ошибка $\rho_2$ — максимальная относительная ошибка

$$\rho_2 = \max_{i=1,\ldots,n}\frac{|y_i - f(\mathbf{x}_i)|}{\max(y_i)}.$$

В строке «Описание» таблицы 21 показана структура модели в виде дерева. В качестве примера рассмотрим модель №2. Эта модель состоит из суперпозиции восьми функций $f_2 = g_1(g_2(g_3(g_4(g_5(x), g_6(x)), g_7(x)), x), g_8(x))$. Функции $g_1 = \div(\varnothing, \cdot, \cdot)$ и $g_2, g_3, g_4 = +(\varnothing, \cdot, \cdot)$, сложения и умножения, имеют первым аргументом пустой вектор параметров; $g_5, g_6, g_7 = h(\mathbf{b}_i, \cdot)$, $i = 1, 2, 3$, и $g_8 = l(\mathbf{b}_4, \cdot)$. Функции

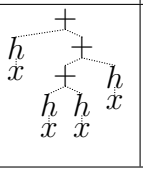
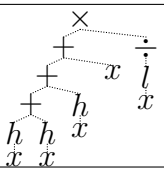
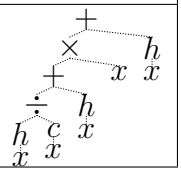
$$h = \frac{\lambda_i}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right) + a_i$$

имеют векторы параметров $\mathbf{b}_i = [\lambda_i, \sigma_i, \xi_i, a_i]^\intercal$, и функция $l = (ax + b)$ имеет вектор параметров $\mathbf{b}_4 = [a, b]^\intercal$.

Таблица 20. Множество порождающих функций.

| № | Функция | Описание | Параметры |
|---|---|---|---|
| Функции двух зависимых переменных, $g(\mathbf{b}, x_1, x_2)$ | | | |
| 1 | plus | $y = x_1 + x_2$ | – |
| 2 | times | $y = x_1 x_2$ | – |
| 3 | divide | $y = x_1/x_2$ | – |
| Функции одной зависимой переменой, $g(\mathbf{b}, x_1)$ | | | |
| 4 | multiply | $y = ax$ | $a$ |
| 5 | add | $y = x + a$ | $a$ |
| 6 | gaussian | $y = \frac{\lambda}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$ | $\lambda, \sigma, \xi, a$ |
| 7 | linear | $y = ax + b$ | $a, b$ |
| 8 | parabolic | $y = ax^2 + bx + c$ | $a, b, c$ |
| 9 | cubic | $y = ax^3 + bx^2 + cx + d$ | $a, b, c, d$ |
| 10 | logsig | $y = \frac{\lambda}{1+\exp\left(-\sigma(x-\xi)\right)} + a$ | $\lambda, \sigma, \xi, a$ |

Таблица 21. Вспомогательные модели, приближающие временной ряд.

| № модели | 1 | 2 | 3 |
|---|---|---|---|
| Ошибка $\rho_1$ | 0.0034 | 0.0037 | 0.0035 |
| Ошибка $\rho_2$ | 0.0421 | 0.0325 | 0.00338 |
| Число параметров | 16 | 14 | 16 |
| Описание |  |  |  |

Легенда: h — gaussian, c — cubic, l — linear,
$+$ — plus, $\times$ — times,$\div$ — divide

Модель $f_2$ можно переписать в виде

$$f(\mathbf{w}, \mathbf{x}) = \left( x + \sum_{i=1}^{3} h(\mathbf{b}_i, x) \right) \div l(\mathbf{b}_4, x),$$

где $\mathbf{x} = x$, и $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots \mathbf{b}_3 \vdots \mathbf{b}_4$. Развернутый вид модели:

$$y = (ax + b)^{-1} \left( x + \sum_{i=1}^{3} \left( \frac{\lambda_i}{\sqrt{2\pi}\sigma_i} \exp\left( -\frac{(x - \xi_i)^2}{2\sigma_i^2} \right) + a_i \right) \right).$$

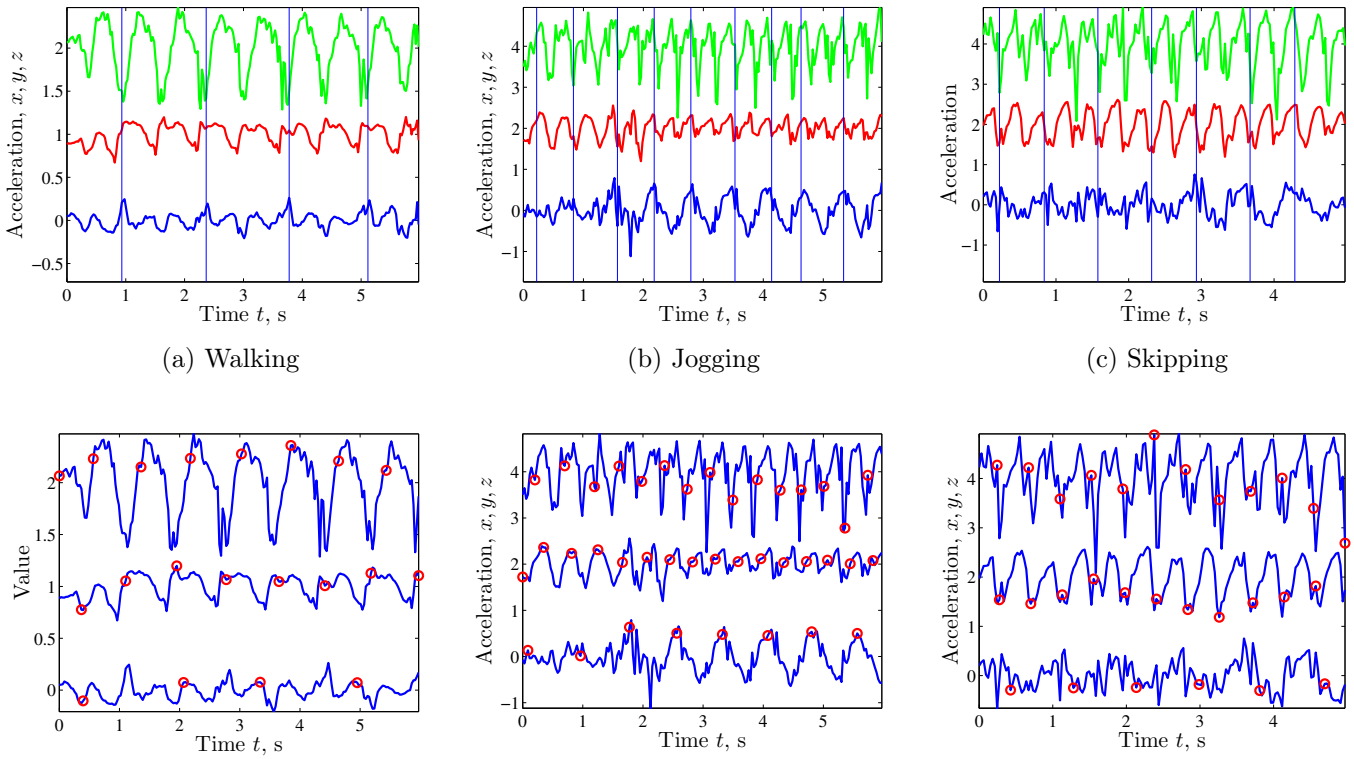
## 6.2. Разметка временных рядов в задачах прогнозирования

Процедура разметки временных рядов разбивает ось времени на сегменты так, что что временной ряд приближается внутри одного сегмента некоторой регрессионной моделью из конечного набора. Нахождение разметки является одной из задач регрессионного анализа. Для решения, например, часть временного рядя, находящегося внутри сегмента приближается отрезком, а весь временной ряд — ломаной. Более сложные способы приближения рассмотрены в разделе «Многомерные адаптивные регрессионные сплайны».

### 6.2.1. Локальное прогнозирование и аппроксимация временных рядов

Метод локального прогнозирования используется для прогнозирования апериодических временных рядов, содержащих повторяющийся сегмент [225, 343, 346, 158, 298]. Задан временной ряд, в котором можно выделить некоторое число временных интервалов таких, что «поведение» ряда на этих интервалах можно охарактеризовать как закономерное. При прогнозировании предлагается выбрать интервалы с «похожей» предысторией, а окончание прогнозируемого интервала вычислить как среднее продолжений найденных интервалов. Такое прогнозирование позволяет избежать использования вспомогательных регрессионных моделей для локальной аппроксимации, описанных в предыдущем разделе. Поэтому этим методом можно прогнозировать временные ряды достаточно сложной формы при единственном условии: прогнозируемый сигнал должен достаточно регулярно повторяться. В качестве

примера приведем прикладные задачи прогнозирования пульсовой волны, энцефалограммы, электрокардиограммы [204, 111, 205, 246, 68], личной подписи [213] и физической активности человека, см. рис. 82.




(a) Walking (b) Jogging (c) Skipping

Рис. 82. Сверху: размеченные временные ряды; вертикальными линиями обозначены границы сегментов. Снизу: результаты сегментирования; красным выделены начало/конец сегмента.

### 6.2.2. Нахождение локального прогноза

Рассмотрим временной ряд $\mathbf{s} = [s_1, \ldots, s_T]^\intercal$ и его всевозможные локальные сегменты $\mathbf{x}_i = [s_i, \ldots, s_{i+n-1}]^\intercal$ длинной $n$ начинающиеся с элемента с номером $i$. Всего таких сегментов $m = T - n + 1$. Представим полученные сегменты в виде матрицы $X = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_m^\intercal]^\intercal$. Для того, чтобы вычислить расстояние между сегментами, введем функцию расстояния $\rho(\mathbf{x}_i, \mathbf{x}_k)$. В качестве примера такой функции приведем взвешенную евклидову метрику

$$\rho(\mathbf{x}_i, \mathbf{x}_k) = (\mathbf{x}_i - \mathbf{x}_k)^\intercal \mathrm{diag}(\lambda_1, \ldots, \lambda_n)(\mathbf{x}_i - \mathbf{x}_k). \tag{207}$$

Функция расстояния и диагональная матрица весовых коэффициентов $\mathrm{diag}(\boldsymbol{\lambda})$, где $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]^\intercal$, задаются экспертно в зависимости от вида прикладной задачи или же оптимизируются согласно назначенной функции ошибки. Необходимо, чтобы выбираемая функция расстояния между словами $\rho(\mathbf{x}, \mathbf{y})$ была метрикой.

Решим задачу кластеризации, используя EM-алгоритм [121] или алгоритм $k$-средних [151], и поставим в соответствие каждому сегменту $\mathbf{x}_i$ метку кластера $y_i$ из конечного алфавита, полученного по результатам кластеризации.

Построение прогноза заключается в следующем. Пусть известны последние $H$ элементов временного ряда $\mathbf{s}$, причем $H < n$. Обозначим их $\mathbf{u} = [s_{T-H}, \ldots, s_T]^\intercal$. Найдем $k$ векторов $\bar{\mathbf{x}}_i$, ближайших к вектору $\mathbf{u}$. Здесь $\bar{\mathbf{x}}_i$ означает, что вектор содержит только $H$ первых элементов вектора $\mathbf{x}_i$. Прогнозируемые значения временного ряда определяются как последние $n-H-1$ элементов линейной комбинации $k$ векторов $\mathbf{x}_i$,

$$\mathbf{x} = [\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}]\mathbf{w} \quad \text{при ограничении} \quad \|\mathbf{w}\|_1 = 1.$$

Вектор весов $\mathbf{w}$ может быть оптимизирован или задан экспертно, например $\mathbf{w} = [\frac{1}{k}, \ldots, \frac{1}{k}]^\intercal$.

Таблица 22. Результаты прогноза объемов потребления электроэнергии.

| Metrics | (208) | (209) | (210) |
|---|---|---|---|
| best $k$ | 6 | 26 | 6 |
| $\lambda$ | 1 | 0,6 | 1 |
| $p$ | 2 | 2 | 5 |
| SMAPE,% | 7,44 | 5,74 | 7,37 |

В таблице 22 и на рисунке 83 приведены результаты локального прогнозирования. Слева рисунке 83 зеленым цветом показан исходный ряд, синим — прогноз, красным — невязка. При прогнозировании были использованы следующие метрики:

евклидова:

$$\rho_E(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})}. \tag{208}$$

диагонально взвешенная евклидова:

$$\rho_{wE}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{\Lambda}^2(\mathbf{x}-\mathbf{y})}, \text{ где } \mathbf{\Lambda} = \operatorname{diag}(\lambda). \tag{209}$$

метрика Минковского $L_p$:

$$\rho_{L_p}(\mathbf{x}, \mathbf{y}) = (\sum_i |x_i - y_i|^p)^{1/p}, \text{ где } p \in \mathbb{N}. \tag{210}$$

Для оценки качества прогноза используется функция ошибки SMAPE(Symmetric Mean Absolute Percent):

$$\text{SMAPE}(s, \hat{s}, n, t) = \frac{1}{t}\sum_{i=1}^{t} \frac{|\hat{s}_{n+i} - s_{n+i}|}{|\hat{s}_{n+i} + s_{n+i}|/2} * 100\%. \tag{211}$$

### 6.2.3. Кусочно-линейная аппроксимация

Рассмотрим один временной ряд $\mathbf{s} = [s_1, \ldots, s_T]^\intercal$ и соответствующий набор отсчетов времени $1, \ldots, T$. Как и ранее, считаем, что отсчеты времени тождественны индексам элементов вектора $\mathbf{s}$. Построим кусочно-линейную аппроксимацию ряда, считая дисперсию $\hat{\sigma}^2_{\mathbf{r}}$ вектора регрессионных остатков $\mathbf{r} = \mathbf{s} - \mathbf{f}$ известной и ненулевой. Для нахождения границ сегмента $(\tau_k, \tau_{k+1})$ используем метод поэлементного добавления значений временного ряда в локальную регрессионную выборку.

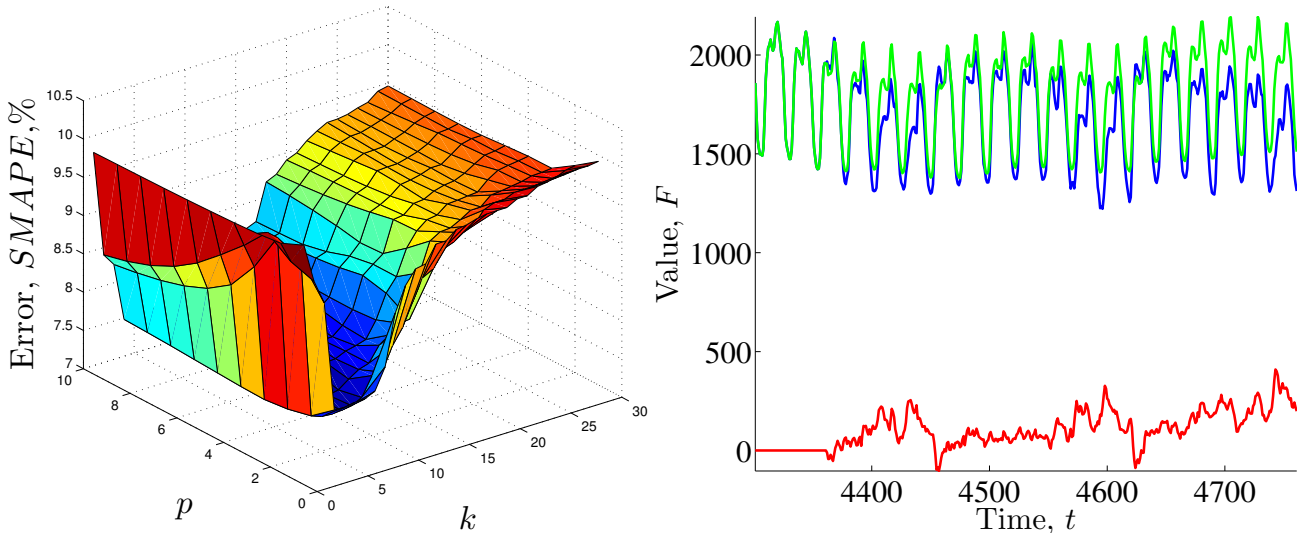



Рис. 83. Величина ошибки и построение прогноза объема потребления электроэнергии.

**Определение 21.** *Локальная регрессионная выборка — подмножество* $\mathfrak{D}' \subseteq \mathfrak{D} = (\mathbf{X}, \mathbf{y}) = \{(\mathbf{x}_i, y_i)\}$ *пар значений зависимой и независимых переменны, заданное множеством из индексов* $\{i\}$*, таким образом, что в локальную выборку* $\mathfrak{D}'$ *попадает любая пара* $(\mathbf{x}_i, y_i)$ *— элемент множества* $\mathfrak{D}$ *с индексом* $i$*,*

$$\mathfrak{D}' = \{(\mathbf{x}_i, y_i) \in \mathfrak{D} : i_{\min} \leqslant i \leqslant i_{\max}\},$$

*где границы* $i_{\min}, i_{\max}$ *заданы.*

Пусть левая граница $\tau_k$ сегмента с номером $k$ известна. Пошагово найдем правую границу $\tau_{k+1}$ этого сегмента. На каждом шаге алгоритма локальная выборка $\mathbf{s}_k$ содержит элементы $s_\tau$ временного ряда, $\mathbf{s}_k = \{s_{\tau_k}, \ldots, s_{\tau_{k+1}}\}$. На первом шаге алгоритма она содержит два элемента с индексами $\tau_k$ и $\tau_k + 1 = \tau_{k+1}$. Построим линейную регрессию $f_k(\mathbf{w}, \tau) = w_1 + w_2\tau$ на этой локальной выборке. Вычислим дисперсию $\sigma^2_{\mathbf{r}_k}$

$$\sigma^2_{\mathbf{r}_k} = \frac{\|\mathbf{r}_k - \mathbf{1}\bar{r_k}\|^2}{m_k},$$

регрессионных остатков

$$\mathbf{r}_k = [s_{\tau_k} - f_k(\mathbf{w}, \tau_k), \ldots, s_{\tau_{k+1}} - f_k(\mathbf{w}, \tau_{k+1}),]^\intercal$$

где $\bar{r}_k$ — среднее арифметическое значение элементов вектора $\mathbf{r}_k$ и $m_k$ — число его элементов. При выполнении условия

$$\frac{\hat{\sigma}^2_{\mathbf{r}}}{\sigma^2_{\mathbf{r}_k}} < 1$$

считаем, что правая граница $k$-го сегмента $\tau_{k+1}$ найдена. В противном случае добавляем следующий по порядку элемент в локальную выборку.

При разметке временного ряда ставим в соответствие элементу $s_\tau$ значение $u_\tau = 1$, если коэффициент $w_2$ линейной модели $f_k$, которая приближает сегмент $(\tau_k, \tau_{k+1}) \ni \tau$, больше нуля. В противном случае $u_\tau = 0$. На рисунке 84 показан временной ряд и его разметка. Ломаной показаны сегменты, на которых разметка неизменна.

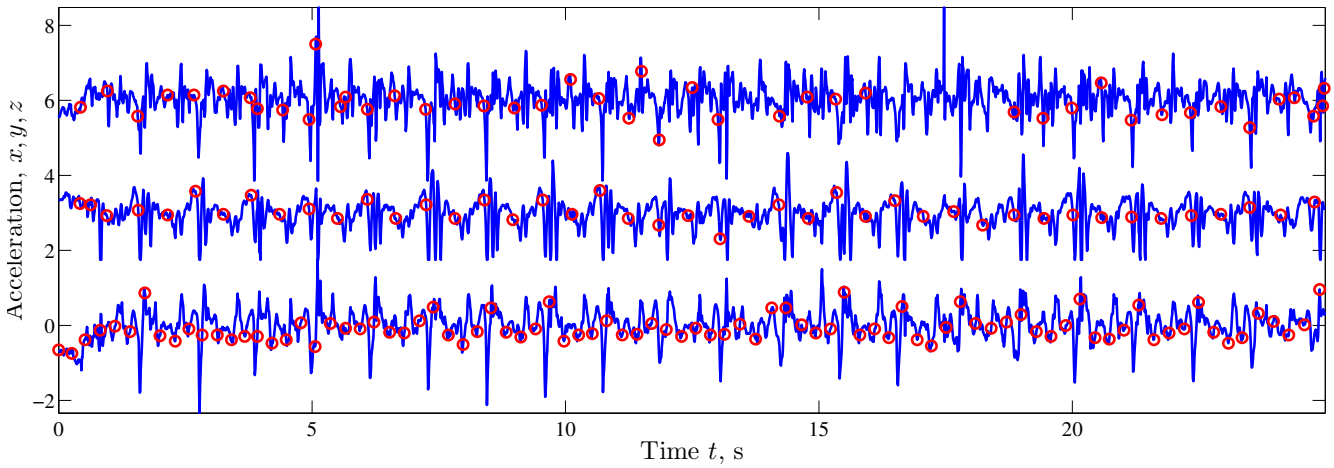



Рис. 84. Размеченный временной ряд акселерометра

### 6.2.4. Сегментация фазовой траектории

Рассмотрим набор временных рядов $\mathbf{s}_1, \ldots, \mathbf{s}_j \ldots, \mathbf{s}_\ell$, удовлетворяющих условиям (201). Значения временных рядов $s_{\tau 1}, \ldots, s_{\tau \ell}$ в моменты времени $\tau \in \{1, \ldots, T\}$ образуют ломаную фазовую траекторию или годограф в пространстве $\mathbb{R}^\ell$. Предлагается так разбить отрезок времени $(1, T)$ на сегменты, чтобы норма разности проекции локальной выборки на подпространство заданной размерности и самой локальной выборки не превосходила заданное значение $\hat{\sigma}^2_{\mathbf{r}}$.

Рассмотрим пример-иллюстрацию использования сингулярного разложения. Пусть поведение некоторой биосистемы описывается набором параметров, образующих фазовое пространство. Например, пусть $x_1, x_2$ — концентрация кислорода в крови и частота сердечных сокращений пациента. Эти параметры, изменяясь во времени, образуют траекторию его жизни. Фазовое пространство разбито на три непересекающихся области: жизни $\mathcal{A}$ — *alive*, смерти $\mathfrak{D}$ — *dead* и границу между ними $\mathcal{B}$ — *boundary*, рис. 85. Гипотеза: в точке, максимально

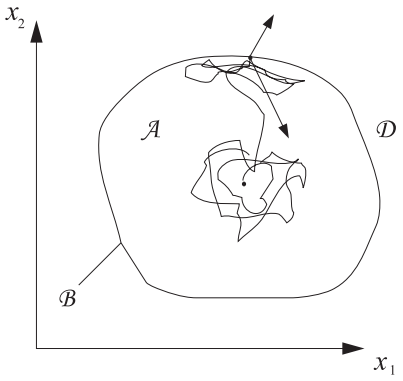



Рис. 85. Поведение биосистемы в экстремальных условиях.

удаленной от границ $\mathcal{B}$ внутри области $\mathcal{A}$ энтропия системы максимальна, в то время как у границы поведение системы становится ригидным, жестким, эффективная размерность траектории снижается.

Для нахождения границ сегментов фазовой траектории найдем сингулярное разложение матрицы $\mathbf{X}$, состоящей из соединенных векторов $\mathbf{s}_1, \ldots, \mathbf{s}_\ell$:

$$[\mathbf{s}_1, \ldots, \mathbf{s}_\ell] = \mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}, \quad \text{где} \quad \mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_\ell).$$

Под эффективной размерностью $d$ матрицы $\mathbf{X}$ будем понимать количество сингулярных чисел, превосходящих заданное $\lambda_d$. Требуется найти значения локальной выборки, лежащие в пространстве эффективной размерности $\mathbb{R}^d$ или меньшей размерности. При этом задано среднеквадратичное отклонение $\hat{\sigma}_{\mathbf{r}}$ и размерность $d$.

Найдем $k$-й сегмент. Пусть левая граница $\tau_k$ сегмента с номером $k$ известна. На первом шаге локальная выборка $\mathbf{s}_k$ содержит два элемента временного ряда, $\mathbf{s}_k = \{s_{\tau_k}, s_{\tau_k+1}\}$. Найдя сингулярное разложение, получим значение сингулярного числа $\lambda_d$. Если значение

$$\lambda_d(k) - \hat{\sigma}_{\mathbf{r}}^2 \leqslant 0,$$

то добавляем следующий элемент временного ряда в локальную выборку. В противном случае считаем правую границу $\tau_{k+1}$ найденной.

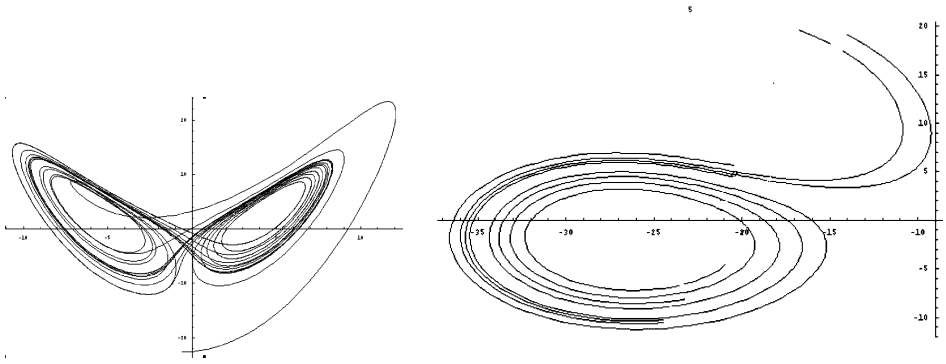


Рис. 86. Фазовая траектория системы с аттрактором Лоренца и её сегменты.

В качестве примера рассмотрим сегментацию фазовой траектории аттрактора Лоренца. Траектория задана системой обыкновенных дифференциальных уравнений

$$\begin{cases} x_1'(t) = -3(x_1(t) - x_2(t)) \\ x_2'(t) = -x_1(t)x_3(t) + 26.5x_1(t) - x_2(t) \\ x_3'(t) = x_1(t)x_2(t) - x_3(t) \end{cases} \tag{212}$$

с начальными условиями $x_1(0) = x_3(0) = 0$, $x_2(0) = 1$. Данная фазовая траектория лежит с пространстве $\mathbb{R}^3$. Временные ряды $\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$ получены путем дискретизации решения данного уравнения, а именно, значение элемента $j$-го временного ряда $s_{\tau j} = x_j(t)$, $j \in \{1, 2, 3\}$. При этом значение $\tau = \tau(t)$ принадлежит конечному множеству $\{1, \ldots, T\}$. На рис. 86 показана двумерная проекция траектории системы с аттрактором Лоренца из трехмерного пространства и одно из подмножеств ее сегментов, лежащее в двумерном пространстве.

### 6.2.5. Прогнозирование размеченных апериодических временных рядов

Пусть каждому элементу $s_\tau$ временного ряда $\mathbf{s}$ поставлен в соответствие элемент $u_\tau$ из множества меток $\mathcal{M}$ некоторого конечного алфавита. Например, множество двух ме-

ток $\mathcal{M} = \{$«ряд возрастает», «ряд не возрастает»$\}$, в дальнейшем для удобства множество $\mathcal{M} = \{0, 1\}$. Требуется по предыстории этого временного ряда и, возможно, набора дополнительных временных рядов, спрогнозировать значение временного ряда $\mathbf{u} = u_1, \ldots, u_{T-1}$ в момент времени $T$. При этом предполагается, что существует по крайней мере одна зависимость между значением метки и ее предысторией. Другими словами, не каждая предыстория заведомо влечет появление заданной метки.

Пусть временной ряд удовлетворяет тем же условиям (201), что и в предыдущем разделе с той лишь разницей, что значения $u_\tau$ прогнозируемого ряда принадлежат множеству $\mathcal{M}$, а значения $s_{ij}$ дополнительных временных рядов $\mathbf{s}_1, \ldots, \mathbf{s}_\ell$ принадлежат множеству $\mathbb{R}$. Требуется спрогнозировать значение метки $u_T$ временного ряда $\mathbf{u}$ в момент времени $T$. Для этого построим матрицу $\mathbf{X}^\dagger$ так, чтобы столбец матрицы с индексом $j$ являлся $j-1$-м временным рядом, $j \in \{2, \ldots, \ell+1\}$, а первым столбцом — временной ряд $\mathbf{u}$. Разбиение матрицы $\mathbf{X}^\dagger$ на подматрицы имеет вид

$$\mathbf{X}^\dagger = \left[\begin{array}{c|ccc} u_1 & s_{11} & \ldots & s_{1\ell} \\ \ldots & \ldots & \ddots & \ldots \\ u_{T-1} & s_{(T-1)1} & \ldots & s_{(T-1)\ell} \\ \hline u_T & & & \end{array}\right].$$

Пусть значение $u_T$ зависит от исторических значений временного ряда за последние $H$ отсчетов времени $\tau$, и пусть эта зависимость линейна. Построим выборку $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$ следующим образом. Вектор зависимых переменных $\mathbf{y}$ тождественно равен вектору $\mathbf{u}$. Строка $\mathbf{x}_\tau^\intercal$ матрицы плана $\mathbf{X}$ состоит из соединенных наборов значений временных рядов

$$\mathbf{x}_\tau^\intercal = [s_{(\tau-1)1}, \ldots, s_{(\tau-H-1)1}, \ldots, s_{(\tau-1)\ell}, \ldots, s_{(\tau-H-1)\ell}].$$

Другими словами, строка с номером $\tau$ матрицы плана $\mathbf{X}$ есть векторизованная подматрица, состоящая из значений временных рядов

$$\left[\begin{array}{ccc} s_{(\tau-H-1)1} & \ldots & s_{(\tau-H-1)\ell} \\ \ldots & \ddots & \ldots \\ s_{(\tau-2)1} & \ldots & s_{(\tau-2)\ell} \\ s_{(\tau-1)1} & \ldots & s_{(\tau-1)\ell} \end{array}\right].$$

Принимая логистическую регрессию как модель зависимости $\mathbf{y} = \boldsymbol{\mu}(\mathbf{X}\mathbf{w})$, после оценки вектора параметров $\hat{\mathbf{w}}$, получаем прогнозируемое значение

$$s_T = \mu(\mathbf{x}_m^\intercal \hat{\mathbf{w}}) = \frac{1}{1 + \exp(-\mathbf{x}_m^\intercal \hat{\mathbf{w}})}.$$

Как и в предыдущем разделе, даны

1) матрица плана $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_i^\intercal, \ldots, \mathbf{x}_{m-1}^\intercal]$, иначе $\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_n]^\intercal$, множество индексов столбцов $\mathcal{J} = \{1, \ldots, n\}$, множество индексов строк $\mathcal{I} = \{1, \ldots, m-1\}$,

2) вектор значений зависимой переменной $\mathbf{y}$, который вместе с матрицей плана образует регрессионную выборку $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$,

3) класс моделей $\{\mathbf{f}_\mathcal{A} = \boldsymbol{\mu}(\mathbf{X}_\mathcal{A}\mathbf{w}_\mathcal{A}) | \mathcal{A} \subseteq \mathcal{J}\}$,

4) гипотеза порождения данных $\mathbf{y} \sim \mathcal{B}(\mathbf{f}, \mathbf{B})$ и функция ошибки $S(\mathbf{w}_\mathcal{A}|\mathbf{f}_\mathcal{A}, \mathfrak{D})$, заданная согласно этой гипотезе.

Так как предполагается, что только часть элементов размеченного временного ряда описываются регрессионной моделью, предлагается разбить множество индексов элементов выборки на две части, прогнозируемую $\mathcal{B}$ и непрогнозируемую $\mathcal{B}^\emptyset$. Требуется найти множество индексов $\hat{A}$ столбцов и множество индексов $\hat{\mathcal{B}}$ строк матрицы плана $\mathbf{X}$ такие, что

$$(\hat{\mathcal{A}}, \hat{\mathcal{B}}) = \underset{\mathcal{A}\subseteq\mathcal{J},\, \mathcal{B}\subseteq\mathcal{I},\, |\mathcal{B}|\geqslant b}{\arg\min} S(\hat{\mathbf{w}}_\mathcal{A}|\mathbf{f}_\mathcal{A}, \mathfrak{D}_\mathcal{B}). \tag{213}$$

Как и ранее проблема нахождения модели оптимальной сложности (переобучения) решается введением соответствующей функции ошибки или методами скользящего контроля.

Аргумент $\mathfrak{D}_\mathcal{B}$ означает, что при вычислении значения функции ошибки используются только элементы выборки $\mathfrak{D}$ с индексами $\mathcal{B}$. Константа $b$ назначается экспертно. При постановке задачи считаем, что оценка $\hat{\mathbf{w}}_\mathcal{A}$ параметра модели была получена ранее согласно гипотезе порождения данных на подвыборке $\mathfrak{D}_\mathcal{B}$.

## 6.3. Кластеризация с использованием наборов парных расстояний в ранговых шкалах

Для решения задачи локального прогнозирования временных рядов требуется быстрый алгоритм кластеризации. Причем этот алгоритм должен выявлять единственный кластер на множестве объектов — подпоследовательностей, или сегментов, временных рядов. Для выявления кластеров используются парные расстояния между подпоследовательностями. Отличительной особенностью алгоритма является то, что не требуется строить полную матрицу парных расстояний, что снижает сложность вычислений. При кластеризации рассматриваются только ранги расстояний между подпоследовательностями.

Предлагаемый алгоритм был разработан в ходе решения задачи прогнозирования вторичной структуры белка по первичной [336, 337]. Предлагалось найти соответствие между «типичными» последовательностями аминокислотных остатков, кодируемых буквами двадцатибуквенного алфавита и между соответствующими им вторичными структурами, кодируемыми буквами трехбуквенного алфавита. Для этого предполагалось составить словарь часто встречающихся, «типичных» последовательностей аминокислотных остатков, то есть, решить задачу кластеризации. Особенностью задачи является то, что база данных остатков, подпоследвательности которых требуется кластеризовать, содержит 11 миллионов записей длиной 20–33000 символов каждая [4, 5]. Такой объем данных выдвигает ограничение на сложность алгоритма кластеризации; предполагается возможность параллельного запуска этого алгоритма. Ранее были предложены алгоритмы быстрой кластеризации объектов, описанных в номинальных шкалах [144, 58, 244, 254, 275].

Основная идея предложенного похода заключается в следующем. Каждая последовательность аминокислотных остатков разбивается на слова одинаковой длины. Длина слова задается до начала кластеризации и выбирается исходя из результатов анализа записей о вторичных белковых структурах. Множество полученных слов является множеством кластери-

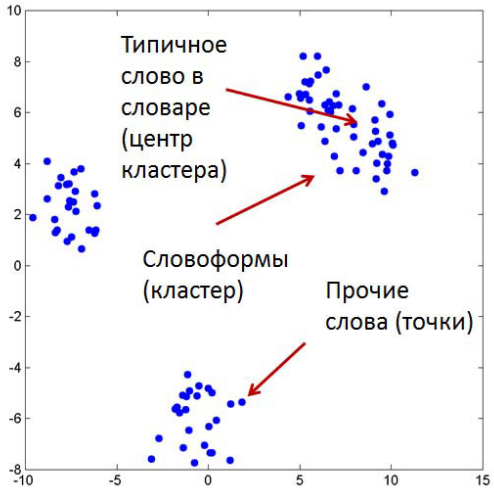



Рис. 87. Точки на плоскости, как пример последовательности аминокислотных остатков.

зуемых объектов. На этом множестве задана метрика, и далее его объекты будут называться точками, погруженными в метрическое пространство.

Вышеперечисленные алгоритмы кластеризации требуют матрицу парных расстояний между всеми парами точек из множества, что существенно повышает сложность алгоритма. Предложенный алгоритм требует только расстояния между выделенными точками, называемыми далее $\rho$-сетью и всеми остальными точками. При этом расстояния от некоторой выделенной точки до прочих ранжируются, и кластеризация выполняется по ранговым значениям. Таким образом алгоритм состоит из следующих основных шагов:

1) разбиение описаний первичных структур белков,

2) задание опорного множества ($\rho-$сети),

3) вычисление расстояния между некоторыми парами объектов,

4) нахождение метрических сгущений, кластеризация.

Далее в работе описаны метрики, используемые при кластеризации последовательностей аминокислотных остатков и их свойства, указан способ построения $\rho$-сети, описан предложенный алгоритм кластеризации и указана его сложность. Работу завершает вычислительный эксперимент, который содержит описание данных, базового алгоритма и принятого функционала качества кластеризации. Работу завершает сравнение и анализ результатов работы двух алгоритмов.

### 6.3.1. Функции расстояния между словами

Опишем способ получения множества объектов кластеризуемой выборки. Задана цепочка букв двадцатибуквенного алфавита, $x_1, \ldots, x_i, \ldots, x_p$ длиной $p$, соответствующая первичной структуре некоторого белка. Множеством объектов будем считать множество $\{x_i, \ldots, x_{i+n-1} | i = 0, \ldots, p-n-1\}$ слов заданной длины $n$. При наличии нескольких цепочек букв множества слов, полученные для каждой цепочки объединяются. Представим

каждое полученное слово в виде точки на плоскости. Такое представление позволяет погрузить $n+1$ точку в $n$-мерное пространство и, задав метрику между парами точек, найти наиболее близкие пары. Множества точек, имеющие относительно малые парные расстояния, будем называть *метрическим сгущением*.

Рассмотрим два слова: $\mathbf{x} = (x_1, ..., x_n)$ и $\mathbf{y} = (y_1, ..., y_m)$. В общем случае слова $\mathbf{x}$ и $\mathbf{y}$ могут быть разной длины. Необходимо, чтобы выбираемая нами функция расстояния между словами $\rho(\mathbf{x}, \mathbf{y})$ была метрикой. Для этого должны быть выполнены следующие условия:

1) условие тождества, $\rho(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$;

2) условие симметрии, $\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}, \mathbf{x})$;

3) неравенство треугольника $\rho(\mathbf{x}, \mathbf{z}) \leqslant \rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z})$.

**Симметрическая разность на неупорядоченных множествах.** Данная функция расстояния между словами $\mathbf{x}$ и $\mathbf{y}$ определена как

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{|\mathbf{x}| + |\mathbf{y}| - 2S(\mathbf{x}, \mathbf{y})}{|\mathbf{x}| + |\mathbf{y}| - S(\mathbf{x}, \mathbf{y})},$$

где $S(\mathbf{x}, \mathbf{y})$ — пересечение наборов $\mathbf{x}$ и $\mathbf{y}$ как неупорядоченных множеств: каждому элементу набора $\mathbf{x}$ ставится в соответствие тождественный ему элемент набора $\mathbf{y}$ без учета индексов последнего. Число полученных пар является значением функции $S$. При этом множества $X$, $Y$ считаются неупорядоченными. Знак $|\cdot|$ означает мощность множества, в данном случае — число букв в слове. Элементы слов $\mathbf{x}$ и $\mathbf{y}$ индексированы, на множестве индексов задано отношение полного порядка.

Для данного расстояния, очевидно, выполнено условие симметрии. Также выполнено неравенство треугольника. Докажем его для случая $|\mathbf{x}| = |\mathbf{y}| = |\mathbf{z}|$, потому что предложенный алгоритм будет использовать только одинаковые длины слов.

Обозначим

$$a = \frac{S(\mathbf{x}, \mathbf{y})}{|\mathbf{x}| + |\mathbf{y}|}, b = \frac{S(\mathbf{y}, \mathbf{z})}{|\mathbf{y}| + |\mathbf{z}|}, c = \frac{S(\mathbf{x}, \mathbf{z})}{|\mathbf{x}| + |\mathbf{z}|}.$$

Тогда

$$\rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z}) - \rho(\mathbf{x}, \mathbf{z}) = \frac{|\mathbf{x}| + |\mathbf{y}| - 2S(\mathbf{x}, \mathbf{y})}{|\mathbf{x}| + |\mathbf{y}| - S(\mathbf{x}, \mathbf{y})} + \frac{|\mathbf{y}| + |\mathbf{z}| - 2S(\mathbf{y}, \mathbf{z})}{|\mathbf{y}| + |\mathbf{z}| - S(\mathbf{y}, \mathbf{z})} - \frac{|\mathbf{x}| + |\mathbf{z}| - 2S(\mathbf{x}, \mathbf{z})}{|\mathbf{x}| + |\mathbf{z}| - S(\mathbf{x}, \mathbf{z})} =$$

$$= 1 - \frac{S(\mathbf{x}, \mathbf{y})}{|\mathbf{x}| + |\mathbf{y}| - S(\mathbf{x}, \mathbf{y})} - \frac{S(\mathbf{y}, \mathbf{z})}{|\mathbf{y}| + |\mathbf{z}| - S(\mathbf{y}, \mathbf{z})} + \frac{S(\mathbf{x}, \mathbf{z})}{|\mathbf{x}| + |\mathbf{z}| - S(\mathbf{x}, \mathbf{z})} =$$

$$= 1 - \frac{a}{1-a} - \frac{b}{1-b} + \frac{c}{1-c} = \frac{1 - 2a - 2b + 3ab + ac + bc - 2abc}{(1-a)(1-b)(1-c)}.$$

Чтобы неравенство треугольника выполнялось, необходимо, чтобы эта дробь была неотрицательной. Поскольку все $a, b, c \in [0; \frac{1}{2}]$, знаменатель является положительным числом. Заметим также, что для $a, b$ и $c$ выполнено соотношение $c \geqslant a + b - \frac{1}{2}$. Это так, потому что наибольшее по мощности множество букв, состоящее из объединения пересечений слов $\mathbf{x}, \mathbf{y}$ и $\mathbf{y}, \mathbf{z}$, не содержащихся в пересечении $\mathbf{x}, \mathbf{z}$, равно $|\mathbf{x}|$. Обозначим мощность этого объединения

за $u$, а мощность множества букв, состоящего из объединения пересечений слов $\mathbf{x}, \mathbf{y}$ и $\mathbf{y}, \mathbf{z}$, содержащихся в пересечении $\mathbf{x}, \mathbf{z}$, за $u'$. Тогда:

$$\frac{S(\mathbf{x},\mathbf{y})}{|\mathbf{x}|+|\mathbf{y}|}+\frac{S(\mathbf{y},\mathbf{z})}{|\mathbf{y}|+|\mathbf{z}|}=\frac{u+u'}{2|\mathbf{x}|}\leqslant\frac{1}{2}+c.$$

Поэтому числитель дроби

$$1-2a-2b+3ab+ac+bc-2abc\geqslant 1-\frac{5}{2}a-\frac{5}{2}b+6ab+a^2+b^2-2a^2b-2ab^2.$$

Заметим, что эта дробь симметрична по $a$ и $b$, поэтому для глобального минимума должно выполняться $a=b$. Симметризуя это выражение, получаем:

$$1-5a+8a^2-4a^3,$$

которое $\geqslant 0$ при $a\in[0,\frac{1}{2}]$. Значит, неравенство треугольника в этом случае выполнено.

Однако, условие тождества не выполнено, потому что данное расстояние между любой парой слов, состоящией из одинакового набора букв, равно нулю.

**Симметрическая разность на упорядоченных множествах.** Данная метрика определена как

$$\rho(\mathbf{x},\mathbf{y})=\frac{|\mathbf{x}|+|\mathbf{y}|-2G(\mathbf{x},\mathbf{y})}{|\mathbf{x}|+|\mathbf{y}|-G(\mathbf{x},\mathbf{y})},$$

где $G(\mathbf{x},\mathbf{y})$ — мощность наибольшей общей подпоследовательности символов в словах $\mathbf{x}$ и $\mathbf{y}$. Мощность пересечения двух упорядоченных наборов символов (наибольшей общей подпоследовательности) равна длине диагонального пути наименьшей стоимости, определенного в (214).

Данное расстояние является метрикой, потому что для него выполнены условие симметрии и неравенство треугольника (аналогично предыдущему случаю), а также верно условие тождества:

$$\rho(\mathbf{x},\mathbf{y})=0\Leftrightarrow G(\mathbf{x},\mathbf{y})=|\mathbf{x}|=|\mathbf{y}|,$$

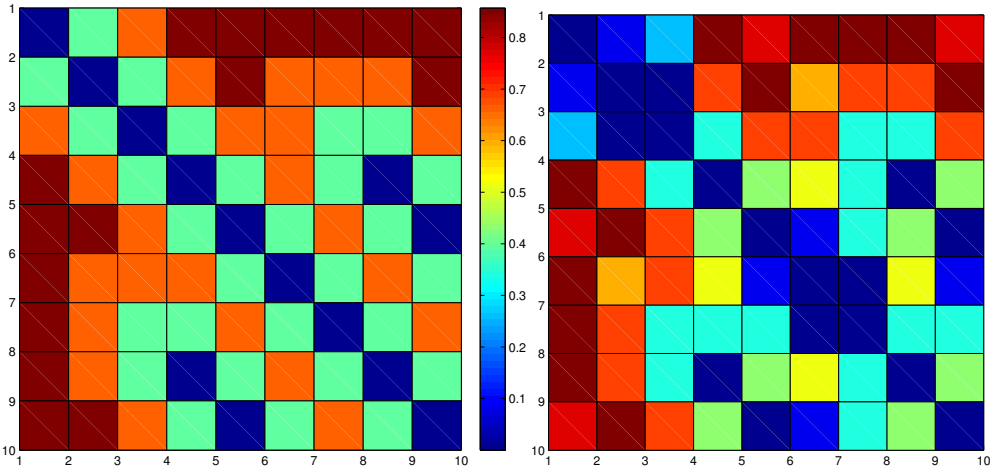
а это возможно только в том случае, когда наибольшая общая подпоследовательность совпадает со всем словом, то есть два слова тождественны. Область значения данной функции расстояния находится на отрезке $[0;1]$. На рис. 88 слева показана матрица парных расстояний $\mathfrak{D}$ для этой метрики. Каждый элемент матрицы есть значение функции расстояния для соответствующей пары слов.

**Оптимальное выравнивание.** Подсчет этого расстояния сводится к поиску оптимального выравнивания между двумя словами. Расстоянием между двумя буквами $x_i$, $y_j$ этих слов является булева функция:

$$d_{i,j}=\begin{cases}1, & \text{если } x_i\neq y_j,\\ 0, & \text{иначе.}\end{cases}$$

Для вычисления расстояния между словами составим $M(n+1\times m+1)$-матрицу стоимости. Обозначим индекс первой строки $i=0$ и индекс первого столбца $j=0$. Присвоим

$$M(0,0)=0;$$



(a) Симметрическая разность (b) Оптимальное выравнивание

Рис. 88. Матрица парных расстояний, пример.

для всех $i = 1, ..., n$ и $j = 1, ..., m$ присвоим

$$M(0, j) = M(i, 0) = \infty;$$

для всех $i = 1, ..., n$ и $j = 1, ..., m$ вычислим последовательно все элементы матрицы $M$ по формуле

$$M(i, j) = d(x_i, y_j) + \min\big(M(i-1, j-1), M(i-1, j), M(i, j-1)\big).$$

Искомым расстоянием между словами $x$ и $y$ будет последний элемент этой матрицы:

$$\rho(x, y) = M(n, m). \tag{214}$$

Стоит отметить, что данное расстояние является частным случаем расстояния Левенштейна [316], то есть, является метрикой. На рис. 88 справа показана матрица парных расстояний для случая $d(x, y) \in [0, 1]$, длина слов $m = n = 8$. На рис. 89 показана матрица стоимости $M$ алгоритма оптимального выравнивания. Путь наименьшей стоимости показан точками. Его начало и конец фиксированы в элементах с индексами (0,0) и $(n, m)$.

### 6.3.2. Описание алгоритма кластеризации $\rho-$сетью

Опишем алгоритм, позволяющий быстро кластеризовать объекты в произвольном метрическом пространстве. На рис. 90 показана симметричная относительно главной диагонали матрица парных расстояний для 40 точек. При создании нижеописанного алгоритма преследовалась цель существенно снизить сложность процедуры кластеризации относительно квадратичной, требуемой для построения матрицы парных расстояний между всеми объектами кластеризуемого множества.

Обозначим $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ — множество, состоящее из $N$ точек. Задана функция расстояния $\rho(\mathbf{x}_i, \mathbf{x}_j)$, определенная на всех парах точек из $X$, для которой выполняются условия метрики. Требуется найти множество $K \subset X$ — подмножество $X$, образующее метрическое сгущение. Сгущением называется множество близких, в смысле заданной метрики, точек, образующих компактные области. Считается, что множеству точек, образующих сгущение,

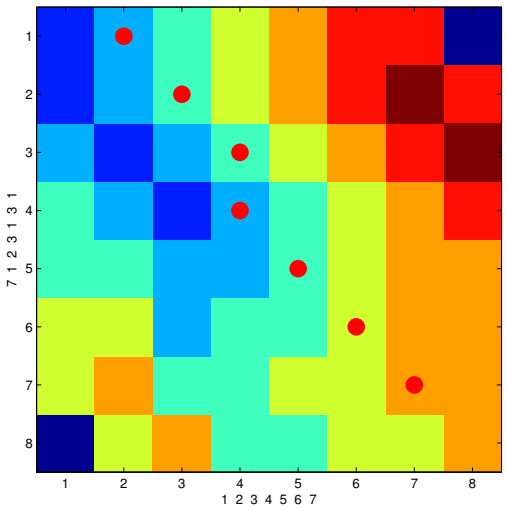



Рис. 89. Матрица стоимости оптимального выравнивания.

принадлежат все точки выпуклой комбинации этого множества. Предполагается, что будет найдена последовательность сгущений $K$ посредством итеративной процедуры следующего вида. Из заданного набора $X$ вычитаем множество точек $K$, образующих сгущение, $X^* = X \setminus K$. Находим сгущение $K^*$ на полученном наборе $X^*$. Процедура повторяется до нахождения всех сгущений $\{K\}$.

Для отыскания множества $K$ введем понятие $\rho-$сети и построим матрицу $\mathfrak{D}$ парных расстояний между точками, принадлежащими $\rho-$сети, и всеми точками множества $X$: $\mathfrak{D} = \{d_{i,j}\}$, где $i \in \{1, ..., n\} = I$ — индекс объекта $\rho-$сети, а $j \in \{1, ..., N\} = J$ — индекс объекта из $X$.

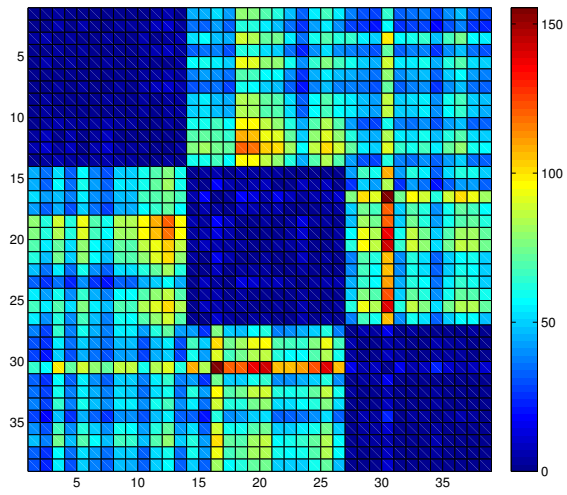



Рис. 90. Матрица парных расстояний для 40 точек.

$\rho-$сеть — это множество $X' = \{\mathbf{x}_k | k \in I\}$ фиксированной мощности $n$, собственное подмножество $X$, состоящее из объектов, которые находятся на максимальном расстоянии друг от друга, т.е.

$$I = \arg\max_{j \in J} \min_{i \in I/j} \rho(\mathbf{x}_i, \mathbf{x}_j), \qquad I \subset J.$$

Точки, входящие в $\rho-$сеть $X'$, также принадлежат множеству $X$, $X' \subset X$, причем предполагается, что $N = |X| \gg n = |X'|$. Множество точек $\rho-$сети отыскивается с помощью следующей процедуры.

### 6.3.3. Выбор точек для $\rho$-сети

Положим изначально $X' = \emptyset$ — множество точек $\rho$-сети.

1. Взять произвольный элемент $\mathbf{y} \in X$.
2. Вычислить $\mathbf{x}' = \arg\max\limits_{\mathbf{x} \in X} \rho(\mathbf{x}, \mathbf{y})$, присвоить $X' := X' \cup \mathbf{x}'$.
3. Пока $|X'| < n$: вычислить $\mathbf{x}' = \arg\max\limits_{\mathbf{x} \in X} \min\limits_{z \in X'} \rho(\mathbf{x}, z)$, присвоить $X' := X' \cup \mathbf{x}'$.

Отметим, что предложенный алгоритм имеет сложность $O(n^2 N)$, где $n^2 \ll N$, то есть линейную по числу объектов.

**Построение матрицы парных расстояний.** Построим матрицу $\mathfrak{D}$ парных расстояний между точками, принадлежащими $\rho$-сети и всеми точками из $X$: $\mathfrak{D} = \{d_{ij}\}, d_{ij} = \rho(\mathbf{x}_i, \mathbf{x}_j)$, где $i \in I$ — индекс объекта сети, а $j \in J$ — индекс объекта из $X$. Матрица $\mathfrak{D} \in \mathbb{R}_+^{n \times N}$ содержит в своих строках расстояния от каждого объекта $\rho$-сети до каждого объекта из всего множества $X$.

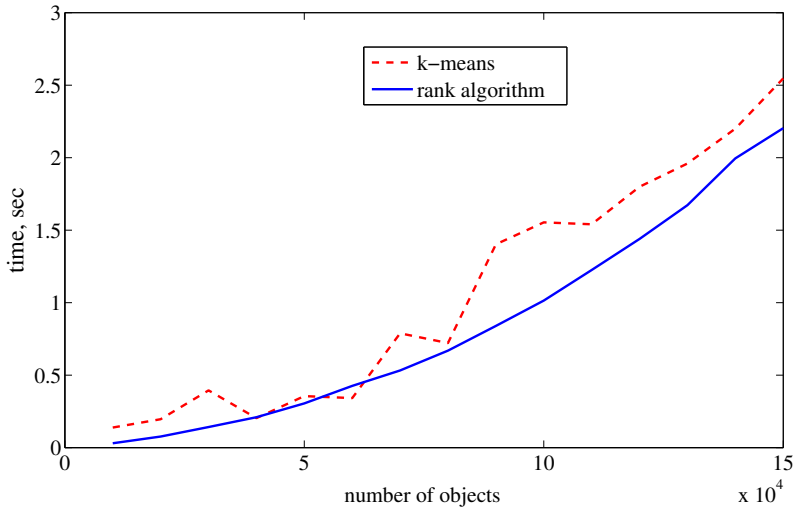



Рис. 91. Сложность алгоритма поиска сгущения относительно количества слов.

**Сортировка матрицы парных расстояний.** Для каждой строки $i$ матрицы $\mathfrak{D}$ зададим функцию $\phi_i$, которая индексам элементов строки ставит в соответствие индексы отсортированных по возрастанию расстояний от $i$-й точки $\rho$-сети до всех точек множества $X$:

$$\phi_i : \{\rho_{ij} | j \in J\} \mapsto \{\text{sort}(\rho_{ik}) | k \in J\}.$$

Функция $\phi_i$ задает преобразование $J \to J$ — биекцию

$$\phi_i : j \mapsto k, \qquad j, k \in I.$$

Построим матрицу, содержащую в строках индексы $r_{ij} \in \mathbb{N}$ отсортированных значений расстояний

$$R = \{r_{ij} | r_{ij} = \phi_i(j)\}$$

и индексы $r'_{ij} \in \mathbb{N}$ обратных относительно операции сортировки значений

$$R' = \{r'_{ik} | r'_{ik} = \phi_i^{-1}(k)\}.$$

Другими словами, матрица $R \in \mathbb{N}^{n \times N}$ содержит в своих строках индексы расстояний от $i$-й точки $\rho$-сети до $j$-й точки из множества $X$, отсортированных по возрастанию. Пример для фиксированного $i$ и $j \in \{1, 2, 3\}$ показан в табл. 23.

Таблица 23. Пример строки матрицы парных расстояний и соответствующих ранговых значений.

| Индексы точек | 1 | 2 | 3 |
|---|---|---|---|
| $\rho_{ij}$ | 0.7 | 0.3 | 0.5 |
| $\mathrm{sort}(\rho_{ij})$ | 0.3 | 0.5 | 0.7 |
| $r_{ij}$ | 3 | 1 | 2 |
| $r'_{ij}$ | 2 | 3 | 1 |

### 6.3.4. Поиск метрического сгущения

На строках матрицы $R'$ зададим окно заданной ширины, включающее $dN = \lfloor \frac{1}{2}\kappa \cdot N \rfloor$ элементов строки. Здесь $\kappa$ — задаваемый параметр, описывающий выраженность сгущения. За центр окна примем $k$-й столбец матрицы $R'$, индекс $k \in \{dN + 1, ..., N - dN - 1\}$ .

Найдем кластер $K$ с наибольшим количеством элементов, $|K| \to \max$. Для этого для каждого номера точки $j \in J$ в каждой строке с номером $i$ матрицы $R'$ найдем окрестность $K_i \subset J$, соседние элементы $j$-го столбца, мощностью $2dN + 1$. Кластером $K$ будет являться пересечение множеств $K_i$ по всем $i$:

$$K = \bigcap_{i=1}^{n} K_i.$$

Опишем процедуру поиска сгущения. Примем изначально $K := J$. Далее

1) для всех индексов точек множества $X$ $j \in J$ и для всех индексов точек $\rho$-сети $i \in I$:

2) найти ближайших соседей $K_i$ точки $\mathbf{x}_j \in X$ относительно точки $\rho-$сети $\mathbf{x}_i \in X'$:

$$K_i = \{r'_{is} : s \in \{r_{ij} - dN, ..., r_{ij} + dN\}\}. \tag{215}$$

3) Присвоить $K := K \cap K_j$.

**Примечание:** на шаге 2) алгоритма возникнет ситуация, когда

$$r_{ij} - dN < 0, \text{ или } r_{ij} + dN > N.$$

В первом случае, надо брать $s$: $s \in \{1, ..., 2dN + 1\}$, а во втором $s$: $s \in \{N - 2dN, ..., N\}$.

**Сложность алгоритма.** Предлагаемый алгоритм имеет сложность $O(n^2N)$ при построении матрицы расстояний $\mathfrak{D}$, $O(nN\log N)$ при сортировке строк матрицы $\mathfrak{D}$ и $O(nN)$ при поиске метрических сгущений. На рис. 91 показана сложность предложенного алгоритма, в сравнении со сложностью алгоритма k-Means.

**Функция ошибки кластеризации.** Для оценки качества кластеризации была введена следующая функция ошибки. Среднее внутрикластерное расстояние должно быть как можно меньше:

$$F_0 = \frac{\sum\limits_{i<j}[k_i = k_j]\rho(\mathbf{x}_i, \mathbf{x}_j)}{\sum\limits_{i<j}[k_i = k_j]} \to \min, \quad i, j \in \{i, \ldots, N\}.$$

Здесь индикаторная функция $[k_i = k_j]$ означает, что если точки с индексами $i$ и $j$ принадлежат одному и тому же кластеру с номером $k$, то возвращается единица, в противном случае — ноль. Среднее межкластерное расстояние должно быть как можно больше:

$$F_1 = \frac{\sum\limits_{i<j}[k_i \neq k_j]\rho(\mathbf{x}_i, \mathbf{x}_j)}{\sum\limits_{i<j}[\mathbf{y}_i \neq \mathbf{y}_j]} \to \max \quad i, j \in \{i, \ldots, N\}.$$

Зададим функцию ошибки кластеризации как отношение среднего внутрикластерного и среднего межкластерного расстояния:

$$Q = \frac{F_0}{F_1} \to \min.$$

Результатом работы алгоритма кластеризации должен быть кластер максимальной мощности, содержащий слова максимальной длины.

**Базовый алгоритм.** В качестве базового алгоритма, с которым сравнивался предложенный, был выбран алгоритм «$k$ средних», или $k$-Means [244, 254]. В качестве параметра алгоритм принимает на вход количество кластеров, а на первом шаге делает начальное приближение центров кластеров, которые затем итеративно пересчитывает.

Алгоритму также необходимо признаковое описание объекта: набор функций $f_j : X \to \mathbb{R}, j = 1, ..., m$, где $m$ — количество признаков. В случае точек на плоскости, признаковое описание состоит из координат точек. Алгоритм выполняется следующим образом:

1) сформировать начальное приближение центров всех кластеров: $\mu_k, k = 1, ..., K$,

2) отнести каждый объект к ближайшему центру:

$$k_i := \arg\min_{k \in K} \rho(\mathbf{x}_i, \mu_k), i = 1, ..., N,$$

3) вычислить новое положение центров:

$$\mu_{kj} := \frac{\sum\limits_{i=1}^{N}[k_i = k]f_j(\mathbf{x}_i)}{\sum\limits_{i=1}^{N}[k_i = k]},$$

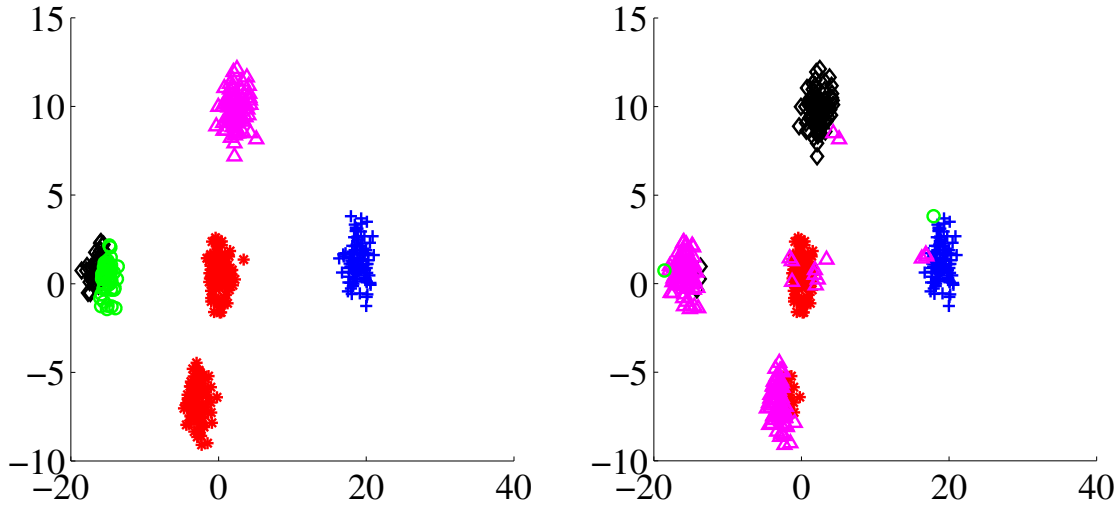



Рис. 92. Сравнение работы алгоритма $k$-Means и алгоритма ранговой кластеризации, пять кластеров.

4) повторять шаги 2, 3 пока значения $k_i$ не перестанут изменяться.

Во второй формуле используется признаковое описание точек, функция $f_j$ — $j$-е значение вектора описания точки. В случае, когда используется только матрица парных расстояний, $j$-м признаком $i$-й точки считается соответствующий элемент $i$-й строки матрицы парных расстояний $f_j(\mathbf{x}_i) = \rho(\mathbf{x}_i, \mathbf{x}_j)$, $j \in \{1, \ldots, N\}$, в которой строка — набор расстояний от этой точки до всех остальных.

**Сравнение работы двух алгоритмов кластеризации.** Приведем сначала результаты визуального сравнения на синтетической выборке, а затем опишем результаты кластеризации аминокислотных последовательностей. На рис. 92 показаны результаты кластеризации. Алгоритм k-Means, получив в качестве входного параметра число кластеров, при «неудачном» начальном приближении центров кластеров разбил один порожденный кластер на две части, а два оставшихся объединил, см. рис 92 а). Предложенный алгоритм не получал число кластеров в качестве входного параметра и выявил четыре кластера, объединив последние два, что для решения рассматриваемой прикладной задачи является корректным результатом.

На рис. 93 показаны два сгущения точек. Так как в алгоритме k-Means, в отличие от рангового, в качестве параметра задано число кластеров, то кластеры обнаружены некорректно, см. рис. 93 а). Ранговый алгоритм на рис. 93 обнаружил два кластера с удовлетворительной ошибкой.

На рис. 94 графике показаны два кластера, содержащие по 250 точек. У одного из кластеров разброс значений значительно меньше, чем у второго, и геометрически он целиком

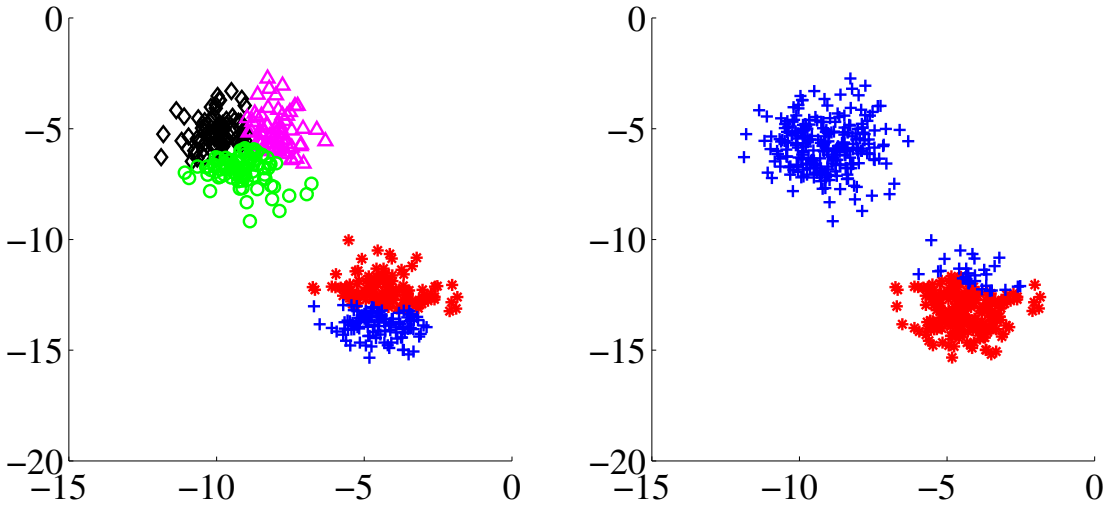



Рис. 93. Сравнение работы алгоритма $k$-Means и алгоритма ранговой кластеризации, два кластера.

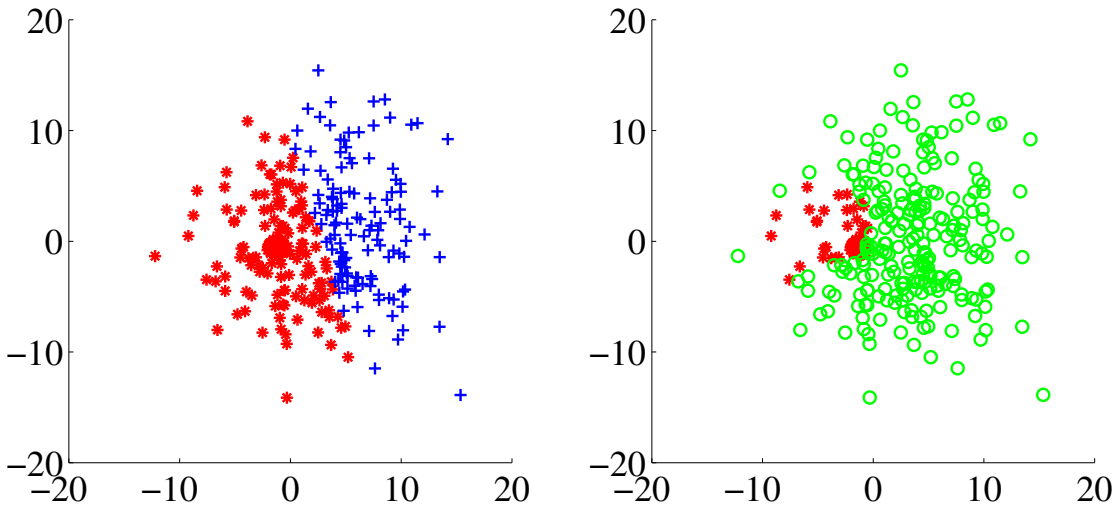



Рис. 94. Сравнение работы алгоритмов, случай вложенных кластеров.

лежит внутри второго. Алгоритм K-Means не может корректно отделить такие кластеры, это показано на рис. 94 a). В данном случае алгоритм поместил в один кластер 375, а в другой 125 точек. Алгоритм ранговой кластеризации гораздо лучше выделил сгущение, см. рис. 94 b), получив в результате 274 точки в одном кластере и 226 в другом.

Предлагаемому алгоритму кластеризации, в отличие алгоритмов кластеризации типа k-Means, не требуется признаковое описание объектов, достаточно только матрицы парных расстояний. В связи с тем, что используются только ранговые значения набора расстояний от некоторой точки до всех остальных, предложенный алгоритм нечувствителен к «небольшим» изменениям функции расстояний, что важно, если у исследователя нет точной информации о виде этой функции.

Таблица 24. Сравнение результатов работы алгоритмов на последовательностях аминокислотных остатков.

| Алгоритм | Точек в кластерах | Найдено кластеров | Качество кластеризации | Сложность алгоритма |
|---|---|---|---|---|
| k-Means | 1 | 3 | 3 | $O(N^{mK+1}\log N)$ |
| Ранговый | 1 | 3 | 3 | $O(nN\log N)$ |

Предложенный метод был использован для классификации временных рядов давления в камере внутреннего сгорания дизельного двигателя. Непосредственное вычисление значения пути оптимальной стоимости не позволило решить задачу классификации исследуемых временных рядов, так как стоимость двух несовпадающих путей временных рядов из разных классов часто оказывалась одинаковой. Создание моделей, аппроксимирующих эквивалентные временные ряды, также не позволило решить данную задачу, так как классификацию при этом приходилось выполнять в пространстве параметров, которое имело большую размерность. Предложенный метод позволил разделить данные временные ряды на кластеры, так как классификация выполняется в пространстве параметров небольшой размерности.

## 6.4. Прямая и обратная задача авторегрессионного прогнозирования

Рассмотрим *обратную задачу* макроэкономического управления, т. е. задачу определения множества таких траекторий управляемых переменных (инструментов экономической политики), которые, — при заданных ограничениях на диапазон варьирования и гладкость управляющих воздействий, — обеспечивают выход ключевых индикаторов социально-экономического развития страны (региона) на заданные уровни за определенное число тактов (кварталов, лет). В качестве модели объекта управления выберем эконометрическую модель экономики страны в форме системы одновременных уравнений (см. СОУ-модель в [13]).

### 6.4.1. Модель управления с обратной связью

Для описания *модели управления* введем следующие понятия.

*Субъект управления* — орган (лицо), принимающий решение. Другими словами, под субъектом управления мы будем понимать орган (в частном случае состоящий из одного лица), который определяет цели управления, выбирает управляющее воздействие, наблюдает за последствиями управления и оценивает результат.

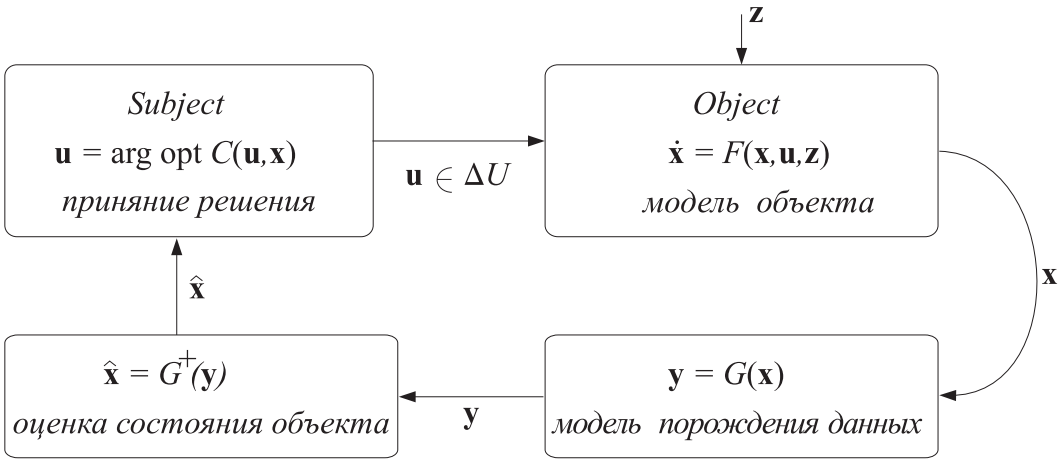



Рис. 95. Схема управления с обратной связью.

*Объект управления* — социально-экономическая система (страна, регион, см. рис. 95), эффективность функционирования которой описывается набором результирующих (эндогенных) переменных $\mathbf{y} = [y^{(1)}, ..., y^{(m)}]^\intercal$. Значения этих показателей $\mathbf{y}_t = [y_t^{(1)}, ..., y_t^{(m)}]^\intercal$ в некоторый момент времени $t$ *определяют состояние объекта управления*. Объект изменяет свое состояние под влиянием *управляющих воздействий* $\mathbf{u}_t = [u_t^{(1)}, ..., u_t^{(p)}]^\intercal$. На состояние объекта влияют также *переменные* $\mathbf{z}_t = [z_t^{(1)}, ..., z_t^{(k)}]^\intercal$ *внешней среды*, в которую он погружен. Совокупность управляемых (или частично управляемых) переменных $\mathbf{u}$ и переменных внешней среды $\mathbf{z}$ будем называть *экзогенными переменными* $X$. При отсутствии *экспертно задаваемой* целевой функции (скалярного измерителя степени эффективности функционирования системы) $\varphi(\mathbf{y})$ мы будем использовать в качестве скалярного *индекса состояния* системы *первую главную компоненту результирующих (эндогенных) переменных*.

$$q = \sum_{i=1}^{m} \theta_{1i}(y^{(i)} - \bar{y}^{(i)}). \tag{216}$$

В соотношении (216) вектор $[\theta_{11}, ..., \theta_{1m}]^\intercal$ это собственный вектор ковариационной матрицы $\Sigma_{\mathbf{y}}$ результирующих переменных $\mathbf{y}$, соответствующий *наибольшему* ее значению, а $\bar{y}^{(1)}, ..., \bar{y}^{(m)}$ — средние значения наблюдаемых результирующих переменных (усреднение — по базовому наблюдаемому периоду времени).

Пусть мы располагаем данными $(\mathbf{u}_t, \mathbf{z}_t, \mathbf{y}_t)$ о поведении объекта в течение $t_0$ тактов времени $t = 1, 2, ..., t_0$ и пусть заданы *горизонт управления* — число тактов времени $n$ и множество допустимых траекторий $\mathbf{u}(t_1, t_n)$ в $p$-мерном фазовом пространстве управляющих воздействий. Допустимость траектории определяется ограничениями на общий диапазон и гладкость варьирования переменных $u^{(1)}, ..., u^{(p)}$ на рассматриваемом отрезке времени $t = t_1, ..., t_n$, так, что элементом множества $\mathbf{u}(t_1, t_n)$ является $p$-мерная траектория $\mathbf{u}^{(t_1,t_n)} = \{[u_t^{(1)}, ..., u_t^{(p)}]^\intercal, t = t_1, ..., t_n\}$.

Пусть, наконец, мы располагаем целевыми значениями результирующих переменных $\bar{\mathbf{y}}_{t_n}$ или одного из скалярных индикаторов $\bar{\varphi}_{t_n} = \varphi(\mathbf{y}_{t_n})$ или $q_{t_n} = \sum_{j=1}^{m} \theta_{1j}(y_{t_n}^{(j)} - \bar{y}^{(j)})$, а также *допустимыми окрестностями* этих значений, — соответственно $\varepsilon(\bar{\mathbf{y}}_{t_n})$, $\varepsilon(\bar{\varphi}_{t_n})$ или $\varepsilon(\bar{q}_{t_n})$.

Тогда обратная задача сводится к определению такого подмножества $\Delta\mathbf{u}(t_1, t_n)$ множества допустимых траекторий $\mathbf{u}(t_1, t_n)$, которые при заданных (спрогнозированных) значениях переменных внешней среды $\mathbf{z}_t, t = t_1, ..., t_n$ обеспечивали бы следующие включения:

$$\begin{array}{rl|l} & \mathbf{y}_{t_n}(\mathbf{u}_{t_n}, \mathbf{z}_{t_n}) \in \varepsilon(\bar{\mathbf{y}}_{t_n}) & \text{при } (\mathbf{u}_{t_1}, ..., \mathbf{u}_{t_n}) \in \Delta\mathbf{u}_{(t_1,t_n)}. \\ \text{или} & \mathbf{y}_{t_n}(\mathbf{u}_{t_n}, \mathbf{z}_{t_n}) \in \varepsilon(\bar{\varphi}_{t_n}) & \\ \text{или} & \mathbf{y}_{t_n}(\mathbf{u}_{t_n}, \mathbf{z}_{t_n}) \in \varepsilon(\bar{q}_{t_n}) & \end{array}$$

Таким образом, для решения рассматриваемой задачи макроэкономического управления необходимо выполнение следующих этапов.

1. Выбираются субъект и объект управления. Составляется список управляющих воздействий, или альтернатив управления. Выбирается цель управления, которая в дальнейшем будет служить критерием для принятия решений. На основании отчетов о функционировании объекта управления назначаются три типа переменных:

   – управляемые переменные $\mathbf{u}$ — те переменные, которые зависят непосредственно от принятых альтернатив управления;

   – переменные внешнего воздействия на объект $\mathbf{z}$: эти переменные не зависят от управления, а определяются внешней средой, так что их значения в моменты $t_1, ..., t_n$ приходится предсказывать, поскольку от них зависит состояние объекта;

   – результирующие, определяющие состояние объекта, т. е. те переменные $\mathbf{y}$, которые характеризуют эффективность функционирования объекта.

   Определяется зависимость переменных состояния объекта $\mathbf{y}$ от экзогенных переменных $\mathbf{u}$ и $\mathbf{z}$. Эта зависимость описывается *математической моделью объекта*. Для определения связи между переменными модели выполняется тест причинно-следственной связи Грэнджера, который заключается в следующем. Каждая переменная, включаемая в модель управления либо сама влияет на состояние объекта, либо изменяется под влиянием других переменных. Если изменения переменной $a$ предшествуют изменениям переменной $b$ (при наличии статистической связи между ними), но не наоборот, то переменная $b$ зависит от переменной $a$. При выборе переменных для модели управления исключаются безразличные переменные — те, от которых не зависит состояние объекта, и которые не изменяются под влиянием других переменных. Таким образом множество переменных разбивается на подмножества управляемых, неуправляемых и зависимых переменных.

   Определяется зависимость управляемых переменных $\mathbf{u}$ от альтернатив управления. Такая зависимость называется *математической моделью субъекта*. Назначается цель управления; она может задаваться значением вектора состояния $\bar{\mathbf{y}}_{t_n}$ в заданный момент времени $t_n$, либо значениями скалярных индикаторов $\bar{\varphi}_{t_n}, \bar{q}_{t_n}$, либо их целевыми траекториями.

Результат выполнения вышеописанной процедуры показан в таблице 25.

Таблица 25. Переменные эконометрической модели экономики страны.

| | |
|---|---|
| Объект управления | экономика Российской Федерации |
| Субъект управления | правительство России |
| Цель управления | за небольшое число шагов привести показатели экономики в оптимальное состояние, определяемое индикатором |
| Альтернативы управления | `a` — принять новую программу государственных социальных расходов,<br>`b` — изменить тарификацию экспортируемых товаров. |
| Управляемые переменные | `gt` — государственные социальные расходы, млрд. руб.,<br>`tr` — средневзвешенные тарифы на экспорт, млрд. руб. |
| Неуправляемые переменные | `in` — инвестиции, млрд. руб.,<br>`oi` — цены на нефть, долл. за баррель,<br>`ex` — курс доллара США, руб.,<br>`gd` — обслуживание государственного долга, млрд. руб. |
| Переменные состояния | `y` — ВВП, млрд. руб.,<br>`x` — экспорт, млрд. долл.,<br>`p` — инфляция, % к предыдущему периоду,<br>`n` — доходы населения, млрд. руб.,<br>`m` — импорт, млрд. долл.,<br>`co` — конечное потребление, млрд. руб. |
| Модель объекта | $Y = Y(U, Z)$, где<br>$Y$ — вектор состояния, $Y = [\texttt{y,x,p,n,m,co}]^T$,<br>$U$ — вектор управления, $U = [\texttt{gt,tr}]^T$,<br>$Z$ — вектор внешнего воздействия, $Z = [\texttt{in,oi,ex,gd}]^T$ |
| Модель субъекта | $U = U(\texttt{a,b})$ строится на основе экспертных оценок влияния принимаемых альтернатив управления на управляемые переменные |
| Индикаторы состояния объекта | $\varphi, q$ — скалярные величины, характеризующие состояние объекта управления в целом |

В качестве эконометрической модели экономики, описанной в [13], предложена система *одновременных линейных уравнений*

$$\begin{aligned}
\mathtt{y} &= \mathtt{c}_{20} + \mathtt{c}_{21}\mathtt{i}_{(-4)} + \mathtt{c}_{22}\mathtt{e} - \mathtt{c}_{22}\mathtt{e}_{(-1)} + \mathtt{c}_{23}\mathtt{y}_{(-1)} + \mathtt{c}_{24}\mathtt{gd}_{(-2)} + \mathtt{c}_{25}\mathtt{dummy} + \epsilon_2,\\
\mathtt{x} &= \mathtt{c}_{10} + \mathtt{c}_{11}\mathtt{e} + \mathtt{c}_{12}\mathtt{tr} + \mathtt{c}_{13}\mathtt{o}_{(-1)} + \mathtt{c}_{14}\mathtt{y}_{(-1)} + \mathtt{c}_{15}\mathtt{x}_{(-1)} + \mathtt{c}_{16}\mathtt{dummy} + \epsilon_1,\\
\mathtt{p} &= \mathtt{c}_{30} + \mathtt{c}_{31}\mathtt{e}_{(-1)} + \mathtt{c}_{32}\mathtt{o}_{(-1)} + \mathtt{c}_{33}\mathtt{dummy} + \epsilon_3,\\
\mathtt{n} &= \mathtt{c}_{40} + \mathtt{c}_{41}\mathtt{y} + \mathtt{c}_{42}\mathtt{n}_{(-1)} + \mathtt{c}_{43}\mathtt{gt} + \mathtt{c}_{44}\mathtt{dummy} + \epsilon_4,\\
\mathtt{m} &= \mathtt{c}_{50} + \mathtt{c}_{51}\mathtt{p} + \mathtt{c}_{52}\mathtt{y} + \mathtt{c}_{53}\mathtt{m}_{(-1)} + \mathtt{c}_{54}\mathtt{x} + \mathtt{c}_{55}\mathtt{dummy} + \epsilon_5,\\
\mathtt{co} &= \mathtt{c}_{60} + \mathtt{c}_{61}\mathtt{p} + \mathtt{c}_{62}\mathtt{y} + \mathtt{c}_{63}\mathtt{m} + \mathtt{c}_{64}\mathtt{n} + \mathtt{c}_{65}\mathtt{co}_{(-1)} + \mathtt{c}_{66}\mathtt{dummy} + \epsilon_6,
\end{aligned} \tag{217}$$

где $\mathtt{c}_{ij}$ — параметры модели. Добавочная переменная $\mathtt{dummy} \in \{0, 1\}$ отражает состояние экономики до и после сентября 1998г., $\epsilon$ — авторегрессионный остаток.

Коэффициенты $\mathtt{c}_{10}, ..., \mathtt{c}_{66}$ модели оцениваются в результате оптимизации функции ошибки, включающей квадрат регрессионных остатков. Результатом является идентифицированная модель, с помощью которой прогнозируется состояние объекта.

Очевидно, что экзогенные переменные $\mathtt{x}, \mathtt{y}, \mathtt{p}, \mathtt{n}, \mathtt{m}, \mathtt{co}$ влияют на эндогенные переменные $\mathtt{i}, \mathtt{o}, \mathtt{ex}, \mathtt{gt}, \mathtt{gd}, \mathtt{tr}$ в большей или меньшей степени, что определяется коэффициентами $\mathtt{c}$. Не исключено и нулевое влияние некоторых переменных на другие.

Разделим экзогенные переменные на управляемые и переменные внешнего воздействия. Тогда модель прогноза на один квартал при заданном управлении $\mathtt{gt}, \mathtt{tr}$ и предполагаемом сценарном внешнем воздействии $\mathtt{i}, \mathtt{o}, \mathtt{ex}, \mathtt{gd}$ может быть представлена как

$$\begin{pmatrix} \mathtt{x} \\ \mathtt{y} \\ \mathtt{p} \\ \mathtt{n} \\ \mathtt{m} \\ \mathtt{co} \end{pmatrix} = \mathtt{C}_1 \begin{pmatrix} \mathtt{gt} \\ \mathtt{tr} \end{pmatrix} + \mathtt{C}_2 \begin{pmatrix} \mathtt{in} \\ \mathtt{oi} \\ \mathtt{ex} \\ \mathtt{gd} \end{pmatrix}, \tag{218}$$

где $\mathtt{C}_1, \mathtt{C}_2$ — матрицы коэффициентов, вычисляемые для заданного времени прогноза с помощью выражения (217).

### 6.4.2. Векторная авторегрессионная модель

Рекурсивная форма векторной авторегрессионной модели имеет вид

$$\mathbf{y}_t = \sum_{\tau=0}^{r} (\mathbf{A}_\tau \mathbf{y}_{t-\tau} + \mathbf{B}_\tau \mathbf{u}_{t-\tau} + \mathbf{C}_\tau \mathbf{z}_{t-\tau}) + \mathbf{m} + \boldsymbol{\varepsilon}_t. \tag{219}$$

Здесь вектор управляющих воздействий $\mathbf{u}^\intercal$ и присоединенный к нему справа вектор внешних воздействий $\mathbf{z}^\intercal$ образуют транспонированный вектор экзогенных переменных, а матрица коэффициентов $\mathbf{B}$ и присоединенная к ней справа матрица $\mathbf{C}$ образуют матрицу коэффициентов, на которую вектор экзогенных переменных умножается слева.

В выражении (219) переменная $t$ — дискретное время $t = 1, ..., t_0$, $t_0$ — последний наблюдаемый такт времени. Переменная $\tau$ обозначает глубину лагирования, причем $\tau = 1, ..., r < t_0$. Также переменная $\mathbf{m}$ есть регрессионное среднее и $\boldsymbol{\varepsilon}_t$ — регрессионный остаток, в общем

различный в каждый момент времени. Так как состояние $\mathbf{y}$ объекта управления описано $m$ переменными, а управляющие $\mathbf{u}$ и неуправляемые $\mathbf{z}$ внешние воздействия описаны соответственно $q$ и $k$ переменными, то матрицы $\mathbf{A} \in \mathbb{R}^{m\times m}, \mathbf{B} \in \mathbb{R}^{m\times p}, \mathbf{C} \in \mathbb{R}^{m\times k}$ и векторы $\mathbf{y}, \mathbf{m}, \boldsymbol{\varepsilon} \in \mathbb{R}^m, \mathbf{u} \in \mathbb{R}^p, \mathbf{z} \in \mathbb{R}^k$.

Соответствие коэффициентов системы одновременных линейных уравнений (217), описанных в эконометрической модели и элементов матриц $\mathbf{A}, \mathbf{B}$ и $\mathbf{C}$ модели векторной авторегрессии (218) показано в таблице 26.

В первом столбце таблицы показаны значения лаговой переменный $\tau$, которые соответствуют матрицам коэффициентов напротив. Левая часть уравнения (219) — вектор $\mathbf{y}$ показан в верхней строке таблицы. Он получается путем суммирования всех матриц, транспонированных и умноженных слева на соответствующие векторы, которые показаны в правом столбце таблицы, а также транспонированного вектора регрессионного среднего $\mathbf{m}$ (нижняя строка таблицы) и вектора авторегрессионного остатка $\boldsymbol{\varepsilon}_t$ (в таблице не показан). Из таблицы видно, что заполняемость ненулевыми коэффициентами невысока. В частности, все элементы матриц $\mathbf{A}_2, \mathbf{A}_3, \mathbf{A}_4, \mathbf{C}_3$ равны нулю, а матрицы $\mathbf{C}_2, \mathbf{C}_4$ имеют только по одному ненулевому элементу. В данной работе обсуждается только этот способ идентификации моделей (217) и (219), поэтому анализ заполняемости матриц $\mathbf{A}, \mathbf{B}, \mathbf{C}$ и нахождение оптимальной глубины лагирования $\tau$ останется за рамками статьи.

Представим выражение (219) в приведенной форме. Для этого перенесем вектор $\mathbf{y}_t$ состояния объекта управления в левую часть и получим выражение

$$\mathbf{I}\mathbf{y}_t - \mathbf{A}_0\mathbf{y}_t = \mathbf{B}_0\mathbf{u}_t + \mathbf{C}_0\mathbf{z}_t + \sum_{\tau=1}^{r}(\mathbf{A}_\tau\mathbf{y}_{t-\tau} + \mathbf{B}_\tau\mathbf{u}_{t-\tau} + \mathbf{C}_\tau\mathbf{z}_{t-\tau}) + \mathbf{m} + \boldsymbol{\varepsilon}_t,$$

здесь $\mathbf{I}$ — единичная матрица. Матрица линейного оператора $\mathbf{A} : \mathbf{y} \longrightarrow \mathbf{y}$ квадратная, диагональная вследствие предлагаемой эконометрической модели; следовательно, матрица $(\mathbf{I}-\mathbf{A}_0)$ не вырождена. Найдем обратную матрицу $(\mathbf{I} - \mathbf{A}_0)^{-1}$ и получим выражение

$$\mathbf{y}_t = (\mathbf{I} - \mathbf{A}_0)^{-1}\left(\mathbf{B}_0\mathbf{u}_t + \mathbf{C}_0\mathbf{z}_t + \sum_{\tau=1}^{r}(\mathbf{A}_\tau\mathbf{y}_{t-\tau} + \mathbf{B}_\tau\mathbf{u}_{t-\tau} + \mathbf{C}_\tau\mathbf{z}_{t-\tau}) + \mathbf{m} + \boldsymbol{\varepsilon}_t\right). \qquad (220)$$

Пусть известно состояние $\mathbf{y}_t$ объекта управления и внешние воздействия $\mathbf{u}_t, \mathbf{z}_t$ в течение времени $t = 1, ..., t_0$. Чтобы спрогнозировать состояние объекта управления для момента времени $t = t_1$ необходимо подставить в выражение (220) значения векторов измерений экзогенных переменных $\mathbf{u}_t, \mathbf{z}_t$ в моменты времени $t = t_0, t_0 - 1..., t_0 - r$, вектора $\mathbf{y}_t$ измерений эндогенных переменных в моменты времени $t = t_0 - 1..., t_0 - r$, а также значения матриц коэффициентов $\mathbf{A}_\tau, \mathbf{B}_\tau, \mathbf{C}_\tau$, где $\tau$=0, ..., $r$.

Таблица 26. Соответствие коэффициентов авторегрессионной модели экономики и элементов матриц модели векторной авторегрессии.

| $\tau=$ | | $Y^{\tau}=$ | y | x | p | n | m | co | $Y=$ |
|---|---|---|---|---|---|---|---|---|---|
| 0 | $A^{\tau}=$ | $A_0{}^{\tau}=$ | 0 | 0 | 0 | c41 | c52 | c62 | y |
| | | | 0 | 0 | 0 | 0 | c54 | 0 | x |
| | | | 0 | 0 | 0 | 0 | c51 | c61 | p |
| | | | 0 | 0 | 0 | 0 | 0 | c64 | n |
| | | | 0 | 0 | 0 | 0 | 0 | c63 | m |
| | | | 0 | 0 | 0 | 0 | 0 | 0 | co |
| 1 | | $A_1{}^{\tau}=$ | c23 | c14 | 0 | 0 | 0 | 0 | y |
| | | | 0 | c15 | 0 | 0 | 0 | 0 | x |
| | | | 0 | 0 | 0 | 0 | 0 | 0 | p |
| | | | 0 | 0 | 0 | c42 | 0 | 0 | n |
| | | | 0 | 0 | 0 | 0 | c53 | 0 | m |
| | | | 0 | 0 | 0 | 0 | 0 | c65 | co |
| 2 | | $A_2{}^{\tau}=$ | 0 | | | | | | |
| 3 | | $A_3{}^{\tau}=$ | 0 | | | | | | |
| 4 | | $A_4{}^{\tau}=$ | 0 | | | | | | |
| | $B^{\tau}=$ | $B_0{}^{\tau}=$ | 0 | 0 | 0 | **c43** | 0 | 0 | gt |
| | | | 0 | **c12** | 0 | 0 | 0 | 0 | tr |
| 0 | $C^{\tau}=$ | $C_0{}^{\tau}=$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | | | 0 | 0 | 0 | 0 | 0 | 0 | o |
| | | | c22 | c11 | 0 | 0 | 0 | 0 | e |
| | | | 0 | 0 | 0 | 0 | 0 | 0 | gd |
| | | | c25 | c16 | c33 | c44 | c55 | c66 | dummy |
| 1 | | $C_1{}^{\tau}=$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | | | 0 | c13 | c32 | 0 | 0 | 0 | o |
| | | | c22 | 0 | c31 | 0 | 0 | 0 | e |
| | | | 0 | 0 | 0 | 0 | 0 | 0 | gd |
| 2 | | $C_2{}^{\tau}=$ | c24 | 0 | 0 | 0 | 0 | 0 | gd |
| 3 | | $C_3{}^{\tau}=$ | 0 | | | | | | |
| 4 | | $C_4{}^{\tau}=$ | c21 | 0 | 0 | 0 | 0 | 0 | 1 |
| | | $M=$ | c20 | c10 | c30 | c40 | c50 | c60 | |

### 6.4.3. Модель субъекта управления

Модель субъекта определяет связь между списком альтернатив принимаемых решений и переменных, которые управляют субъектом. Для заданных элементов $a$ из множества альтернатив $\mathcal{A} = \{a\}$ определяются значения управляемых переменных, $\mathbf{u} = \mathbf{u}(\mathcal{A})$. Принятие той или иной управляющей альтернативы определяет состояние субъекта управления и индикатора состояния объекта. И наоборот, задавая индикатор состояния или переменные состояния объекта мы определяем значения управляемых переменных и находим ближайшую соответствующую этим значениям альтернативу.

Согласно вышесказанному, цель управления объектом может быть задана двумя способами. Первый способ: лицо, принимающее решение указывает, какие показатели объекта должны быть получены в результате управления. Второй способ: лицо принимающее решение указывает, какое оптимальное значение индикатора состояния объекта должно быть достигнуто.

### 6.4.4. Нахождение оптимального управляющего воздействия

При моделировании систем управления различают две задачи: *прямую и обратную*. Прямая задача заключается в нахождении состояния объекта управления при заданных управляющих воздействиях, см. (220). Обратная задача заключается в нахождении управляющих воздействий, которые требуются для достижения заданного состояния объекта при некоторых условиях, которые будут описаны ниже.

Прямая задача нахождения состояния $\mathbf{y}_t$ объекта управления по экзогенным переменным $\mathbf{u}_t, \mathbf{z}_t$, согласно эконометрической модели (217) решается посредством выражения (220). Для решения задачи управления, то есть, нахождения таких управляющих воздействий $\mathbf{u}$, которые бы привели объект управления в заданное состояние $\bar{\mathbf{y}}$, рассмотрим зависимость состояния $\mathbf{y}_t$ от управляющих воздействий $\mathbf{u}_t, ..., \mathbf{u}_{t-r}$. Для этого выберем из множества элементов $\{u^{(1)}_{t,\tau}, ..., u^{(k)}_{t,\tau}, t = t_0, \tau = 0, ..., r\}$ векторов $\mathbf{u}_{t-\tau}$, такие элементы $u^{*(j)}$, составляющие вектор управления $\mathbf{u}_t = [u^{*(1)}, ..., u^{*(k)}]^\intercal$ что для $i = 1, ..., p$ и $j = 1, ..., k$ выполняется условие

$$b_{ij,\tau} \neq 0, \tau = \min(0, ..., r),$$

где $\mathbf{B}_\tau = \{b_{ij,\tau}\}$. Другими словами выберем такие элементы вектора управляющих воздействий, которые для данного прогнозируемого состояния в момент времени $t$ являются существенными, имеют ненулевые коэффициенты. При этом необходимо учитывать, что управляющее воздействие было последним по времени относительно состояния объекта управления. Например, в таблице 26 эти ненулевые коэффициенты `c12` и `c43` выделены. Также рассмотрим в качестве примера влияние управляемых переменных `gt` и `tr` на вектор `y` состояния объекта управления. Для этого используем коэффициенты, описанные в табл. 26. На рис. 96 показано прямое и косвенное влияние управляющего воздействия, выраженное значениями коэффициентов `c`.

Подставляя в выражение (220) значения векторов фазовых траекторий $(\mathbf{y}_{t_0-1}, ..., \mathbf{y}_{t_0-r})$, $(\mathbf{z}_{t_0-1}, ..., \mathbf{z}_{t_0-r})$ и $(\mathbf{u}_{t_0-1}, ..., \mathbf{u}_{t_0-r})$ за исключением элементов вектора $\mathbf{u}_t$ и упрощая это выражение, получаем

$$\mathbf{y}_t = \mathbf{G}_r \mathbf{u}_t + \mathbf{h}_{t,r}, \tag{221}$$

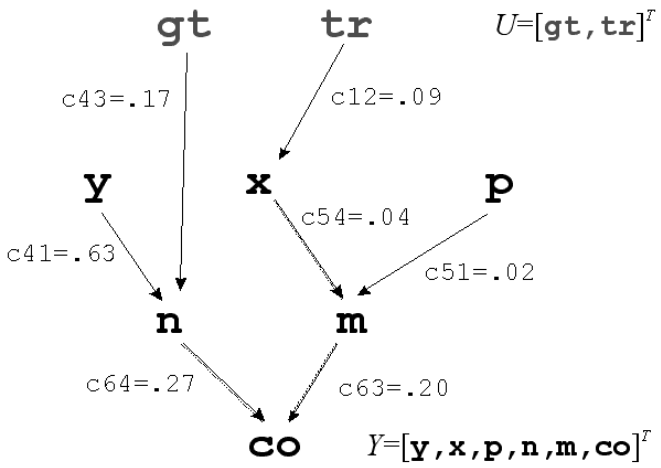



Рис. 96. Связи между управляемыми переменными и переменными состояния.

где $\mathbf{G}_r \in \mathbb{R}^{p\times k}$ — новая матрица коэффициентов для управляемых переменных $\mathbf{u}_t$, значение которой вычисляется для заданного $r$ и $\mathbf{h}_{t,r} \in \mathbb{R}^m$ — вектор, вычисляемый для заданного момента времени по известным значениям фазовых траекторий.

Уравнение обратной задачи

$$\mathbf{u}_t = \mathbf{G}_r^+(\mathbf{y}_t - \mathbf{h}_{t,r}) \tag{222}$$

получается путем псевдообращения оператора $\mathbf{G}$. Так как $\mathbf{G} \in \mathbb{R}^{m\times p}$, то псевдообратная матрица $\mathbf{G}^+ \in \mathbb{R}^{p\times m}$ при выполнении условия $\mathbf{G}^+\mathbf{G} = \mathbf{I}_p$. Отметим, что полученная модель является эконометрической по определению, см. с. 612 в [290], так как для решения обратной задачи необходимо с помощью измеряемых данных настраивать модель управления в каждый момент времени.

Для псевдообращения используется сингулярное разложение матрицы $\mathbf{G} = \mathbf{W}\mathbf{\Lambda}\mathbf{V}^\intercal$. Так как $\mathbf{W}$ и $\mathbf{V}$ являются ортогональными матрицами, а $\mathbf{\Lambda}$ — диагональная матрица, то справедливо равенство $\mathbf{G}^+ = \mathbf{V}^\intercal\mathbf{\Lambda}^{-1}\mathbf{W}$, причем $\mathbf{G}^+\mathbf{G} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{W}^\intercal\mathbf{W}\mathbf{\Lambda}\mathbf{V}^\intercal = \mathbf{I}_k$, см. [341].

*Задача управления* в данной работе ставится следующим образом. Требуется подобрать такую последовательность управляющих воздействий $(\mathbf{u}_{t_1}, ..., \mathbf{u}_{t_n})$, при ограничениях на управление $\mathbf{u}_t \in \Delta\mathbf{u}_t$, которая бы при некотором заданном сценарии внешних воздействий обеспечивала бы через $n$ шагов состояние $\bar{\mathbf{y}}_{t_n} \in \Delta\mathbf{u}_{t_n}$. В рамках данной задачи определим две: задачу *наискорейшего приближения* к целевому состоянию и задачу *оптимальноого управления*.

Задача наискорейшего приближения не является оптимальной в том смысле, что для ее решения не назначается функция общей стоимости управления; требуется подобрать такие векторы управления $(\mathbf{u}_{t_0}, ..., \mathbf{u}_{t_n})$ при ограничениях $\mathbf{u}_t \in \Delta\mathbf{u}_t$, которые бы минимизировали расстояние между целевым вектором $\bar{\mathbf{y}}_{t_n}$ и вектором текущего состояния $\mathbf{y}_t$ на каждом шаге.

Для этого на каждом шаге, начиная с $t_0$, отыскивается такое новое состояние $\mathbf{y}_{t+1} = \alpha\bar{\mathbf{y}}_{t_n} + (1-\alpha)\mathbf{y}_t$ объекта управления, что

$$\alpha = \arg\min_{\mathbf{u}_{t+1}\in\Delta\mathbf{u}_{t+1}} ||\bar{\mathbf{y}}_{t_n} - \mathbf{y}_{t+1}||^2,$$

где параметр $\alpha \in [0, 1]$. Данный алгоритм стремится достичь заданное состояние «любой ценой», независимо от характера заданных внешних воздействий $(\mathbf{z}_{t_1}, ..., \mathbf{z}_{t_n})$.

В задаче оптимального управления, как и в предыдущей, заданы сценарий внешних воздействий $(\mathbf{z}_{t_1}, ..., \mathbf{z}_{t_n})$, ограничения $\Delta\mathbf{u}_t$ на управляющие воздействия $\mathbf{u}_t$ и целевой вектор $\bar{\mathbf{y}}_n$. Требуется найти такую последовательность векторов $(\mathbf{u}_{t_1}, ..., \mathbf{u}_{t_n})$, при ограничениях $\mathbf{u}_t \in \Delta\mathbf{u}_t$, которая приводила бы объект управления из начального состояния $\mathbf{y}_{t_0}$ в целевое состояние $\bar{\mathbf{y}}_{t_n}$ за $n$ шагов при минимальной стоимости управления $F(\mathbf{u}_{t_1}, ..., \mathbf{u}_{t_n}) \to \min$.

В основу процедуры оптимизации мы положим принцип оптимальности Р. Беллмана: любое оптимальное управление может быть образовано только оптимальными управляющими воздействиями на каждом шаге. Иначе, при любом состоянии системы перед очередным шагом необходимо выбирать управление так, чтобы стоимость управления на данном шаге и стоимость управления на всех последующих шагах была минимальной.

Для решения задачи используем теорему оптимальности Л. Миттена. Приведем ее в принятых нами обозначениях.

**Определение 22.** *Функция* $F$ строго разложима, *если* $F$ *представима в виде*

$$F(\mathbf{u}_1, \mathbf{u}_2) = f_1\big(\mathbf{u}_1, f_2(\mathbf{u}_2)\big)$$

*и если* $f_1$ *— монотонная функция по своему второму аргументу. Общий класс разложимых функций образован функциями вида*

$$F(\mathbf{u}_1, .., \mathbf{u}_n) = f_1(\mathbf{u}_1) \circ f_2(\mathbf{u}_2) \circ ... \circ f_n(\mathbf{u}_n).$$

**Теорема 18.** *Пусть* $F$ *— вещественная функция от* $\mathbf{u}_1$ *и* $\mathbf{u}_2$. *Если* $F$ *разложима и* $F(\mathbf{u}_1, \mathbf{u}_2) = f_1\big(\mathbf{u}_1, f_2(\mathbf{u}_2)\big)$, *то тогда*

$$\min_{\mathbf{u}_1, \mathbf{u}_2} F(\mathbf{u}_1, \mathbf{u}_2) = \min_{\mathbf{u}_1} \left( f_1 \left[ \mathbf{u}_1, \min_{\mathbf{u}_2} \left( f_2(\mathbf{u}_2) \right) \right] \right).$$

Для упрощения индексных обозначений будем считать, что начальное состояние объекта управления $t_0 = 0$, а конечное состояние — $t_n = n$ Объект управления в момент времени $t$ описывается вектором $\mathbf{y}_t$. В моменты времени $t_1, ..., t_n$ к объекту применяются управляющие воздействия $\mathbf{u}_1, ..., \mathbf{u}_n$. Поведение объекта будем описывать *функциями перехода* $h_1, ..., h_n$, где для $t = 1, ..., n$ вектор $\mathbf{y}_t = h_t(\mathbf{u}_t, \mathbf{y}_{t-1})$ есть результат применения к объекту управляющего воздействия $\mathbf{u}_t$. В данном случае функция перехода $h_t$ соответствует модели $\mathbf{y}_t = \mathbf{G}_{t,r}\mathbf{u}_t + \mathbf{h}_{t,r}$, где $\mathbf{h}_{t,r}$ зависит от $\mathbf{y}_{t-1}$. Каждому управляющему воздействию $\mathbf{u}_t$ соответствует стоимость $f_t = f(\mathbf{u}_t, \mathbf{y}_{t-1})$.

Из состояния $\mathbf{y}_0$ в момент времени $t_0$ мы хотим привести объект управления в целевую область $\Delta\bar{\mathbf{y}}_n \ni \bar{\mathbf{y}}_n = g(\mathbf{u}_1, ..., \mathbf{u}_n)$ минимизируя при этом полную стоимость

$$F^* = \min_{\mathbf{u}_1 \in \Delta\mathbf{u}_1, ..., \mathbf{u}_n \in \Delta\mathbf{u}_n} F(\mathbf{u}_1, ..., \mathbf{u}_n).$$

Для множества управлений $\{\mathbf{u}_1, ..., \mathbf{u}_n\}$ конечное состояние системы $g(\mathbf{u}_1, ..., \mathbf{u}_n)$ определяется равенствами

$$\begin{aligned}
&\mathbf{y}_1 = h_1(\mathbf{u}_1, \mathbf{y}_0),\\
&\cdots\cdots\cdots\cdots\\
&\mathbf{y}_n = h_n(\mathbf{u}_n, \mathbf{y}_{n-1}),\\
&\mathbf{y}_n = g(\mathbf{u}_1, ..., \mathbf{u}_n).
\end{aligned}$$

Полная стоимость управления равна $F(\mathbf{u}_1, ..., \mathbf{u}_n)$ и определяется соотношением

$$F(\mathbf{u}_1, ..., \mathbf{u}_n) = f_1(\mathbf{u}_1, \mathbf{y}_0) + f_2(\mathbf{u}_2, \mathbf{y}_1) + ... + f_n(\mathbf{u}_n, \mathbf{y}_{n-1}).$$

В нашем случае функция полной стоимости $F$ является аддитивной и очевидным образом строго разложима, что дает возможность применить к задаче алгоритм *динамического программирования* [322].

Алгоритм нахождения оптимального управления заключается в следующем. Под действием управления $\mathbf{u}_t$ объект принимает состояние $\mathbf{y}_t = h_t(\mathbf{u}_t, \mathbf{y}_{t-1}) = \mathbf{G}_r\mathbf{u}_t + \mathbf{h}_{t,r}$, причем стоимость управления на каждом шаге определяется как $f(\mathbf{u}_t, \mathbf{y}_{t-1})$.

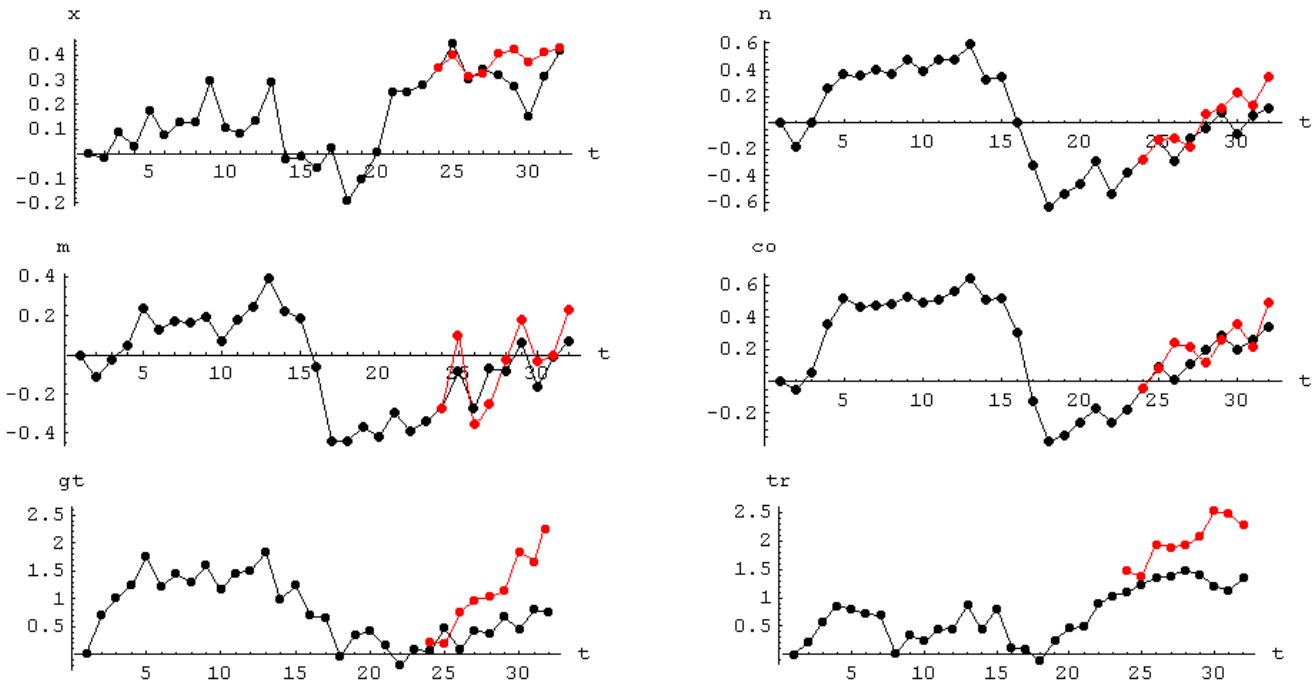



Рис. 97. Прогноз изменения состояния объекта при оптимальном управлении.

Рекуррентное уравнение динамического программирования выражает стоимость $F_t(\mathbf{y})$ условного оптимального управления начиная с $t$-го шага до конца через уже известную функцию $F_{t+1}(\mathbf{y})$:

$$F_t(\mathbf{y}) = \min_{\mathbf{u}_t \in \Delta\mathbf{u}_t} [f(\mathbf{u}_t, \mathbf{y}_{t-1}) + F_{t+1}(h_t(\mathbf{u}_t, \mathbf{y}_{t-1}))].$$

Этой стоимости соответствует условное оптимальное управление $\mathbf{u}_t$ на шаге $t$.

Далее производится условная оптимизация последнего шага $n$ для множества состояний $\mathbf{y}_{n-1}$ таких, что $\mathbf{y}_n = h(\mathbf{u}_n, \mathbf{y}_{n-1})$ при $\mathbf{u}_n \in \Delta\mathbf{u}_n$ и вычисляется условная стоимость

$$F_n(\mathbf{y}_t) = \min_{\mathbf{u}_n \in \Delta\mathbf{u}_n} f(\mathbf{u}_n, \mathbf{y}_{n-1})$$

и находится оптимальное управление $\mathbf{u}_n$.

После этого производится условная оптимизация для всех $t$, $n - 1 > t > 0$. Так как начальное состояние $\mathbf{y}_0$ известно, то искомая величина $F^* = F(\mathbf{u}_1, ..., \mathbf{u}_n) = F_1(\mathbf{y}_0)$.

Из вышеописанной процедуры оптимизации следует, что целевое множество множество $\Delta\mathbf{y}_0$ достижимо, если найдутся такие векторы $\mathbf{y}_1, ..., \mathbf{y}_n$ состояния объекта, что для всех $t = 1, ..., n$ существует управление $\mathbf{u}_t = \mathbf{G}^+_{t,r}(\mathbf{y}_t - \mathbf{h}_{t,r})$, лежащее в $\Delta\mathbf{u}_t$.

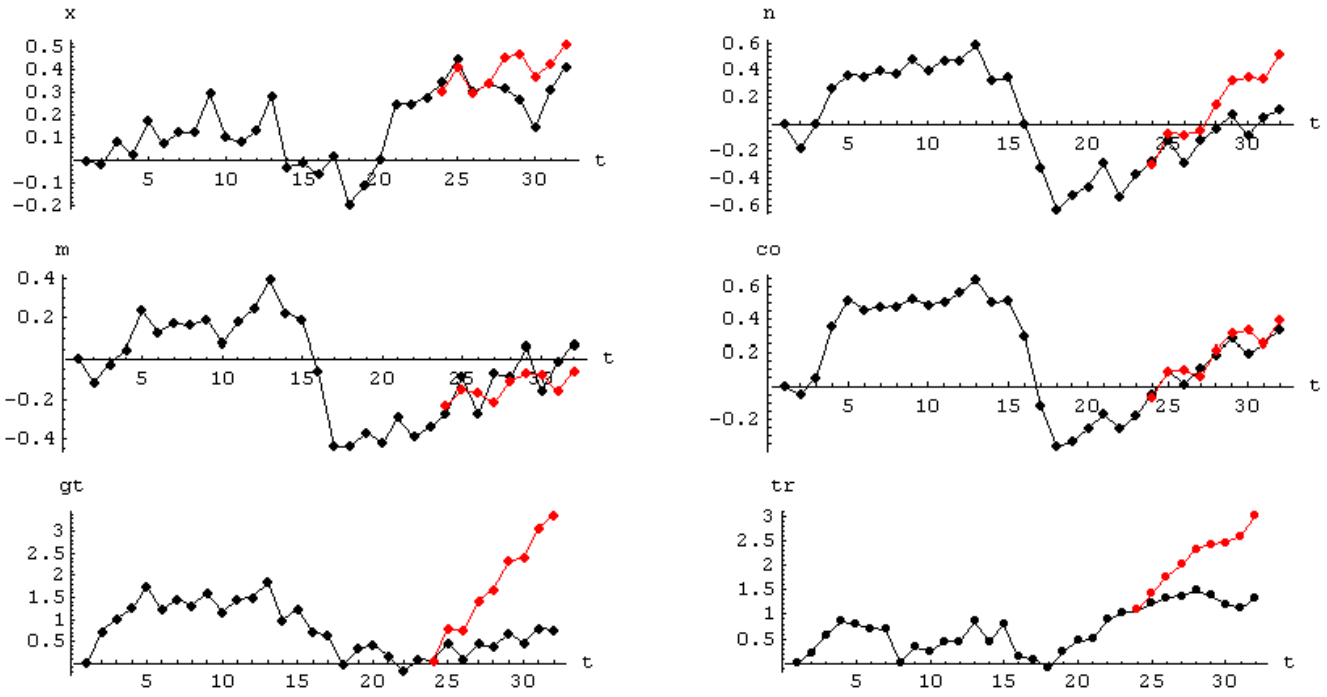



Рис. 98. Прогноз изменения состояния объекта при максимальных затратах управления.

Приведем пример нахождения оптимального управления в рамках рассматриваемой эконометрической модели. На рисунках 98 и 97 показаны результаты ретроспективного оптимального управления на восемь кварталов. Черной линией обозначены исходные данные, а красной — полученные в результате оптимизации и моделирования. Функция стоимости управления назначена как линейная комбинация разности последующих во времени векторов управляющего воздействия, то есть учтены последовательные изменения переменных `gt` и `tr`. На верхних четырех графиках каждого рисунка показаны значения переменных состояния $[\mathtt{x}, \mathtt{n}, \mathtt{m}, \mathtt{co}]^{\mathtt{T}}$, отложенные по осям ординат. На нижних двух графиках каждого рисунка показаны значения переменных управления. Значения всех переменных показаны в унифицированной шкале. По осям абсцисс всех графиков отложено время в кварталах.

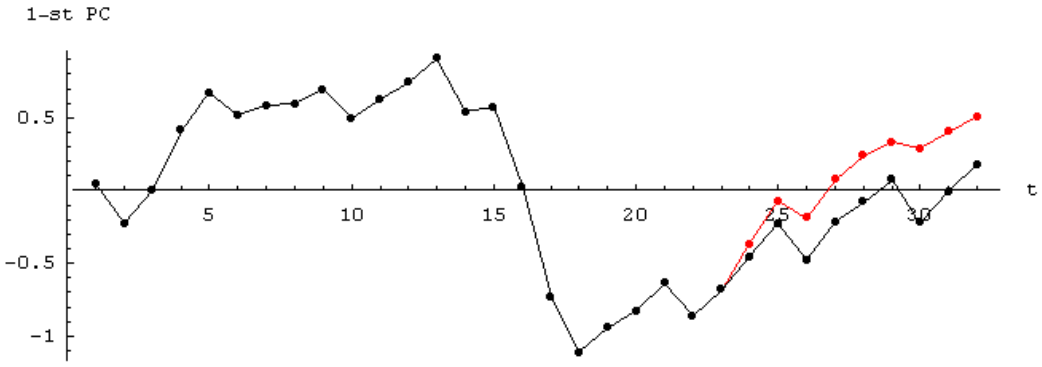



Рис. 99. Изменение интегрального индикатора состояния экономики страны.

На рис. 97 показано изменение индикатора $q$ состояния экономики. Черная линия показывает фактическое значения индикатора за исследуемый период времени. Красная линия показывает значение индикатора при оптимальном управлении.

В данной работе на сквозном тестовом примере показаны определение элементов системы. Также показаны две основные функции системы поддержки принятия решений: прогноз состояния объекта управления и нахождения оптимальных управляющих воздействий. Показано, что существует множество различных траекторий $\{\mathbf{y}_{t_1}, ..., \mathbf{y}_{t_n}\}$, которые позволяют достичь целевое состояние объекта управления $\bar{\mathbf{y}}_{t_n}$. Стоимость управления при выборе траектории можно оптимизировать. Не всегда максимальная стоимость управления приводит к оптимальному результату.

# Список основных обозначений

Матрицы обозначены заглавными буквами, векторы — полужирными прописными буквами, множества — каллиграфическими буквами.

$\mathbb{R}$ — множество действительных чисел

$\mathbb{N}$ — множество натуральных чисел

$\mathsf{E}$ — математическое ожидание случайной величины

$\mathsf{D}$ — дисперсия случайной величины

$\mathbf{x}$ — набор свободных переменных, многомерная случайная величина $\mathbf{x} = [x_1, ..., x_j, ..., x_n]^\mathsf{T} \in \mathbb{R}^n$

$\mathbf{y}$ — вектор зависимых переменных, $\mathbf{y} = [y_1, ..., y_i, ..., y_m]^\mathsf{T} \in \mathbb{R}^m$

$\mathbf{x}_i$ — $i$-й объект выборки, реализация многомерной случайной величины $\mathbf{x}$, $\mathbf{x}_i \in \mathbb{R}^n$

$\boldsymbol{\chi}_j$ — реализации $j$-й свободной переменной, признак, $\boldsymbol{\chi}_j = [x_{1j}, ..., x_{mj}]^\mathsf{T} \in \mathbb{R}^m$

$\mathbf{X}$ — матрица плана, $\mathbf{X} = [\mathbf{x}_1^\mathsf{T}, ..., \mathbf{x}_m^\mathsf{T}]^\mathsf{T}$, $\mathbf{X} = [\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n]$

$\mathfrak{D}$ — выборка, множество пар $\{(\mathbf{x}_i, y_i) |\ i = 1, ..., m\}$, также $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$

$\mathcal{I}$ — множество индексов элементов выборки (объектов)

$\mathcal{B}$ — множество индексов опорных объектов, $\mathcal{B} \subset \mathcal{I}$

$\mathcal{J}$ — множество индексов свободных переменных (признаков)

$\mathcal{A}$ — множество индексов активных признаков, $\mathcal{A} \subset \mathcal{J}$

$\mathbf{X}_\mathcal{A}$ — подмножество признаков, заданное индексным множеством $\mathcal{A}$

$m$ — число зависимых переменных, размерность пространства зависимых переменных, $m = |\mathcal{I}|$

$n$ — число свободных переменных, размерность пространства свободной переменной, $n = |\mathcal{J}|$

$f$ — регрессионная модель, $f = f(\mathbf{w}, \mathbf{x})$, по определению $f : (\mathbf{w}, \mathbf{x}) \mapsto y$

$\mathbf{f}$ — вектор значений регрессионной модели, $\mathbf{f} = [f(\mathbf{w}, \mathbf{x}_1), ..., f(\mathbf{w}, \mathbf{x}_m)]^\mathsf{T}$,
вектор-функция $\mathbf{f}(\mathbf{w}, \mathbf{X}) \mapsto \mathbf{y}$

$\mathbf{w}$ — вектор параметров $\mathbf{w} = [w_1, ..., w_n]^\mathsf{T}$ модели

$\boldsymbol{\varepsilon}$ — многомерная случайная величина $\boldsymbol{\varepsilon} = [\varepsilon_1, ..., \varepsilon_m]^\mathsf{T}$, вектор регрессионных остатков $\hat{\boldsymbol{\varepsilon}}$

$\sigma_\varepsilon^2$ — дисперсия элементов многомерной случайной величины $\boldsymbol{\varepsilon}$, описываемых ковариационной матрицей $\sigma_\varepsilon^2 \mathbf{I}$

$\mathbf{A}^{-1}$ — ковариационная матрица многомерной случайной величины $\mathbf{w}$

$\mathbf{B}^{-1}$ — ковариационная матрица многомерной случайной величины $\mathbf{y}$

$\mathbf{J}$ — матрица Якоби функции $f$ с элементами $J_{ij} = \left[\frac{\partial f(\mathbf{w}, \mathbf{x}_i)}{\partial w_j}\right], i \in \mathcal{I}, j \in \mathcal{J}$

$S$ — функция ошибки, $S = S(\mathbf{w})$, полный вариант $S = S(\mathbf{w}|\mathfrak{D}, f)$ при заданной выборке $\mathfrak{D}$ и фиксированной модели $f(\mathbf{w}, \mathbf{x})$

$\nabla S$ — градиент функции ошибки $S(\mathbf{w})$ в пространстве параметров $\mathcal{W} \ni \mathbf{w}$, $\nabla S(\mathbf{w}) = \left[\frac{\partial S(\mathbf{w})}{\partial w_j}\right], j \in \mathcal{J}$

$\mathbf{H}$ — матрица Гессе функции $f$ с элементами $H_{ij} = \left[\frac{\partial^2 S(\mathbf{w})}{\partial w_j \partial w_k}\right], j, k \in \mathcal{J}$, $\mathbf{H} = \nabla^2 S(\mathbf{w})$

$g$ — порождающая функция, $g = g(\mathbf{w}, \cdot)$

$\mathfrak{G}$ — множество порождающих функций, $\mathfrak{G} = \{g\}$

$\mathfrak{F}$ — множество индкутивно-порожденных регрессионных моделей, $\mathfrak{F} = \{f\}$

$[\cdot]$ — элементы матрицы или вектора, например: матрица $\mathbf{X} = [x_{ij}]$, вектор $\mathbf{y} = [y_1, ..., y_m]^\mathsf{T}$

$\|\cdot\|$ — евклидова норма вектора $\|\cdot\|_2$, если нижним индексом не указано иное
$\langle\cdot,\cdot\rangle$ — скалярное произведение двух векторов

**Замечания.** Число элементов вектора параметров $\mathbf{w}$ может не совпадать с числом элементов свободной переменной $\mathbf{x}$ в существенно нелинейных регрессионных моделях.

Использование при анализе случайной величины $\mathbf{y}$, обозначений ее оценки $\hat{\mathbf{y}}$ и ее фактического значения $\overline{\mathbf{y}}$ необязательно, так как оценка многомерной случайной величины $\mathbf{y}$ есть значение функции регрессии, $\mathrm{E}(\mathbf{y}) = \hat{\mathbf{y}} = \mathbf{f} = \mathbf{f}(\mathbf{w_0}, \mathbf{X})$, а фактическое значение $\overline{\mathbf{y}}$ используется в тексте только совместно с фактическим значением $\mathbf{X}$ случайной переменной $\mathbf{x}$. Эта пара является выборкой и обозначается $\mathfrak{D} = (\mathbf{X}, \mathbf{y})$. При этом считается, что речь идет о некоторой измеряемой реализации двух этих величин.

# Список иллюстраций

# Список таблиц

# Литература


[1] ГОСТ 8.207-76, Государственная система обеспечения единства измерений. Прямые измерения с многократными наблюдениями. Методы обработки результатов наблюдений. Основные положения, 1976.

[2] Gropup method for data handling. http://www.gmdh.net, 2000.

[3] Международная конвергенция измерения капитала и стандартов капитала: новые подходы. Technical report, Банк международных расчетов, Базель, Швейцария, 2004.

[4] Fasta sequence database, 2011.

[5] Fasta sequence database, example of a record, 2011.

[6] Игра в цыфирь, или как теперь оценивают труд ученого. *Сборник статей о библиометрике*, 2011.

[7] The dblp computer science bibliography, 10 2012.

[8] Robert Adler, John Ewing, Peter Taylor, et al. Citation statistics. *Statistical Science*, 24(1):1, 2009.

[9] Michael Affenzeller and Stephan Winkler. *Genetic algorithms and genetic programming: modern concepts and practical applications*. CRC Press, 2009.

[10] A. A. Afifi, V. Clark, and S. May. *Computer-aided multivariate analysis*. CRC Press, 2004.

[11] Alan Agresti. *An introduction to categorical data analysis*, volume 423. Wiley-Interscience, 2007.

[12] Leona S. Aiken, Stephen G. West, and Raymond R. Reno. *Multiple regression: testing and interpreting interactions*. SAGE, 1991.

[13] S. A. Aivazian, S. V. Borisova, E. A. Lakalin, and V. L. Makarov. Econometric modelling of the russian economy. *Acta Applicandae Mathematica*, 78(1-3):3–19, 2003.

[14] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[15] H. Akaike. A bayesian analysis of the minimum aic procedure. *Ann. Inst. Statist. Math.*, 2(30):9–15, 1978.

[16] John Aldrich. R. a. fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, 1997.

[17] A. Alessandri, C. Cervellera, D. Maccio, and M. Sanguineti. Optimization based on quasi-monte carlo sampling to design state estimators for non-linear systems. *Optimization*, 59:963–984, 2010.

[18] Mukhtar M. Ali and Carmelo Giaccotto. A study of several new and existing tests for heteroscedasticity in the general linear model. *Journal of Econometrics*, 26(3):355 – 373, 1984.

[19] Takeshi Amemiya. Selection of regressors. *International Economic Review, Department of Economics, University of Pennsylvania and Osaka University Institute of Social and Economic Research Association*, 21(2):331–354, 1980.

[20] Senjian An, Wanquan Liu, and Svetha Venkatesh. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognition*, 40:2154–2162, 2007.

[21] Tomohiro Ando and Ruey Tsay. Predictive likelihood for bayesian model selection and averaging. *International Journal of Forecasting*, 26:744–763, 2010.

[22] F. J. Anscombe and J. W. Tukey. The examination and analysis of residuals. *Technometrics*, 5:141–160, 1963.

[23] Thomas J. Archdeacon. *Correlation and regression analysis: a historian's guide*. University of Wisconsin Press, 1994.

[24] Itamar Arel, Derek C. Rose, and Thomas P. Karnowski. Deep machine learning – a new frontier in artificial intelligence research. *IEEE Computational Intelligence Magazine*, November:13–19, 2010.

[25] Sylvain Arlot, Gilles Blanchard, and Etienne Roquain. Some non-asymptotic results on resampling in high dimension, i: confidence regions. *Annals of Statistics (submitted)*, 2009.

[26] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

[27] Dursun Aydin and Memmedaga Memmedli. Optimum smoothing parameter selection for penalized least squares in form of linear mixed effect models. *Optimization*, 61:459–476, 2012.

[28] Vijay Balasubramanian. *MDL, Bayesian Inference and the Geometry of the Space of Probability Distributions*, pages 81–99. MIT Press, 2005.

[29] David Barber and Christopher M. Bishop. Ensemble learning in bayesian neural networks. In *Neural Networks and Machine Learning*, pages 215–237. Springer, 1998.

[30] M. Bekara and G. Fleury. Model selection using cross validation bayesian predictive densities. *Seventh International Symposium on Signal Processing and Its Applications*, 2:507–510, 2003.

[31] M. Bekara, L. Knockaert, A.-K. Seghouane, and G. Fleury. A model selection approach to signal denoising using kullback's symmetric divergence. *Signal Processing*, 86(7):1400–1409, 2006.

[32] D. A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Reggression.* New York: John Wiley and Sons, 1991.

[33] David A. Belsley, Edwin Kuh, and Roy E. Welsh. *Regression diagnistics: Identifying Influential Data and Sources of Collinearity.* Wiley-Intersvience, 2004.

[34] Adi Ben-Israel and Thomas N. E. Greville. *Generalized Inverses.* Springer-Verlag, 2003.

[35] L. Benati and P. Surico. Var analysis and the great moderation. *American Economic Review*, 99(4):1636–52, 2009.

[36] Yoshua Bengio. Learning deep architectures for ai. *Foundations and Trends inMachine Learning*, 2(1):1–127, 2009.

[37] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. Technical report, Department of computer science and operations research, U. Montreal, 2012.

[38] F. Berghen. *LARS Library: Least Angle Regression Stagewise Library.* Addision-Wesley, 2005.

[39] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory.* Wiley, 1994.

[40] Michael J. A. Berry and Gordon Linoff. *Data mining techniques: for marketing, sales, and customer relationship management.* Wiley, 2004.

[41] B. Betro and C. Vercellis. Bayesian nonparametrie inference and monte carlo optimization. *Optimization*, 17:681–694, 2007.

[42] Marco Better, Fred Glover, and Michele Samorani. Classification by vertical and cutting multi-hyperplane decision tree induction. *Decision Support Systems*, 48(3):430–436, 2010.

[43] H. S. Bhat and N. Kumar. On the derivation of the bayesian information criterion. Technical report, School of Natural Sciences, University of California, 2010.

[44] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics, Volume 1: Basic and Selected Topics.* Pearson Prentice–Hall, 2007.

[45] C. Bishop. *Neural networks and Machine Learning.* Springer, 1997.

[46] C. Bishop. *Pattern Recognition And Machine Learning.* Springer, 2006.

[47] C. M. Bishop. A new framework for machine learning. In *Computational Intelligence: Research Frontiers, IEEE World Congress on Computational Intelligence, WCCI 2008, Hong Kong*, pages 1–24. Springer, 2008.

[48] C. M. Bishop and J. Lasserre. Generative or discriminative? getting the best of both worlds. In J. M. et al. Bernardo, editor, *In Bayesian Statistics 8*, pages 3–23. Oxford University Press, 2007.

[49] Christopher M. Bishop and Michael E. Tipping. Bayesian regression and classification. *Advances in Learning Theory: Methods, Models and Applications*, 190:267–285, 2003.

[50] A. Bjorkstrom. Ridge regression and inverse problems. Technical report, Stockholm University, Sweden, 2001.

[51] Jonathan Borwein and Adrian Lewis. *Convex Analysis and Nonlinear Optimization.* Springer, 2000.

[52] Charles S Bos. A comparison of marginal likelihood computation methods. In *Compstat*, pages 111–116. Springer, 2002.

[53] G. E. P. Box and N. R. Draper. *Empirical model-building and response surfaces.* New York: John Wiley & Sons, 1987.

[54] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[55] Michael Bulmer. *Francis Galton: Pioneer of Heredity and Biometry.* Johns Hopkins University Press, 2003.

[56] K. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference.* Springer, 2002.

[57] Kenneth P Burnham and David R Anderson. Multimodel inference understanding aic and bic in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[58] H. Cardot, P. Cenac, and J.-M. Monnez. Fast clustering of large datasets with sequential k-medians: a stochastic gradient approach. *ArXiv*, oai:arXiv.org:1101.4179, 2011.

[59] Jacques Carette. Understanding expression simplification, 2004.

[60] David Mackay Cavendish, David J. C. Mackay, and Cavendish Laboratory. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11:1035–1068, 2003.

[61] Gavin C. Cawley and Nicola L. C. Talbot. Preventing over-fitting during model selection using bayesian regularisation of the hyper-parameters. *Journal of Machine Learning Research*, 8:841–861, 2007.

[62] G. Celeux, J.-M. Marin, and C. P. Robert. Selection bayesienne de variables en regression lineaire. *Journal de la Societe Francaise de Statistique*, 147:59–79, 2006.

[63] Changgee Chang and Ruey S. Tsay. Estimation of covariance matrix via the sparse cholesky factor with lasso. *Journal of Statistical Planning and Inference*, 140:3858–3873, 2010.

[64] Samprit Chatterjee and Ali S. Hadi. *Regression analysis by example.* John Wiley and Sons, 2006.

[65] S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning algorithm for radial basis function network. *Transaction on neural netwark*, 2(2):302–309, 1991.

[66] Y. W. Chen, C. A. Billings, and W. Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 2(50):873–896, 1989.

[67] Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: Ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer, 2010.

[68] Selina Chu, Eammon Keogh, David Hart, and Michael Pazzani. Iterative deepening dynamic time warping for time series. In *Proceedings of the Second SIAM International Conference on Data Mining*, 2002.

[69] David R. Clark and Charles A. Thayer. A primer on the exponential family of distributions. Technical report, Call Paper Program on Generalized Linear Models, 2004.

[70] Jacob Cohen, Patricia Cohen, Stephen G. West, and Leona S. Alken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* New Jersey: Lawrence Erlbaum Associates, 2010.

[71] T. F. Coleman and Y. Li. An interior, trust region approach for nonlinear minimization subject to bounds. *SIAM Journal on Optimization*, 6:418–445, 1996.

[72] David Cossock and Tong Zhang. Subset ranking using regression. In Gabor Lugosi and HansUlrich Simon, editors, *Learning Theory*, volume 4005 of *Lecture Notes in Computer Science*, pages 605–619. Springer Berlin Heidelberg, 2006.

[73] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[74] E Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. *Proceedings of the 24th national conference ACM*, pages 151–172, 1969.

[75] T. Daglish, J. Hull, and W. Suo. Volatility surfaces: Theory, rules of thumb, and empirical evidence. *Quantitative Finance*, 7(5):507–524, 2007.

[76] J. Dattorro. *Convex Optimization and Euclidean Distance Geometry.* Meboo Publishing, USA, 2005.

[77] K. Deb and J. Sachin. Running performance metrics for evolutionary multi-objective optimization. Technical report, Indian Institute of Technology Kanpur, 2002.

[78] A. J. Dobson and A. G. Barnett. *Introduction to Generalized Linear Models.* Boca Raton, FL: Chapman and Hall/CRC, 2008.

[79] Jon Doyle. Prospects for preferences. *Computational Intelligence*, 20(2):111–136, 2004.

[80] N. R. Draper and H. Smith. *Appied Regression Analysis.* John Wihley and Sons, 1998.

[81] J. Durbin and G. S. Watson. Testing for serial correlation in least-squares regression. *Biometrika*, 38:159–178, 1951.

[82] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(3):407–499, 2004.

[83] M. A. Efroymson. *Multiple regression analysis.* New York: Ralston, Wiley, 1960.

[84] J. Eggermont and J. I. van Hemert. Stepwise adaptation of weights for symbolic regression with genetic programming. In *Proceedings of the Twelveth Belgium/Netherlands Conference on Artificial Intelligence (BNAIC'00)*, pages 259–266, 2000.

[85] H. Ehrig, G. Ehrig, U. Prange, and G. Taentzer. *Fundamentals of Algebraic Graph Transformation.* Springer, 2006.

[86] H. Ehrig and G. Engels. *Handbook of Graph Grammars and Computing by Graph Transformation*, volume 1-3. World Scientific Publishing, 1997.

[87] Jo Eidsvik, Andrew O Finley, Sudipto Banerjee, and Håvard Rue. Approximate bayesian inference for large spatial datasets using predictive process models. *Computational Statistics & Data Analysis*, 56(6):1362–1380, 2012.

[88] V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory Probab. Appl.*, 14(1):153–158, 1969.

[89] Y. Ephraim and W. J. J. Roberts. Revisiting autoregressive hidden markov modeling of speech signals. *IEEE Signal Processing Letters*, 12:166–169, 2005.

[90] R. L. Eubank and Will Thomas. Detecting heteroscedasticity in nonparametric regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1):145–155, 1993.

[91] M. Farina. A minimal cost hybrid strategy for pareto optimal front approximation. *Evolutionary Optimization*, 3 (1):41–52, 2001.

[92] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:381–396, 2002.

[93] Joseph L. Fleiss, Bruce A. Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions.* Wiley, 2003.

[94] R. Fletcher. *Practical methods of optimization. Volume 1: unconstrained optimization*. Wiley, 1980.

[95] Dean P. Foster and Robert A. Stine. *The Contribution of Parameters to Stochastic Complexity*, pages 195–213. MIT Press, 2005.

[96] Rudolf J. Freund, William J. Wilson, and Ping Sa. *Regression Analysis*. Elsevier, 2006.

[97] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, 2003.

[98] R. Frisch. *Statistical Confluence Analysis by means of complete regression systems*. Universitetets Okonomiske Institutt, 1934.

[99] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1983.

[100] Johannes Furnkranz and Eyke Hullermeier. Pairwise preference learning and ranking. *Machine Learning: ECML 2003*, pages 145–156, 2003.

[101] Gianfranco Galmaccimola. Collinearity detection in linear regression models. *Computational Economics*, 9:215–227, 1996.

[102] F. Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute*, 15:246–63, 1886.

[103] Nick Galwey. *Introduction to mixed modelling : beyond regression and analysis of variance*. John Wiley & Sons, 2006.

[104] Alexander Gammerman and Vladimir Vovk. Kolmogorov complexity: Sources, theory and applications. *Comput. J*, 42(4):252–255, 1999.

[105] Anthony Garratt, Kevin C Lee, M. Hashem Pesaran, and Yongcheol Shin. A structural cointegrating var approach to macroeconometric modelling. Cambridge Working Papers in Economics 9823, Faculty of Economics, University of Cambridge, 1998.

[106] N. Garshina and C. Vladislavleva. *On development of a complexity measure for symbolic regression via genetic programming. Modeling Report*. Eindhoven, The Netherlands: Technische Universiteit Eindhoven, 2004.

[107] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007.

[108] Christian Genest and Jock MacKay. The joy of copulas: bivariate distributions with uniform marginals. *The American Statistician*, 40(4):280–283, 1986.

[109] Jonathan Gillard. Asymptotic variance–covariance matrices for the linear structural model. *Statistical Methodology*, 8:291–301, 2010.

[110] D. E. Goldberg. *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, 1989.

[111] B. Goldstein, J. McNames, M. Ellenby abd L. Ibsen, S. Jacques, M. Aboy, T. Thong, C. Phillips, and G. Levitte. *Current Concepts in Pediatric Critical Care*. Des Planes, IL, USA, 2004.

[112] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 1996.

[113] D. Gorissen, I. Couckuyt, P. Demeester, T. Dhaene, and K. Crombecq 2010. A surrogate modeling and adaptive sampling toolbox for computer based design. *Journal of Machine Learning Research*, 11:2051–2055, 2010.

[114] Jan Gorodkin, Lars Kai Hansen, Anders Krogh, Claus Svarer, and Ole Winther. A quantitative study of pruning by optimal brain damage. *Int. J. Neural Syst*, 4(2):159–169, 1993.

[115] P. Grünwald. *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2005.

[116] P. D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

[117] P. D. Grünwald. Safe learning: bridging the gap between bayes, mdl and statistical learning theory via empirical convexity. In *Proceedings 24th Conference on Learning Theory (COLT 2011), Budapest*, 2011.

[118] Peter Grünwald. A tutorial introduction to the minimum description length principle. 2005.

[119] Peter Grünwald, Petri Myllymäki, Ioan Tabus, Marcelo Weinberger, Bin Yu, et al. *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*. Tampere University Press, 2008.

[120] Peter Grünwald, In Jae Myung, and Mark Pitt. *Advances in Minimum Description Length*. MIT Press, 2005.

[121] Maya R. Gupta and Yihua Chen. Theory and use of the em algorithm. *Foundations and Trends in Signal Processing*, 4(3):223–296, 2010.

[122] I. Guyon and S. Gunn. *Feature extraction: foundation and applications*. Springer, 2006.

[123] Anders Hald. On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science*, 14 (2):214–222, 1999.

[124] James Douglas Hamilton. *Time series analysis*. Princeton University Press, 1994.

[125] Mark H Hansen and Bin Yu. Minimum description length model selection criteria for generalized linear models. *Lecture Notes-Monograph Series*, pages 145–163, 2003.

[126] S. Har-Peled, D. Roth, and D. Zimak. Constraint classification for multiclass classification and ranking. In *NIPS*, pages 785–792, 2003.

[127] James Hardin and Joseph Hilbe. *Generalized Linear Models and Extensions.* College Station: Stata Press, 2007.

[128] Simar Hardle. *Applied Multivariate Statistical Analysis.* Springer, 2004.

[129] Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 164–171. Morgan Kaufmann, San Mateo, CA, 1993.

[130] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1(1):1–29, 2007.

[131] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer, 2001.

[132] Amaury Hazan, Rafael Ramirez, Esteban Maestre, Alfonso Perez, and Antonio Pertusa. Modelling expressive performance: A regression tree approach based on strongly typed genetic programming. In *Applications of Evolutionary Computing*, volume 3907 of *Lecture Notes in Computer Science*, pages 676–687. Springer Berlin / Heidelberg, 2006.

[133] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America*, 102(46):16569–16572, 2005.

[134] R. R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, 1976.

[135] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometric*, 32(1):1–49, 1976.

[136] R. R. Hocking. *Methods and applications of linear models regression and the analysis of variance.* Hoboken, N. J. : Wiley-Interscience, 2003.

[137] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 3(12):55–67, 1970.

[138] L. Hogben. *Handbook of linear algebra.* CRC Press, 2007.

[139] J. H. Holland. *Adaptation in natural and artificial systems.* University of Michigan Press, 1975.

[140] David W. Hosmer. *Applied survival analysis : regression modeling of time-to-event data.* Hoboken, N. J. : Wiley-Interscience, 2008.

[141] David W. Hosmer and Stanley Lemeshow. *Applied Logistic Regression.* Wiley, 2000.

[142] C. Howson and P. Urbach. *Scientific Reasoning: the Bayesian Approach.* Open Court Publishing Company, 2005.

[143] http://vak.ed.gov.ru/ru/help_desk/list. *Перечень российских рецензируемых научных журналов, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученых степеней доктора и кандидата наук*. 2012.

[144] Z. Huang. A fast clustering algorithm to cluster very large categorical data sets in data mining. In *Cooperative Research Centre for Advanced Computational Systems (Australia, Canberra)*, 1997.

[145] J. C. Hull. *Options, Futures and Other Derivatives*. Prentice Hall, 2000.

[146] Eyke Hullermeier and Johannes Furnkranz. Comparison of ranking procedures in pairwise preference learning. In *Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-04), Perugia, Italy*, 2004.

[147] Eyke Hullermeier, Johannes Furnkranz, Weiwei Cheng, and Klaus Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16):1897–1916, 2008.

[148] A. J. Isenmann. *Modern multivariate statistical techniques*. Springer, 2008.

[149] T. Jaakkola. Scaled structured prediction. Technical report, Yandex seminar, 2012.

[150] T. Jaakola and D. Sontag. Learning bayesian network structure using lp relaxations. *Proceedings of the 13th International Conference on Artifcial Intelligence and Statistics*, 9(1):358–365, 2010.

[151] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[152] E. T. Jaynes. *Probability Theory: The Logic of Science*. CUP, 2003.

[153] Harry Joe. Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, 52:5066–5074, 2008.

[154] I. T. Jolliffe. *Principal component analysis*. Springer, 2002.

[155] D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, 2001.

[156] Martin A. Keane, Jessen Yu, and John R. Koza. Automatic synthesis of both topology and tuning of a common parameterized controller for two families of plants using genetic programming. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, pages 496–504. Morgan Kaufmann, 2000.

[157] M. Keijzer and J. Foster. Crossover bias in genetic programming. In *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2007.

[158] E. J. Keogh and M. J. Pazzani. Derivative dynamic time warping. In *First SIAM International Conference on Data Mining (SDM'2001), Chicago, USA.*, 2001.

[159] David G. Kleinbaum, Lawrence L. Kupper, Keith E. Muller, and Azhar Nizam. *Applied Regression Analysis and Multivariable Methods*. Duxbury Press, 1997.

[160] T. Kloek. Note on a large-sample result in specification analysis. *Econometrica*, 43(5/6):933–936, 1975.

[161] J. Knowles and B. Corne. On metrics for comparing non-dominated sets. *IEEE Service Center, Piscata way, New Jersey*, 1:711–719, 2002.

[162] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145, 1995.

[163] A. N. Kolmogorov. On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. *American Math. Soc. Transl.*, 28:55–63, 1963.

[164] Paul Komarek. Logistic regression for data mining and high-dimensional classification. Technical report, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2004.

[165] W. Kotlowski and P. D. Grünwald. Maximum likelihood vs. sequential normalized maximum likelihood in on-line density estimation. In *Proceedings 24th Conference on Learning Theory (COLT 2011), Budapest*, 2011.

[166] John Koza. Genetic programming inc. http://www.genetic-programming.com, 2012.

[167] John R. Koza, Martin A. Keane, Matthew J. Streeter, William Mydlowec, Jessen Yu, and Guido Lanza. *Genetic Programming IV: Routine Human-Competitive Machine Intelligence*. Springer, 2005.

[168] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[169] Michael H. Kutner, Christopher J. Nachtsheim, and John Neter. *Applied Linear Regression Models*. McGrow-Holl Irwin, 2004.

[170] M. P. Kuznetsov. Integral indicator construction using copulas. *Journal of Machine Learning and Data Analysis*, 1(4):411–419, 2012.

[171] T.-Y. Kwok and D.-Y. Yeung. Constructive algorithms for structure learning in feedforward neural networks for regression problems. *IEEE Transactions on Neural Networks*, 8:630–645, 1997.

[172] C. H. Lampert. Maximum margin multi-label structured prediction. Technical report, Institute of Science and Technology Austria, 2011.

[173] L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Englewood Cliffs: Prentice Hall, 1974.

[174] Lucien Le Cam. Maximum likelihood — an introduction. *ISI Review*, 58 (2):153–171, 1990.

[175] Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer, 2000.

[176] Y. LeCun, J. Denker, S. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*, pages 598–605, San Mateo, CA, 1990. Morgan Kauffman.

[177] Youngjo Lee, John A. Nelder, and Yudi Pawitan. *Generalized linear models with random effects: unified analysis via h-likelihood*. Chapman & Hall/CRC, 2006.

[178] E. L. Lehmann and Joseph P. Romano. *Testing Statistical Hypothesis*. Springer, 2005.

[179] K. Levenberg. A method for the solution of certain non-linear probelms in least squares. *Quart. Appl. Math.*, 2:164–168, 1944.

[180] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145, 1991.

[181] James K. Lindsey. *Applying Generalized Linear Models*. Springer, 1997.

[182] Tie-Yan Liu, Thorsten Joachims, Hang Li, and Chengxiang Zhai. Introduction to special issue on learning to rank for information retrieval. *Information Retrieval*, 13:197–200, 2010.

[183] Yi Liu, Taghi M. Khoshgoftaar, and Jenq-Foung Yao. Building a novel GP-based software quality classifier using multiple validation datasets. In *IRI*, pages 644–650. IEEE Systems, Man, and Cybernetics Society, 2007.

[184] Lennart Ljung. *System Identification: Theory For the Use*. N. J.: PTR Prentice Hall, 1999.

[185] Hedibert Freitas Lopes, Ajax R. Bello Moreirac, and Alexandra Mello Schmidt. Hyperparameter estimation in forecast models. *Computational Statistics & Data Analysis*, 29:387–410, 1999.

[186] Helmut Lutkepohl. Vector autoregressions, 1999.

[187] Yunqian Ma and Vladimir Cherkassky. Characterization of data complexity for svm methods. In *Proceedings of Intemational Joint Conference on Neural Networks*, pages 919–924, 2005.

[188] D MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[189] David Mackay. *Maximum Entropy and Bayesian Methods*, chapter Hyperparameters: optimise or integrate out?, pages 327–335. Kluwer Academic, 1994.

[190] David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1991.

[191] David J. C. Mackay. Choice of basis for laplace approximation. *Machine Learning*, 33:77–86, 1998.

[192] H. R. Madala and A. G. Ivakhnenko. *Inductive Learning Algorithms for Complex Systems Modeling*. CRC Press, 1994.

[193] Janos Madar, Janos Abonyi, and Ferenc Szeifert. Genetic programming for the identification of nonlinear input-output models. *Industrial and Engineering Chemistry Research*, 44(9):3178–3186, 2005.

[194] L. O. Mafeteiu-Scai, V. Negru, D. Zaharie, and O. Aritony. Average bandwidth reduction in sparse matrices using hybrid heuristics. *Studia univvertitade babes bolyai, informatica*, 3:97–102, 2011.

[195] P. C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2 (1):49–55, 1936.

[196] Abdul Majid, Asifullah Khan, and Anwar M. Mirza. Intelligent combination of kernels information for improved classification. In *Proceedings of the Fourth International Conference on Machine Learning and Applications*, 2005.

[197] C. L. Mallows. Some comments on $c_p$. *Technometrics*, 15:661–675, 1973.

[198] A. Marconato, A. Boni, B. Caprile, and D. Petri. Model selection for power efficientanalysis of measurement data. In *Instrumentation and Measurement Technology Conference*, pages 1524–1529, 2006.

[199] P. Marenbach, K. Betterhausen, and S. Freyerm. Signal path oriented approach for generation of dynamic process. In *Genetic Programming: Proceedings of the First Annual Conference*, pages 327–332. MIT Press, 1996.

[200] D. W. Marquardt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, 12(3):605–607, 1996.

[201] A. F. T. Martins. The geometry of constrained structured prediction: Applications to inference and learning of natural language syntax. Technical report, Carnegie Mellon University, 2012.

[202] Vijay K. Mathur. How well do we know pareto optimality? *Journal of Economic Education*, 22(2):172–178, 1991.

[203] Peter McCullagh and John Nelder. *Generalized Linear Models*. Boca Raton: Chapman and Hall/CRC, 1989.

[204] J. McNames. Local averaging optimization for chaotic time series prediction. *Neurocomputing*, 48(1-4):279–297, 2002.

[205] J. McNames. *Microelectrode Recordings in Movement Disorder Surgery*. Thieme, New York, 2004.

[206] Scott W. Menard. *Applied Logistic Regression Analysis*. Sage Publications, 2001.

[207] Terence C Mills. *Time Series Techniques for Economists*. Cambridge University Press, 1990.

[208] Yoichi Miyata. Laplace approximations to means and variances with asymptotic modes. *Journal of Statistical Planning and Inference*, 140:382–392, 2010.

[209] Douglas C. Montgomery. *Introduction to Linear Regression Analysis*. Wiley, 2007.

[210] Douglas C. Montgomery. *Design and analysis of experiments*. John Wiley and Sons, 2008.

[211] McKay Mori, Naoki. Equivalent decision simplification. *Proceedings Workshop on Intelligent and Evolutionary Systems*, 1:1–8, 2007.

[212] Morten Mørup, Kristoffer Hougaard Madsen, and Lars Kai Hansen. Approximate L0 constrained non-negative matrix and tensor factorization. In *ISCAS*, pages 1328–1331. IEEE, 2008.

[213] V. Mottl, M. Lange, V. Sulimova, and A Yermakov. Signature verification based on fusion of on-line and off-line kernels. In *19th International Conference on Pattern Recognition, ICPR*, 2008.

[214] Volker Nannen. *A Short Introduction to Model Selection, Kolmogorov Complexity and Minimum Description Length*. 2010. Comment: 20 pages, Chapter 1 of The Paradox of Overfitting, Master's thesis, Rijksuniversiteit Groningen, 2003.

[215] N. Nikolaev and H. Iba. Accelerated genetic programming of polynomials, genetic programmimg and evolvable machines. *Kluwer Academic Publ.*, 2(3):231–257, 2002.

[216] V. D. Nogin. The edgeworth-pareto principle and relative importance of criteria in the case of a fuzzy preference relation. *Computational Mathematics and Mathematical Physics*, 43(11):1666–1676, 2003.

[217] V. D. Nogin. A simplified variant of the hierarchy analysis on the ground of nonlinear convolution of criteria. *Computational Mathematics and Mathematical Physics*, 44(7):1194–1202, 2004.

[218] Charles W. Ostrom. *Time series analysis: regression techniques 2nd ed.* Sage Publications, Thousand Oaks, California, 1990.

[219] Mikko Packalen and Tony S. Wirjanto. Inference about clustering and parametric assumptions in covariance matrix estimation. *Computational Statistics & Data Analysis*, 56:1–14, 2012.

[220] A. R. Pagan and A. D. Hall. *Diagnostic tests as residual analysis*. Australian National University, 1983.

[221] M. Papagelis and D. Plexousakis. Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents. *Engineering Applications of Artificial Intelligence*, 18(7):781–789, 2005.

[222] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. An effective hash-based algorithm for mining association rules. *SIGMOD Rec.*, 24(2):175–186, 1995.

[223] Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist..*, 33(3):1065–1076, 1962.

[224] Frederic Pascal, Hugo Harari-Kermadec, and Pascal Larzabal. The empirical likelihood method applied to covariance matrix estimation. *Signal Processing*, 90:566–578, 2010.

[225] I. Pavlidis, R. Singh, and N. Papanikolopoulos. Recognition of on-line handwritten patterns through shape metamorphosis. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 3, pages 18–22, 1996.

[226] Donald B. Percival and Andrew T. Walden. *Spectral Analysis for Physical Applications*. Cambridge University Press, 1993.

[227] John Peterson, Guillermo Miro-Quesada, and Enrique del Castillo. A bayesian reliability approach to multiple response optimization with seemingly unrelated regression models. *Journal of Quality Technology and Quantitative Management*, 6 (4):353–369, 2009.

[228] V. V. Podinovsky. *Introduction to the importance factors theory in multicriteria decision problem*. Moscow: Fizmatlit, 2007.

[229] Riccardo Poli, William B. Langdon, and Nicholas F. McPhee. *A Field Guide to Genetic Programming*. Kluwer/Springer, 2008.

[230] Theodore M. Porter. *Karl Pearson: the scientific life in a statistical age*. Princeton University Press, 2004.

[231] C. Radhakrishna Rao. *Linear Statistical Inference and its Applications*. Wiley Series in Probability and Statistics, 2002.

[232] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[233] J. O. Rawlings, S. G. Pantula, and D. A. Dickey. *Applied Regression Analysis: A Research Tool*. New York: Springer-Verlag, 1998.

[234] John Rice. Bandwidth choice for nonparametric regression. *Annalsof Statistics*, 4(12):1215–1230, 1984.

[235] Jorma Rissanen, Teemu Roos, and Petri Myllymäki. Model selection by sequentially normalized least squares. *J. Multivariate Analysis*, 101(4):839–849, 2010.

[236] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27(3):832–837, 1956.

[237] Patrick Royston. *Multivariable model-building : a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*. John Wiley & Sons, 2008.

[238] Andrzej Ruszczynski. *Nonlinear Optimization*. Princeton University Press, 2006.

[239] T. Sasaki. Simplification of algebraic expression by multiterm rewriting rules. In *Proceedings of the 1986 Symposium on Symbolic and Algebriaic Computation*, pages 115–120, 1986.

[240] Gunther Schmidt. *Relational mathematics*, volume 132. Cambridge University Press, 2010.

[241] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[242] S. R. Searle. *Linear models*. New York: John Wiley & Sons, 1971.

[243] G. A. F Seber and C. J. Wild. *Nonlinear Regression*. Wiley-IEEE, 2003.

[244] G. A.F. Seber. *Multivariate Observations*. Hoboken, NJ: John Wiley and Sons, 1984.

[245] George Arthur Frederick Seber. *Linear regression analysis*. Hoboken, N. J. : Wiley-Interscience, 2003.

[246] V. Shakin and G. Ptashko. Decision support system using multimedia case history: quantitative comparision and multivariate statistical analysis. In *IEEE Computer-Based Medical Systems, Dublin*, pages 128–133, 2005.

[247] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.

[248] Sidney Siegel. Nonparametric statistics. *The American Statistician*, 11(3):13–19, 1957.

[249] Christopher A. Sims. Macroeconomics and reality. *Econometrica*, 48(48):1–50, 1980.

[250] M. D. Smith. Modelling sample selection using archimedean copulas. *The Econometrics Journal*, 6:99–123, 2003.

[251] Alex J. Smola and Bernhard Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.

[252] Terence Soule and James A. Foster. Support for multiple causes of code growth in GP. Position paper at the Workshop on Evolutionary Computation with Variable Size Representation at ICGA-97, 1997.

[253] Terence Soule and James A. Foster. Removal bias: a new cause of code growth in tree based evolutionary programming. In *1998 IEEE International Conference on Evolutionary Computation*, pages 781–786, Anchorage, Alaska, USA, 1998. IEEE Press.

[254] H. Spath. *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*. New York: Halsted Press, 1985.

[255] F. M. Speed, R. R. Hocking, and D. P. Hackney. Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, 73(361):105–112, 1978.

[256] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*, 64(4):583–639, 2002.

[257] David R. Stoutemyer. Ten commandments for good default expression simplification. *J. Symb. Comput*, 46(7):859–887, 2011.

[258] David R. Stoutemyer. Simplifying products of fractional powers of powers, 2012. Comment: 34 pages. 17 tables. Includes Mathematica rewrite rules. To appear in Communications in Computer Algebra.

[259] Matthew J. Streeter. The root causes of code growth in genetic programming. In Conor Ryan, Terence Soule, Maarten Keijzer, Edward Tsang, Riccardo Poli, and Ernesto Costa, editors, *Genetic Programming, Proceedings of EuroGP'2003*, volume 2610 of *LNCS*, pages 443–454, Essex, 2003. Springer-Verlag.

[260] V. Strijov and V. Shakin. Index construction: the expert-statistical method. *Environmental research, engineering and management*, 26(4):51–55, 2003.

[261] Vadim Strijov, Goran Granic, Jeljko Juric, Branka Jelavic, and Sandra Antecevic Maricic. Integral indicator of ecological impact of the croatian thermal power plants. *Energy*, 36(7):4144–4149, 2011.

[262] Vadim Strijov, Ekaterina Krymova, and Gerhard Wilhelm Weber. Evidence optimization for consequently generated models. *Mathematical and Computer Modelling*, 57(1-2):50–56, 2013.

[263] Vadim Strijov and Peter Letmathe. Integral indicators based on data and rank-scale expert estimations. In *Intellectual Information Processing. Conference Proceedings*, pages 107–110, 2010.

[264] Vadim Strijov and Gerhard Wilhelm Weber. Nonlinear regression model generation using hyperparameter optimization. *Computers and Mathematics with Applications*, 60(4):981–988, 2010.

[265] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, 2005.

[266] I. V. Tetko, D. J. Livingstone, and A. I. Luik. Neural network studies. comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.*, 35:826–833, 1995.

[267] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 32(1):267–288, 1996.

[268] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[269] I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference*, 2:258–265, 2005.

[270] Siddhaling Urolagin, K. V. Prema, and N. V. Subba Reddy. Extending the principle of optimal brain damage to feature selection. In *Proceedings of the International Conference on Cognition and Recognition*, 2005.

[271] David A. Van Veldhuizen and Gary B. Lamont. Multiobjective evolutionary algorithm test suites. In *Proceedings of Symposium on Applied Computing*, pages 351–357, 1999.

[272] E Vladislavleva. *Model-based Problem Solving through Symbolic Regression via Pareto Genetic Programming: PhD thesis.* Tilburg University, Tilburg, the Netherlands, 2008.

[273] E Vladislavleva, G. Smith, and D. Hertog. Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. *EEE Transactions on Evolutionary Computation*, 13(2):333–349, 2009.

[274] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[275] C.-P. Wei, Lee Y.-H., and C.-M. Hsu. Empirical comparison of fast partitioning-based clustering algorithms for large data sets. *Expert Systems with Applications*, 24(4):351–363, 2003.

[276] S. Weisberg. *Applied linear regression.* Wiley, New York, 1980.

[277] Max Welling and Sridevi Parise. Bayesian random fields: The bethe-laplace approximation. In *UAI*, pages 512–519. AUAI Press, 2006.

[278] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–72, 1997.

[279] D. H. Wolpert and W. G. Macready. Coevolutionary free lunches. *IEEE Transactions on Evolutionary Computation*, 9(6):721–735, 2005.

[280] David Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8:1341–1390, 1996.

[281] Hulin Wu and Jin-Ting Zhang. *Nonparametric regression methods for longitudinal data analysis.* John Wiley and Sons, 2006.

[282] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pages 1192–1199, 2008.

[283] X. Yao. A review of evolutionary artificial neural networks. *International Journal of Intelligent Systems*, 8(4):39–67, 1993.

[284] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *Research and development in information retrieval*, pages 271–278, 2007.

[285] Ivan Zelinka, Zuzana Oplatkova, and Lars Nolle. Analytic programming – symbolic regression by means of arbitrary evolutionary algorithm. *I. J. of Simulation*, 6(9):44–56, 2008.

[286] Yunong Zhang and W. E. Leithead. Exploiting hessian matrix and trust-region algorithm in hyperparameters estimation of gaussian process. *Applied Mathematics and Computation*, 171:1264–1281, 2005.

[287] Г. Г. Азгальдов. *Теория и практика оценки качества товаров (основы квалиметрии).* Экономика, 1982.

[288] С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, and Л. Д. Мешалкин. Прикладная статистика. Классификация и снижение размерности. *Финансы и статистика*, pages 421–424, 1989.

[289] С. А. Айвазян, И. С. Енюков, and Л. Д. Мешалкин. *Прикладная статистика: исследование зависимостей.* М.: Финансы и статистика, 1985.

[290] С. А. Айвазян and В. С. Мхитарян. *Прикладная статистика и основы эконометрики.* ЮНИТИ, 1998.

[291] В. И. Арнольд. *Теория катастроф.* М.: Наука, 1990.

[292] В. Б. Боков. Объединенный анализ теоретических и эмпирических данных планируемого эксперимента. *Заводская лаборатория. Диагностика материалов*, 01 (76):61–68, 2010.

[293] В. Н. Вапник. *Восстановление зависимостей по эмпирическим данным.* М.: Наука, 1979.

[294] К. В. Воронцов. Комбинаторные обоснования обучаемых алгоритмов. *Журнал вычислительной математики и математической физики*, 44 (11):2099–2112, 2004.

[295] К. В. Воронцов. *Комбинаторная теория надёжности обучения по прецедентам: Дис. док. физ.-мат. наук.* Вычислительный центр РАН, 2010.

[296] М. Г. Гафт. *Принятие решений при многих критериях.* М.: Знание, 1979.

[297] М. Г. Гафт and В. В. Подиновский. О построении решающих правил в задачах принятия решений. *Автоматика и телемеханика*, 6:128–138, 1981.

[298] Р. Голдблатт. *Топосы: Категорный анализ логики.* М.: Мир, 1983.

[299] Дж. Голуб and Ч. Ван-Лоан. *Матричные вычисления.* М.: Мир, 1999.

[300] В. А. Гордин. *Как это посчитать? Обработка метеорологической информации на компьютере. Идеи, методы, задачи.* М: МЦНМО, 2006.

[301] Е. З. Демиденко. *Линейная и нелинейная регрессии.* Финансы и статистика, 1981.

[302] Е. З. Демиденко. *Оптимизация и регрессия.* М.: Наука, 1989.

[303] Н. Джонсон and Ф. Лион. *Статистика и планирование эксперимента в технике и науке.* М.: Мир, 1980.

[304] С. В. Емельянов and О. И. Ларичев. *Многокритериальные методы принятия решений.* М.: Знание, 1985.

[305] А. А. Зайцев, В. В. Стрижов, and А. А. Токмакова. Оценка гиперпараметров регрессионных моделей методом максимального правдоподобия. *Информационные технологии*, 2:11–15, 2013.

[306] Н. В. Зубаревич, В. С. Тикунов, В. В. Крепец, В. В. Стрижов, and В. В. Шакин. Многовариантные методы интегральной оценки развития человеческого потенциала в регионах Российской Федерации. In *ГИС для устойчивого развития территорий. Материалы Международной конференции*, pages 84–105, Петропавловск-Камчатский, 2001.

[307] А. Г. Ивахненко. *Индуктивный метод самоорганизации моделей сложных систем.* Киев: Наукова думка, 1981.

[308] А. Г. Ивахненко and В. С. Степашко. *Помехоустойчивость моделирования.* Киев: Наукова думка, 1985.

[309] А. Г. Ивахненко and Ю. П. Юрачковский. *Моделирование сложных систем по экспериментальным данным.* М.: Радио и связь, 1987.

[310] В. А. Ильин. О работах А. Н. Тихонова по методам решения некорректно поставленных задач. *Математическая жизнь в СССР и за рубежом*, 1:168–175, 1966.

[311] Д. Ю. Каневский, П. Ю. Кудинов, and К. В. Воронцов. Прогнозирование с несимметричной функцией потерь при наличии стохастического тренда. In *Интеллектуализация обработки информации (ИОИ-2008): тезисы докладов*, pages 113–115, 2008.

[312] А. И. Кобзарь. *Прикладная математическая статистика.* М.: Физматлит, 2006.

[313] П. С. Краснощёков. *Математические модели в исследовании операций.* М.: Знание, 1984.

[314] П. С. Краснощеков and А. А. Петров. *Принципы построения моделей.* М.: Фазис, 2000.

[315] О. И. Ларичев. *Наука и искусство принятия решений.* М.: Наука, 1979.

[316] В. И. Левенштейн. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады Академий Наук СССР*, 163(4):845–848, 1965.

[317] Б. Г. Литвак. *Экспертная информация: методы получения и анализа.* М.: Радио и связь, 1981.

[318] Л. И. Лопатников. *Экономико-математический словарь: словарь современной экономической науки.* М.: Дело, 2003.

[319] А. В. Лотов. Аппроксимация и визуализация паретовой границы для невыпуклых многокритериальных задач. *ДАН*, 386 (6):738–741, 2002.

[320] А. В. Лотов and И. И. Поспелова. *Многокритериальные задачи принятия решений.* М: МАКС Пресс, 2008.

[321] Я. Р. Магнус, П. К. Катышев, and А. А. Персецкий. *Эконометрика.* М.: Дело, 2004.

[322] М. Мину. *Математическое программирование. Теория и алгоритмы.* М.: Мир, 1990.

[323] Е. М. Миркес. *Нейрокомпьютер. Проект стандарта.* Новосибирск: Наука, Сибирская издательская фирма РАН, 1999.

[324] Б. Г. Миркин. *Проблема группового выбора.* М.: Наука, 1974.

[325] А. П. Мотренко and В. В. Стрижов. Многоклассовая логистическая регрессия для прогноза вероятности наступления инфаркта. *Известия Тульского государственного университета, Естественные науки*, 1:153–162, 2012.

[326] В. С. Муха. *Статистические методы обработки данных.* Минск: Издательский центр БГУ, 2009.

[327] Ю. Е. Нестеров. *Методы выпуклой оптимизации.* 2010.

[328] А. И. Орлов. Современный этап развития теории экспертных оценок. *Заводская лаборатория. Диагностика материалов*, 1:60–65, 1996.

[329] А. И. Орлов. *Эконометрика.* М.: Экзамен, 2002.

[330] А. В. Панюков and А. Н. Тырсин. Взаимосвязь взвешенного и обобщенного вариантов метода наименьших модулей. *Известия Челябинского научного центра*, 1(35):6–11, 2007.

[331] И. Ш. Пинскер. Представление функций многих переменных при помощи суммирующих, множительных и простейших функциональных устройств. In *Семинар по точности в машиностроении и приборостроении, вып. 8.* Труды ИМАШ, 1965.

[332] В. В. Подиновский. Многокритериальные задачи с упорядоченными по важности критериями. *Автоматика и телемеханика*, 11:118–127, 1976.

[333] Ю. В. Прохоров, editor. *Вероятность и математическая статистика: Энциклопедия.* М: Большая Российская энциклопедия, 1999.

[334] С. Р. Рао. *Линейные статистические методы и их применения.* М.: Наука, 1968.

[335] Н. С. Редькина. *Формализованные методы анализа документальных информационных потоков*. Библиосфера, 2005.

[336] К. В. Рудаков and И. Ю. Торшин. Об отборе информативных значений признаков на базе критериев разрешимости в задаче распознавания вторичной структуры белка. *Доклады Академии наук*, 441(1):1–5, 2011.

[337] К. В. Рудаков and И. Ю. Торшин. Анализ информативности мотивов на основе критерия разрешимости в задаче распознавания вторичной структуры белка. *Информатика и её применения*, 6(1):79–90, 2012.

[338] Г. И. Рудой and В. В. Стрижов. Упрощение суперпозиций элементарных функций при помощи преобразований графов по правилам. In *Интеллектуализация обработки информации. Доклады 9-й международной конференции*, pages 140–143, 2012.

[339] В. А. Садовничий. *Теория операторов*. Дрофа, 2001.

[340] Дж. Себер. *Линейный регрессионный анализ*. М.: Мир, 1980.

[341] В. В. Стрижов. *Согласование экспертных оценок для биосистем в экстремальных условиях*. М.: ВЦ РАН, 2002.

[342] В. В. Стрижов. Уточнение экспертных оценок с помощью измеряемых данных. *Заводская лаборатория. Диагностика материалов*, 72(7):59–64, 2006.

[343] В. В. Стрижов. Поиск параметрической регрессионной модели в индуктивно заданном множестве. *Вычислительные технологии*, 1:93–102, 2007.

[344] В. В. Стрижов. *Методы индуктивного порождения регрессионных моделей*. М.: ВЦ РАН, 2008.

[345] В. В. Стрижов and Т. В. Казакова. Устойчивые интегральные индикаторы с выбором опорного множества описаний. *Заводская лаборатория. Диагностика материалов*, 73(7):72–76, 2007.

[346] В. В. Стрижов and Г. О. Пташко. *Алгоритмы поиска суперпозиций при выборе оптимальных регрессионных моделей*. М.: ВЦ РАН, 2006.

[347] В. В. Стрижов and Р. А. Сологуб. Индуктивное порождение регрессионных моделей предполагаемой волатильности для опционных торгов. *Вычислительные технологии*, 14(5):102–113, 2009.

[348] А. Н. Тихонов. О решении некорректно поставленных задач и методе регуляризации. *Доклады академии наук СССР*, 151:501–504, 1963.

[349] А. Н. Тихонов and В. Я. Арсенин. *Методы решения некорректных задач*. М.: Наука, 1986.

[350] А. Н. Тырсин. Об эквивалентности знакового и наименьших модулей методов построения линейных моделей. *Обозрение прикладной и промышленной математики*, 12(4):879–880, 2005.

[351] А. Н. Тырсин. Исследование свойств обобщенного метода наименьших модулей (на примере оценки параметра сдвига). *Заводская лаборатория. Диагностика материалов*, 73(11):71–76, 2007.

[352] Дж. Форсайт and К. Молер. *Численное решение систем линейных алгебраических уравнений*. М.: Мир, 1969.

[353] С. Хайкин. *Нейронные сети, полный курс*. М: Вильямс, 2008.

[354] В. Хардле. *Прикладная непараметрическая регрессия*. М.: Мир, 1993.

[355] В. В. Шакин. Вычислительные процедуры для опознавания векторных функций. In *Опознавание и описание линий*, pages 58–77. М.: Наука, 1972.

[356] В. В. Шакин. *Методика и техника статистической обработки материалов социологических исследований идеологической работы*. Академия общественных наук при ЦК КПСС, 1972.

[357] А. Н. Ширяев. *Вероятность – 1*. МЦНМО, 2004.

[358] А. Н. Ширяев. *Основы стохастической финансовой математики*, volume 1. ФАЗИС, 2004.

[359] В. К. Шитиков, Г. С. Розенберг, and Т. Д. Зинченко. *Количественная гидроэкология: методы системной идентификации*. Тольятти: ИЭВБ РАН, 2003.

[360] А. М. Шурыгин. *Прикладная стохастика: робастность, оценивание, прогноз*. Финансы и статистика, 2000.