ПОИСК РЕГРЕССИОННЫХ МОДЕЛЕЙ В ИНДУКТИВНО ЗАДАННОМ МНОЖЕСТВЕ

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

Дано

$X = \{\mathbf{x}_1, ..., \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^T\}$ — множество временных рядов,
$\mathbf{x}_i = \{x_{it} | t = 1, ..., T\}$.

Задано множество параметрических гладких функций
$G = \{g | g : \mathbb{R} \times ... \times \mathbb{R} \longrightarrow \mathbb{R}\}$,
$g = g(\mathbf{b}, \cdot, \cdot, ..., \cdot)$. и G задает множество суперпозиций
$\mathcal{F} = \{f(\mathbf{w}) : t \longrightarrow t\}$ путем обобщенного индуктивного определения. Вектор параметров $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots ... \vdots \mathbf{b}_r, \mathbf{w} \in \mathbb{R}^W$.

Найти

Такой инвариант $f(\mathbf{w}) \in \mathcal{F}$ и такое множество значений его параметров $W \ni \mathbf{w}$, которые задают требуемый класс эквивалентности $[\mathbf{x}]$ на множестве X.

Выбор инварианта определяется суммарным качеством аппроксимации f пути наименьшей стоимости на всех парах временных рядов из $[\mathbf{x}]$.

Качество аппроксимации задано функцией $S = p(\mathbf{w}|\mathbf{x}_i, \mathbf{x}_j, f)$.

Нахождение пути наименьшей стоимости

Построим матрицу расстояний $\Omega(\mathbf{x}_i, \mathbf{x}_j)$ между всеми парами элементов (x_{ip}, x_{jq}) временных рядов $\mathbf{x}_i, \mathbf{x}_j$:
$\Omega(\mathbf{x}_i, \mathbf{x}_j) = \{\omega_{pq} = S(x_{ip}, x_{jq}) | p, q = 1, ..., T\}$.

Рассмотрим в некоторой матрице Ω все пути $\mathbf{s} = \{s_1, ..., s_K\}$ такие, что $s_1 = \omega_{11}$, $s_K = \omega_{T,T}$ и для произвольного $s_k = \omega_{pq}$, где $k = 1, ..., K-1$ значение $s_{k+1} = \omega_{p+u,q+v}$, где $u + v \in \{1, 2\}$.

Стоимость пути равна $\mathbf{s} = K^{-1} \sum_{k=1}^{K} s_k$. Обозначим $\bar{\mathbf{s}}_{ij}$ — путь наименьшей стоимости для пары временных рядов $(\mathbf{x}_i, \mathbf{x}_j)$.

Аппроксимация пути

Найдем такие параметры $\mathbf{w}_{ij}$ произвольной модели $f \in \mathcal{F}$, при которых эта модель наилучшим образом приближает путь $\bar{\mathbf{s}}_{ij}$. Оптимальные значения параметров заданы как $\bar{\mathbf{w}}_{ij} = \arg\min S\Big(f(\mathbf{w}_{ij}), \mathrm{cod}(\bar{\mathbf{s}}_{ij})\Big)$.

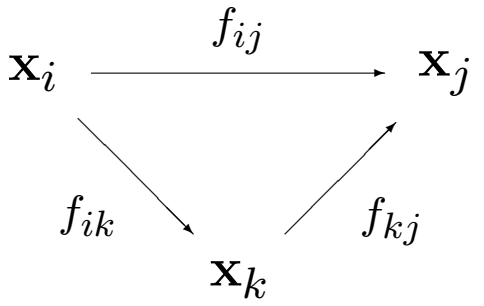
Будем считать искомой моделью ту, для которой сумма значений S по всем парам $(\mathbf{x}_i, \mathbf{x}_j)$ минимальна. $\sum\limits_{i=1}^{N-1} \sum\limits_{j=i}^{N} S\Big(f(\bar{\mathbf{w}}_{ij}), \mathrm{cod}(\bar{\mathbf{s}}_{ij})\Big) \longrightarrow \min$.

Гипотеза $\mathcal{H}$

Функции f, аппроксимирующие путь наименьшей стоимости $\bar{\mathbf{s}}_{ij}$ произвольной пары $(\mathbf{x}_i, \mathbf{x}_j)$ из X являются пучком. Существует отображение которое параметрическому семейству функций f, аппроксимирующих путь $\bar{\mathbf{s}}$ ставит в соответствие параметрическое семейство моделей ϕ, аппроксимирующих ряды $\mathbf{x}$.

Группа преобразований $\{f_{ij}|f : t \longrightarrow t\}$

Для произвольной тройки, удовлетворяющей гипотезе $\mathcal{H}$, справедлива коммутативная диаграмма



и $\quad \mathcal{G} = \begin{pmatrix} \mathsf{id} & f_{12} & \cdots & f_{1N} \\ f_{12}^{-1} & \mathsf{id} & \cdots & f_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1N}^{-1} & f_{2N}^{-1} & \cdots & \mathsf{id} \end{pmatrix}$

является группой, $i, j, k = 1, ..., N$.

NB: В общем случае $\mathsf{dom}\Big(f(\bar{\mathbf{w}})(\mathsf{cod}(\mathbf{x}_i))\Big) \neq \mathbf{x}_j$, так как $f_{ij} : t \longrightarrow t$ аппроксимирует путь $\mathbf{s}_{ij}$, но для пары $(\mathbf{x}_i, \mathbf{x}_j)$ функция f_{ij} является морфизмом.

Классы эквивалентности в пространстве параметов

Множество последовательных отображений $f_{12}, f_{23}, ..., f_{(\ell-1)\ell}$ задает множество фиксированных параметров $\mathbf{w}_{12}, \mathbf{w}_{23}, ..., \mathbf{w}_{(\ell-1)\ell}$, полученных в результате идентификации.

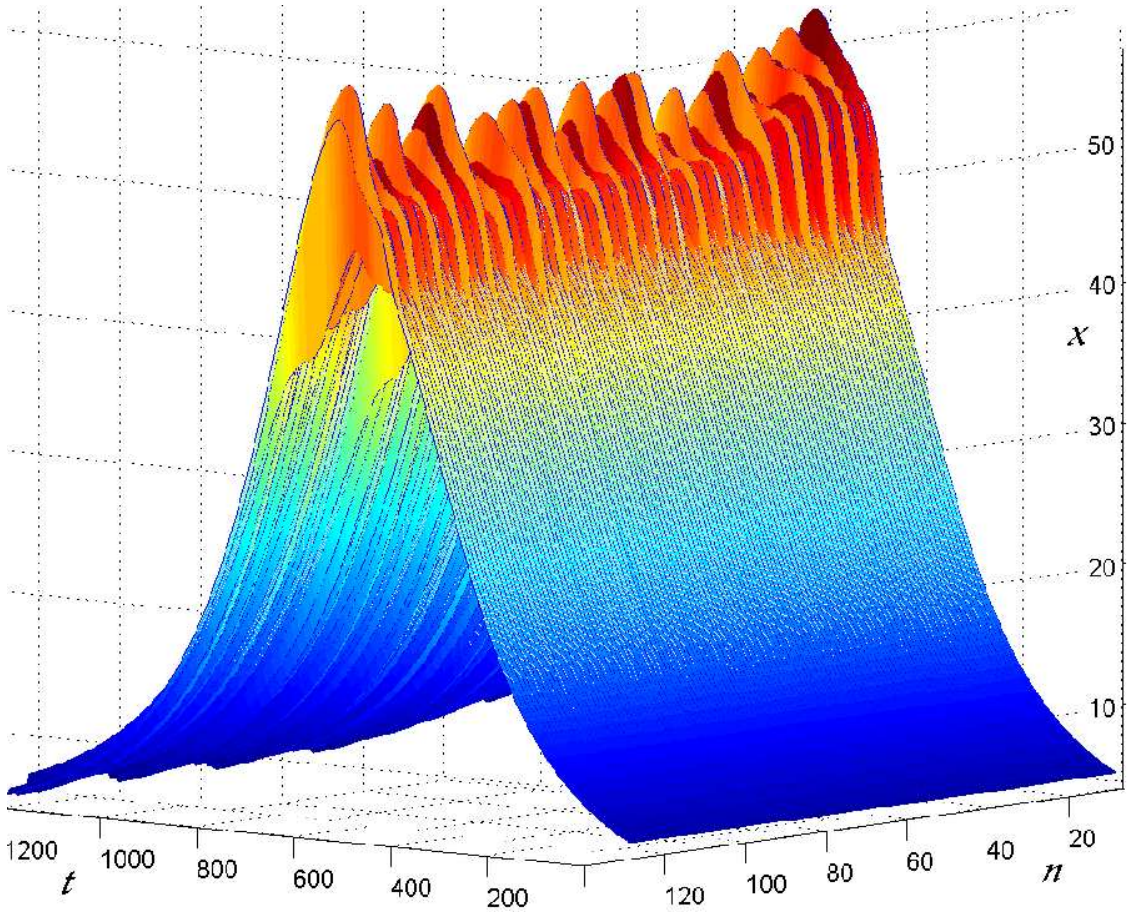
Согласно гипотезе $\mathcal{H}$, класс эквивалентности $[\mathbf{w}] \ni \{\mathbf{w}_{ij}\}$. Определим $[\mathbf{w}]$ как выпуклую оболочку $\{\mathbf{w}_{ij}\}$.

Выбор функции f из $\mathcal{F}$ и кластеризация путем проверки гипотезы $\mathcal{H}$

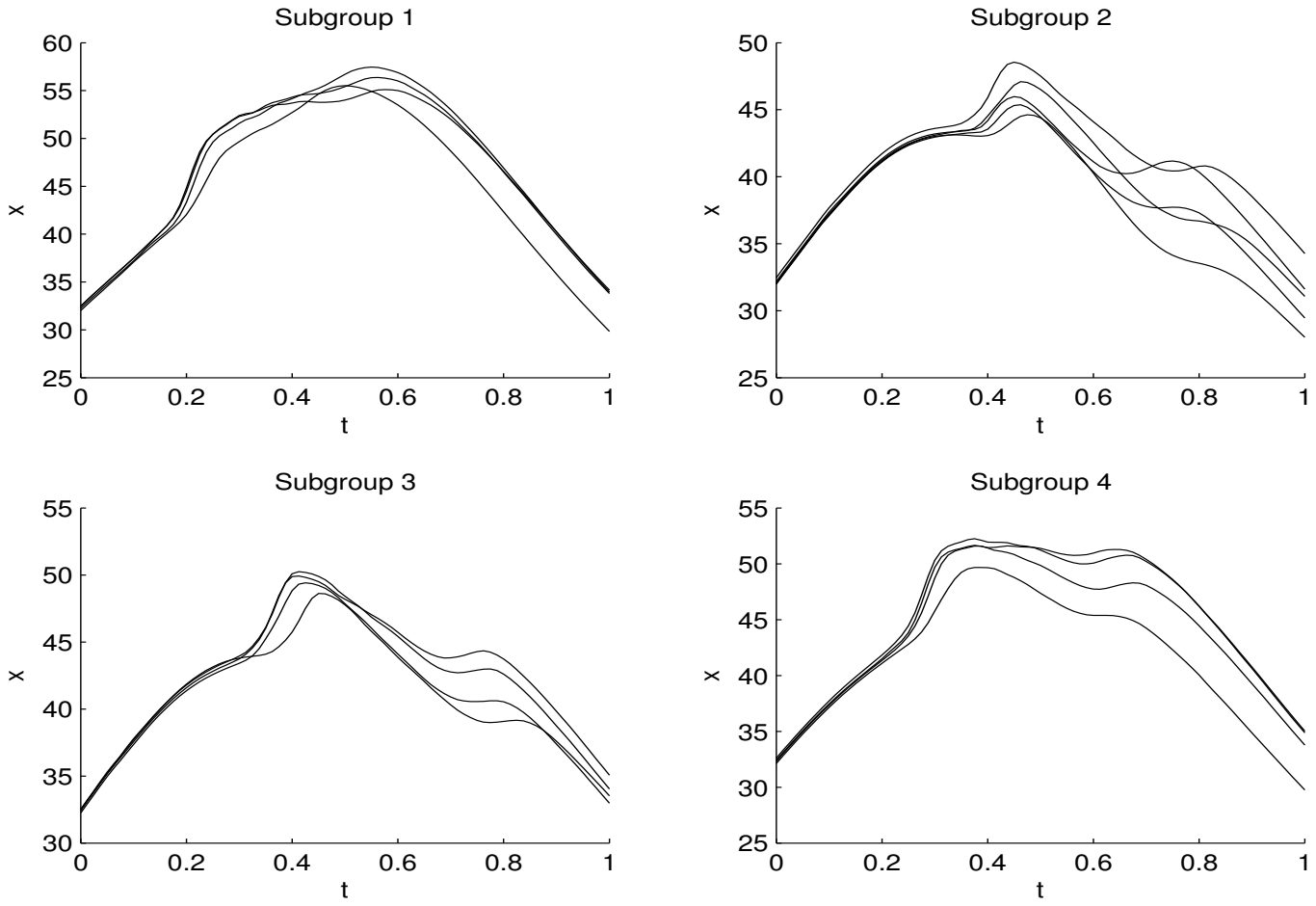
Выберем такую функцию f, доставляющую суммарную минимальную ошибку S, при которой параметры, полученные в результате идентификации на всех рядах, удовлетворяющих гипотезе $\mathcal{H}$, попадают в заданный класс эквивалентности (тривиальная корректность).

Для каждого нового временного ряда $\mathbf{x}_k$ находим $\{\mathbf{w}_{ik}\}$ для $i \in \{1, ..., N\}$. Если полученные векторы $\mathbf{w}_{ik} \in [\mathbf{w}]$, классифицируем новый ряд как удовлетворяющий гипотезе $\mathcal{H}$.

Пример: давление в камере ДВС



Полученная кластеризация кривых



Заключение

- Вычисление значения стоимости оптимального пути DTW не позволяет решить задачу кластеризации исследуемых временных рядов, так как стоимость двух несовпадающих оптимальных путей пар временных рядов из разных кластеров часто оказывается одинаковой.
- Создание моделей, аппроксимирующих исследуемые временные ряды не позволяет решить данную задачу, так как кластеризацию при этом приходится выполнять в пространстве большой размерности.
- Предложенный метод позволяет разделить данные временные ряды на классы, так как кластеризация выполняется в пространстве параметров монотонных функций, и это пространство имеет небольшую размерность.