Построение оптимальных регрессионных моделей

В.В. Стрижов
Вычислительный центр им. А.А. Дородницына РАН

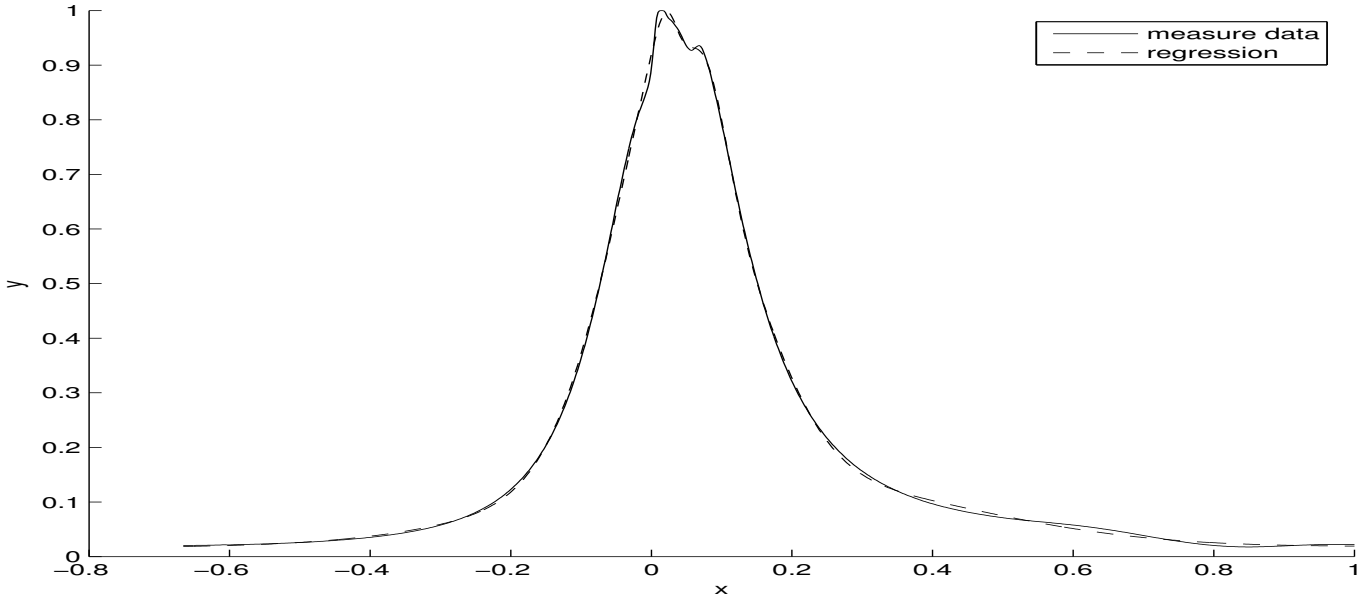
Практическая задача

Требуется найти закономерность (модель) наблюдаемого физического, социального, финансового явления исходя из

- результатов измерений вида «свободная переменная» — «зависимая переменная», например: «воздействие» — «отклик»;
- предположений относительно характера явления, например: «предполагается, что зависимость квадратичная»;
- предположений относительно характера измерений, например: «предполагается, что шум аддитивный гауссов».

Пример: давление в камере двигателя внутреннего сгорания

Выборка: один цикл работы дизельного двигателя Yamaha, давление в камере внутреннего сгорания. Модель — гладкая параметрическая функция одного аргумента.



Регрессионая модель имеет вид

$$y = f(\mathbf{w}, \mathbf{x}) + \varepsilon_{\mathbf{x}}$$

$\mathbf{w} \in \mathbf{R}^W$ — параметры модели, W — число параметров,

$\mathbf{x} \in \mathbf{R}^N$ — свободная переменная, N — число элементов вектора,

$y \in \mathbf{R}$ — свободная переменная,

$\varepsilon_{\mathbf{x}}$ — случайная переменная, в том же пространстве, что y,

f — функция, обладающая некоторыми известными свойствами.

Формальная постановка задачи

Задана выборка

$$D = \{(\mathbf{x}_i, y_i) | i = 1, ..., N\},$$

задана модель $f(\mathbf{w}, \mathbf{x})$, например,

$$f(\mathbf{w}, \mathbf{x}) = w_1 + w_2 \exp\left(-\frac{(x - w_3)^2}{w_4}\right)$$

задано распределение шума, например, гауссов аддитивный:

$$\mathcal{N}(y)$$

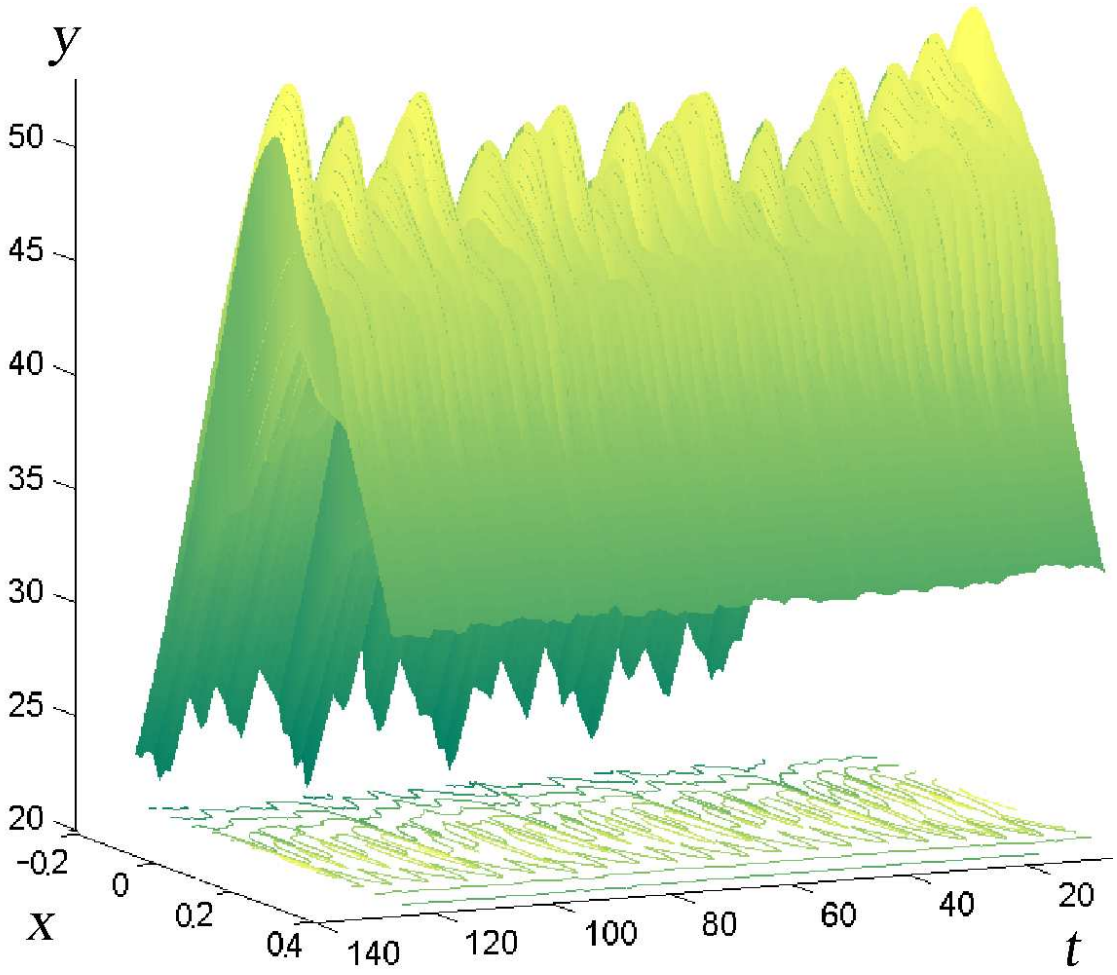
— *гипотеза порождения данных*.

Найти такие значения параметров $\mathbf{w}$, которые доставляли бы минимум заданной невязки:

$$\mathbf{w} = \arg\min \sum_{i=1}^{N} (y_i - f(\mathbf{w}, \mathbf{x}_i))^2$$

— *критерий оптимальности модели*.

Пример: давление в камере ДВС Lambordgini



Переобучение универсальных моделей

Теорема Колмогорова:

$$f(\mathbf{x}) = \sum_{q=1}^{2N+1} h_q \left(\sum_{n=1}^{N} g_q^n(x_n) \right),$$

Пример: однослойная нейронная сеть. При использовании универсальных моделей, оптимизация параметров может дать «переобученный» результат. Чтобы этого избежать, применяется регуляризация.

Качество модели = невязка + штраф за большие значения параметров

Метод группового учета аргументов

Требуется выбрать наиболее простую модель в классе заданных моделей. Для выбора предлагается разбить множество D на обучающую и контрольную выборку, и использовать контрольную выборку для выбора модели.

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k + \dots.$$

Возможны подстановки:

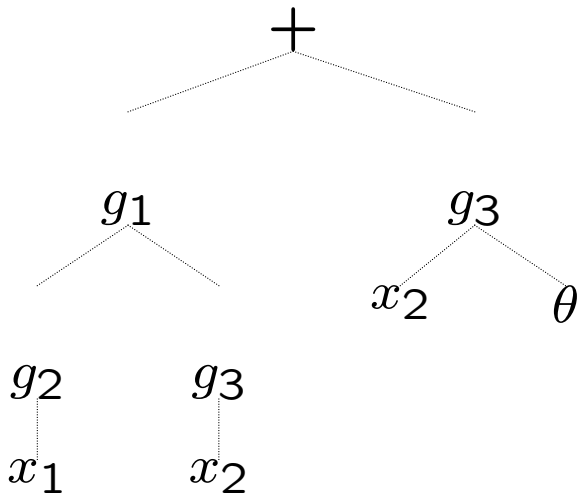
$x \mapsto \sin x$

$x \mapsto e^x$

...

Модель как произвольная суперпозиция

Пример:



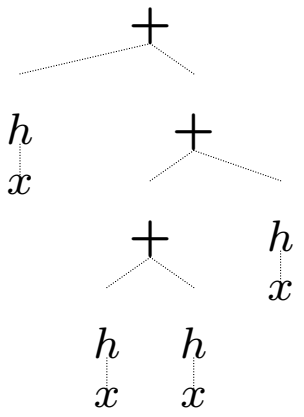
$$f \equiv g_1\left(g_2(x_1), g_3(x_2)\right) + g_4(x_2, \theta)$$

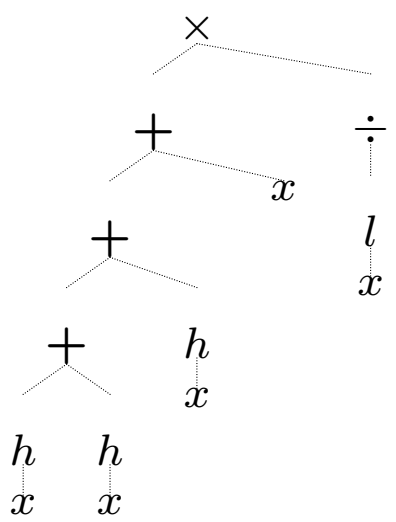
Список порождающих функций

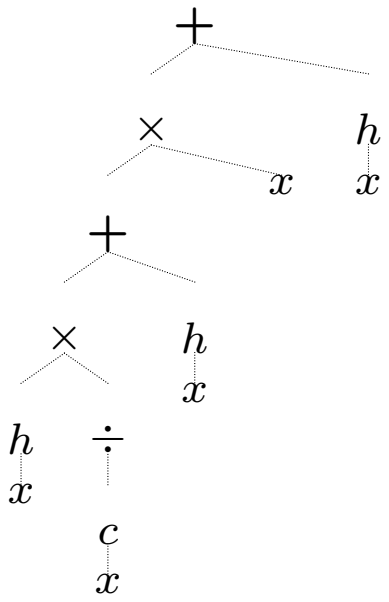
№	Функция	Описание	Параметры
Функции двух переменных аргументов, $g(\mathbf{b}, x_1, x_2)$			
1	plus	$y = x_1 + x_2$	–
2	times	$y = x_1 x_2$	–
3	divide	$y = x_1 / x_2$	–
Функции одного переменного аргумента, $g(\mathbf{b}, x_1)$			
4	multiply	$y = ax$	a
5	add	$y = x + a$	a
6	gaussian	$y = \frac{\lambda}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, 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	$y = \frac{\lambda}{1+\exp(-\sigma(x-\xi))} + a$	λ, σ, ξ, a

Поиск оптимальной регрессионной модели

f	$+(h,h,h,h)$	$\times(\div l,+(h,h,h,e))$	$+(\times(+(\times(h,\div h),h),e),h)$
mse$(y-\widehat{y})$	0.0034	0.0037	0.0035
max$(y-\widehat{y})$	0.0421	0.0325	0.00338
#w	16	16	16







Легенда: h — gaussian, c — cubic, l — logsig

Для получения модели было порождено более 10 000 различных семейств функций. Выбрана модель №2.

$$y = (ax + b)^{-1} \left(x + \sum_{i=1}^{3} \frac{\lambda_i}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - \xi_i)^2}{2\sigma_i^2} \right) + a_i \right).$$

Задача поиска оптимальной регрессионной модели

Задача 1 (оптимизации):
Фиксируем модель, фиксируем гипотезу порождения данных, получаем модель из семейства функций.

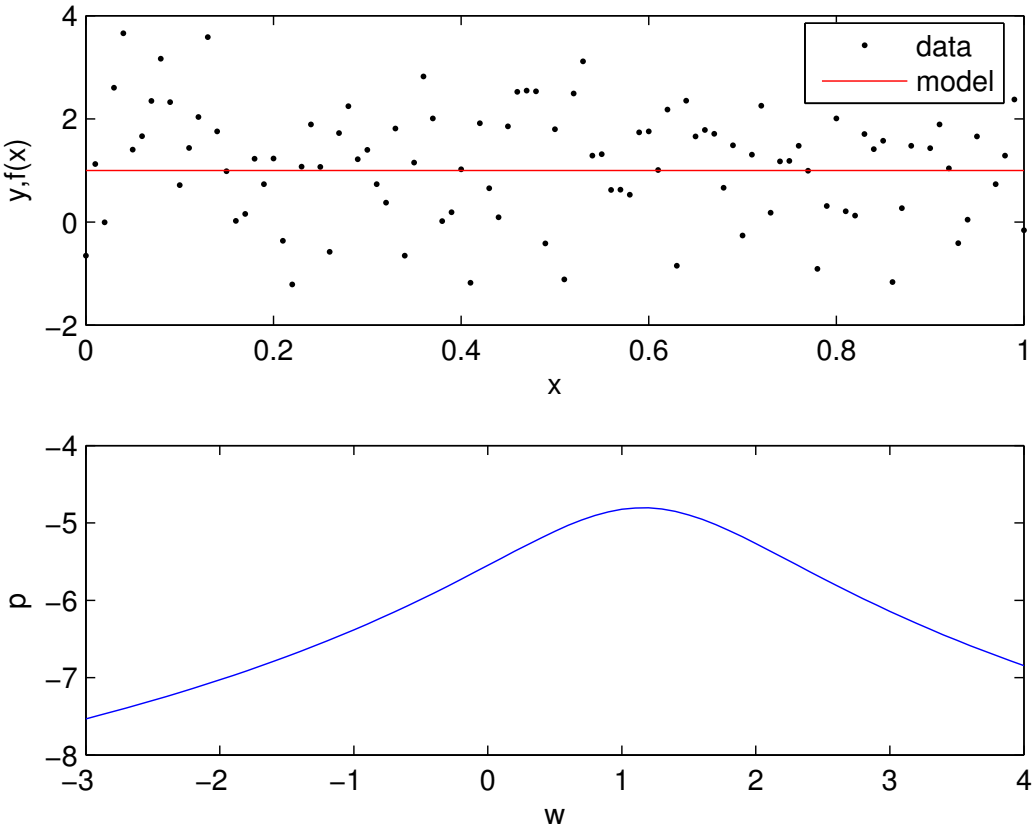
Задача 2 (группового учета аргуметов):
Фиксируем гипотезу порождения данных, получаем модель из множества семейств функций — претендентов.

Задача 3 (совместного поиска):
Получаем множество моделей в пространстве гипотез порождения данных.

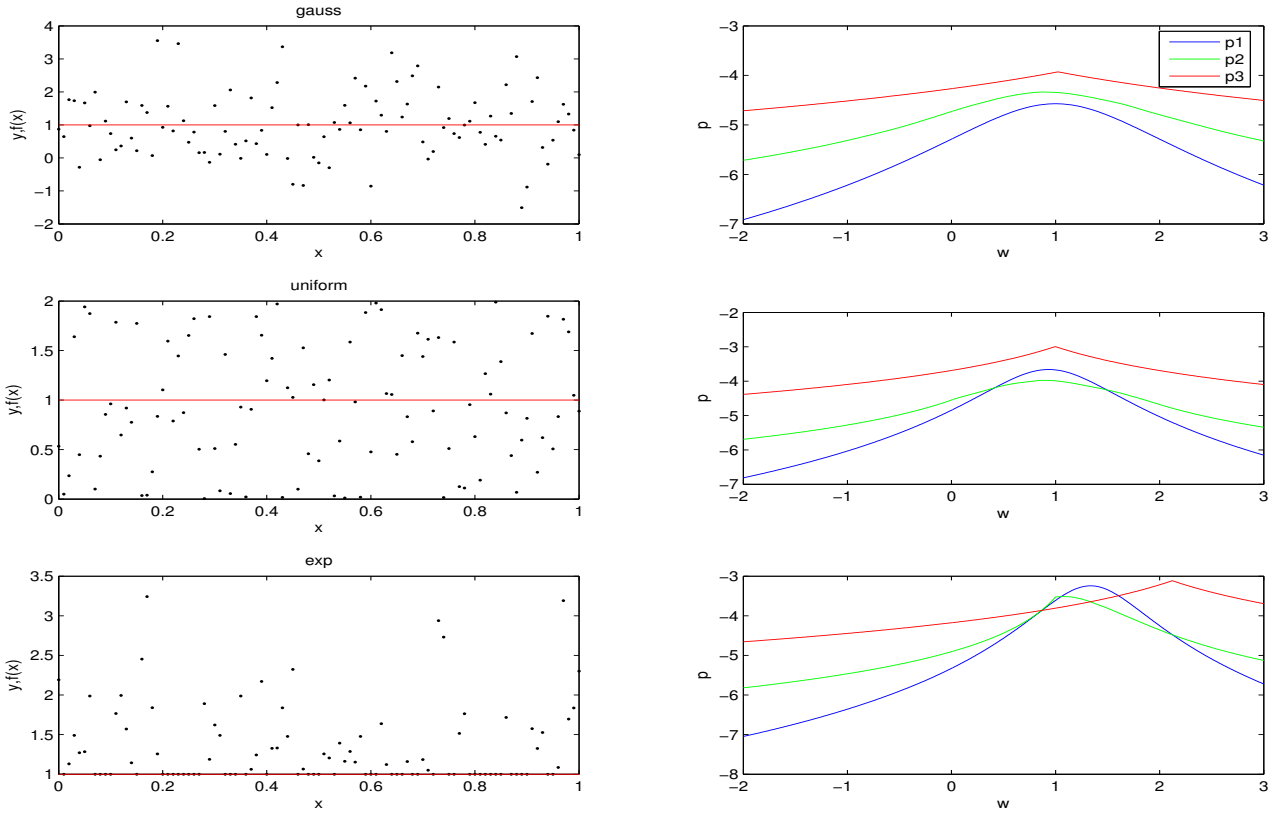
Гипотеза порождения данных: альтернативы

Часто невозможно указать гипотезу распределения случайной переменной. От того, как распределена эта переменная зависит функционал качества модели.

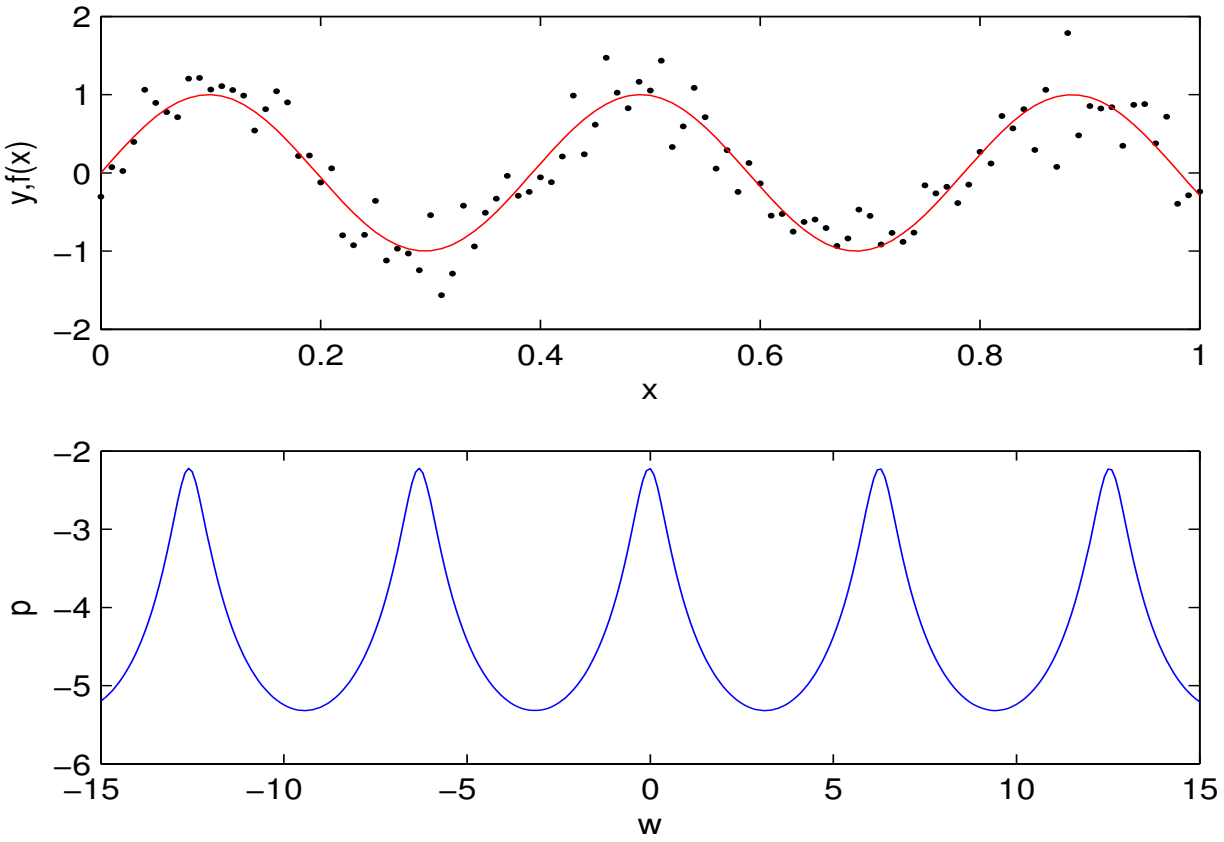
Пример 1. Пространство параметров $W \ni \mathbf{w}$



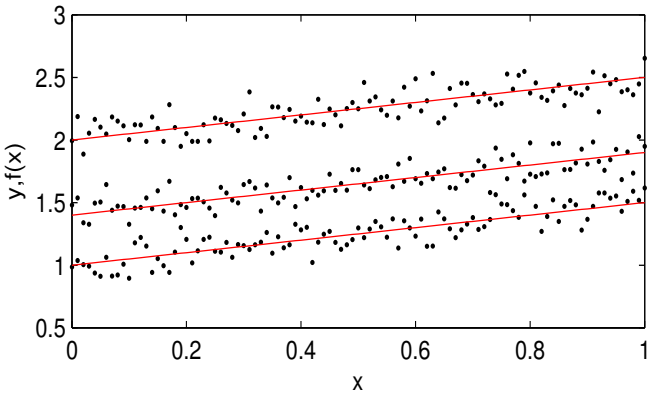
Пример 2. $\mathbf{w} = \arg\max_{\mathbf{w} \in W} p(\mathbf{w}|D, f), p \in \mathcal{P}$

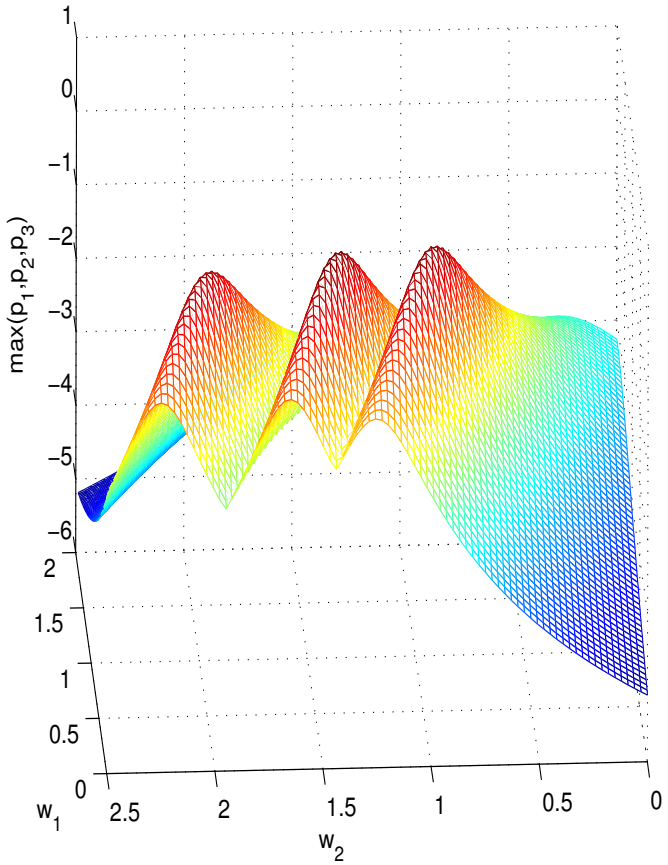


Пример 3. Инвариант на заданном D

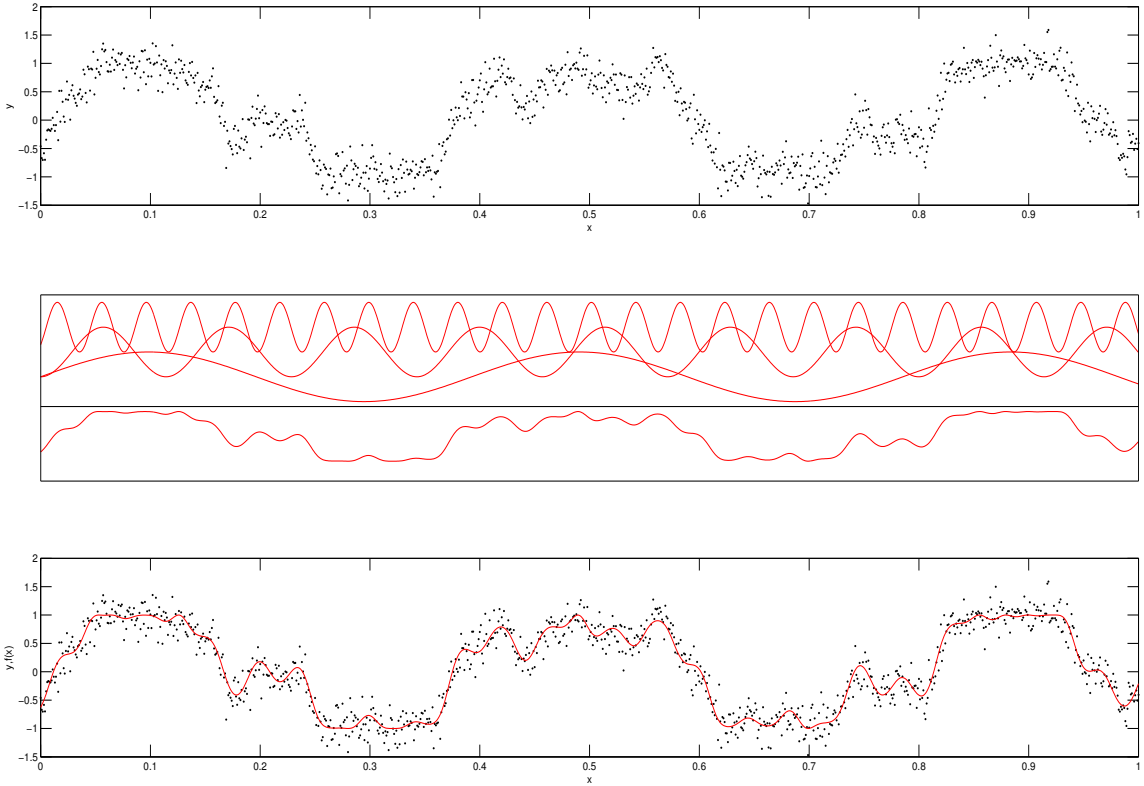


Пример 4. Инвариант на D_1, D_2, D_3

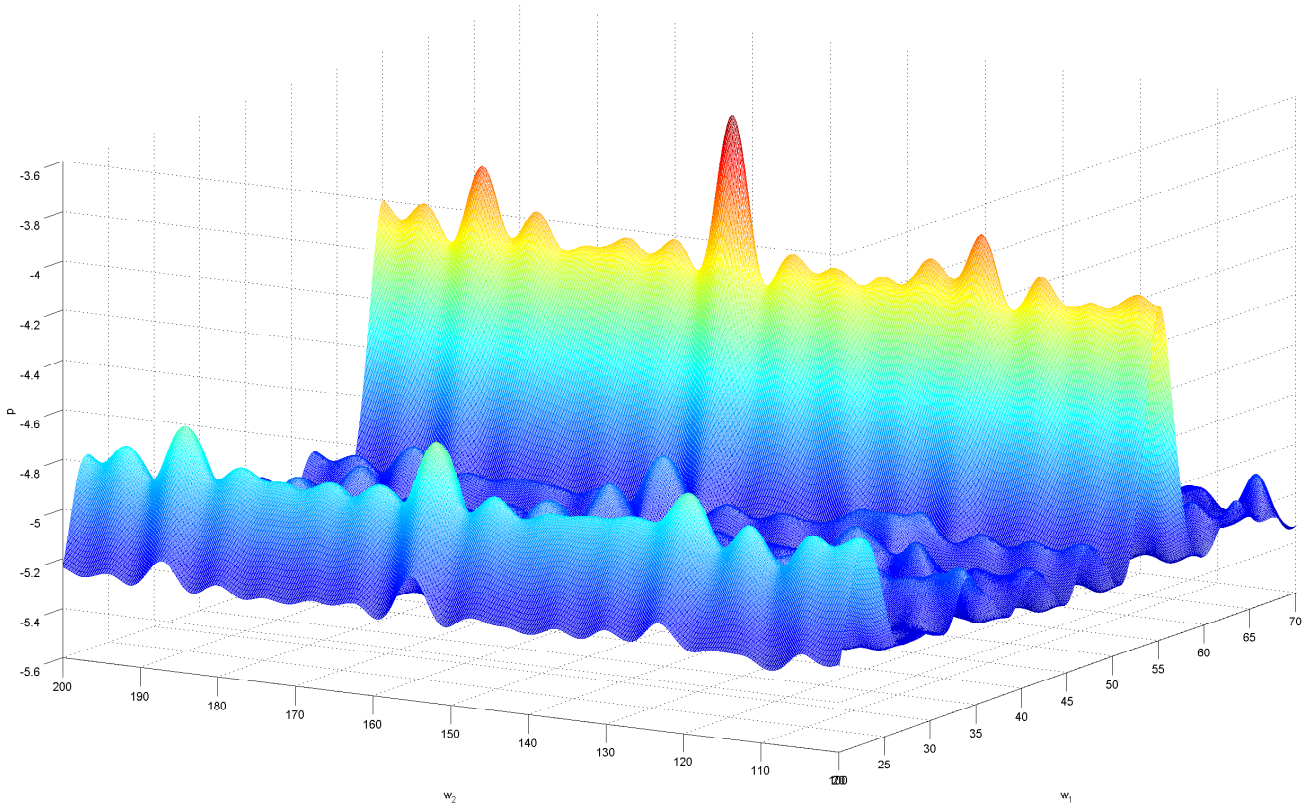




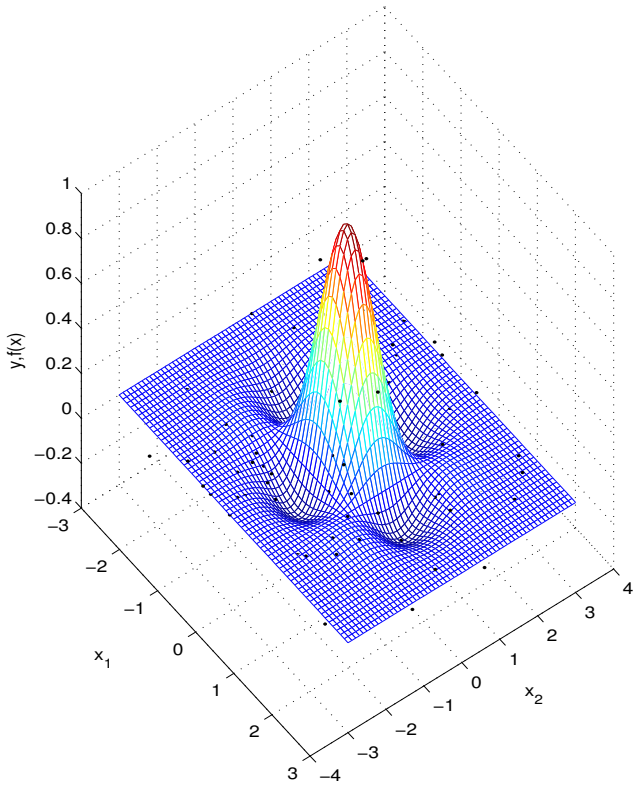
Пример 5a. Устойчивость элементов модели

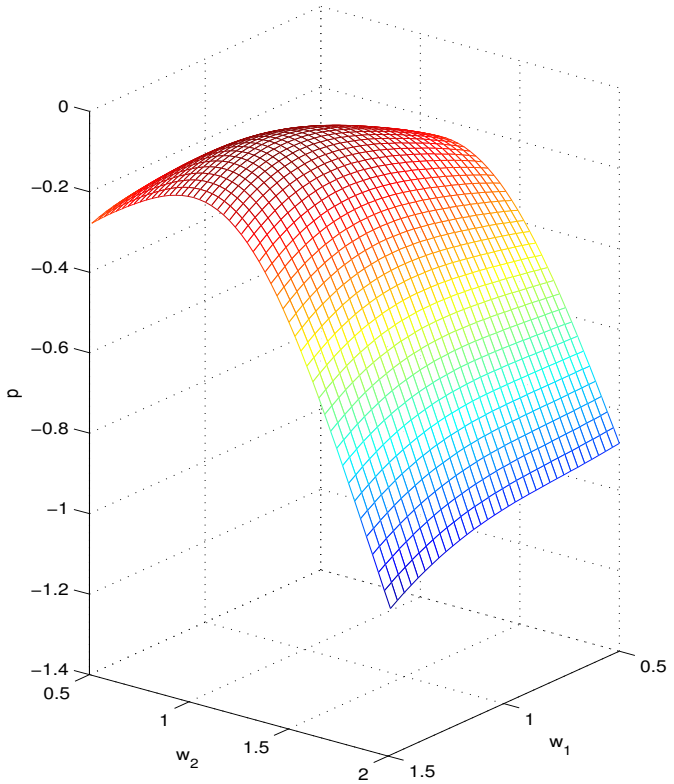


Пример 5b. Устойчивость элементов модели



Пример 5c. Устойчивость элементов, $\mathbf{x} \in \mathbb{R}^2$





Метод автоматического поиска оптимальной модели использовался в проектах:

1. Поиск оптимальной модели давления в камере внутреннего сгорания дизельного двигателя

2. Прогноз концентрации кислорода в выхлопных газах двигателя внутреннего сгорания

3. Поиск модели оценки справедливой стоимости опционов

4. Исследование зависимости геометрического (конформационного) состояния белков от физико-химических свойств аминокислотных остатков

5. Поиск наиболее информативного набора маркеров пациентов CVD

6. Поиск стратегий поведения крупных игроков на мировых финансовых рынках