

# ПОИСК РЕГРЕССИОННЫХ МОДЕЛЕЙ В ИНДУКТИВНО ЗАДАННОМ МНОЖЕСТВЕ

Стрижов В.В.

Вычислительный центр им. А.А. Дородницына РАН

## Дано

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

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

## Найти

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

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

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

## Нахождение пути наименьшей стоимости

Построим матрицу расстояний  $\Omega(x_i, x_j)$  между всеми парами элементов  $(x_{ip}, x_{jq})$  временных рядов  $x_i, x_j$ :

$$\Omega(x_i, x_j) = \{\omega_{pq} = S(x_{ip}, x_{jq}) | p, q = 1, \dots, T\}.$$

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

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

## Аппроксимация пути

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

Будем считать искомой моделью ту, для которой сумма значений  $S$  по всем парам  $(x_i, x_j)$  минимальна.

$$\sum_{i=1}^{N-1} \sum_{j=i}^N S(f(\bar{w}_{ij}), \text{cod}(\bar{s}_{ij})) \longrightarrow \min.$$

## Гипотеза $\mathcal{H}$

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

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

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

$$\begin{array}{ccc}
 \mathbf{x}_i & \xrightarrow{f_{ij}} & \mathbf{x}_j \\
 & \searrow f_{ik} & \\
 & & \mathbf{x}_k \\
 & & \nearrow f_{kj}
 \end{array}
 \quad \text{и} \quad
 \mathcal{G} = \begin{pmatrix}
 \text{id} & f_{12} & \cdots & f_{1N} \\
 f_{12}^{-1} & \text{id} & \cdots & f_{2N} \\
 \vdots & \vdots & \ddots & \vdots \\
 f_{1N}^{-1} & f_{2N}^{-1} & \cdots & \text{id}
 \end{pmatrix}$$

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

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

## Классы эквивалентности в пространстве параметров

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

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

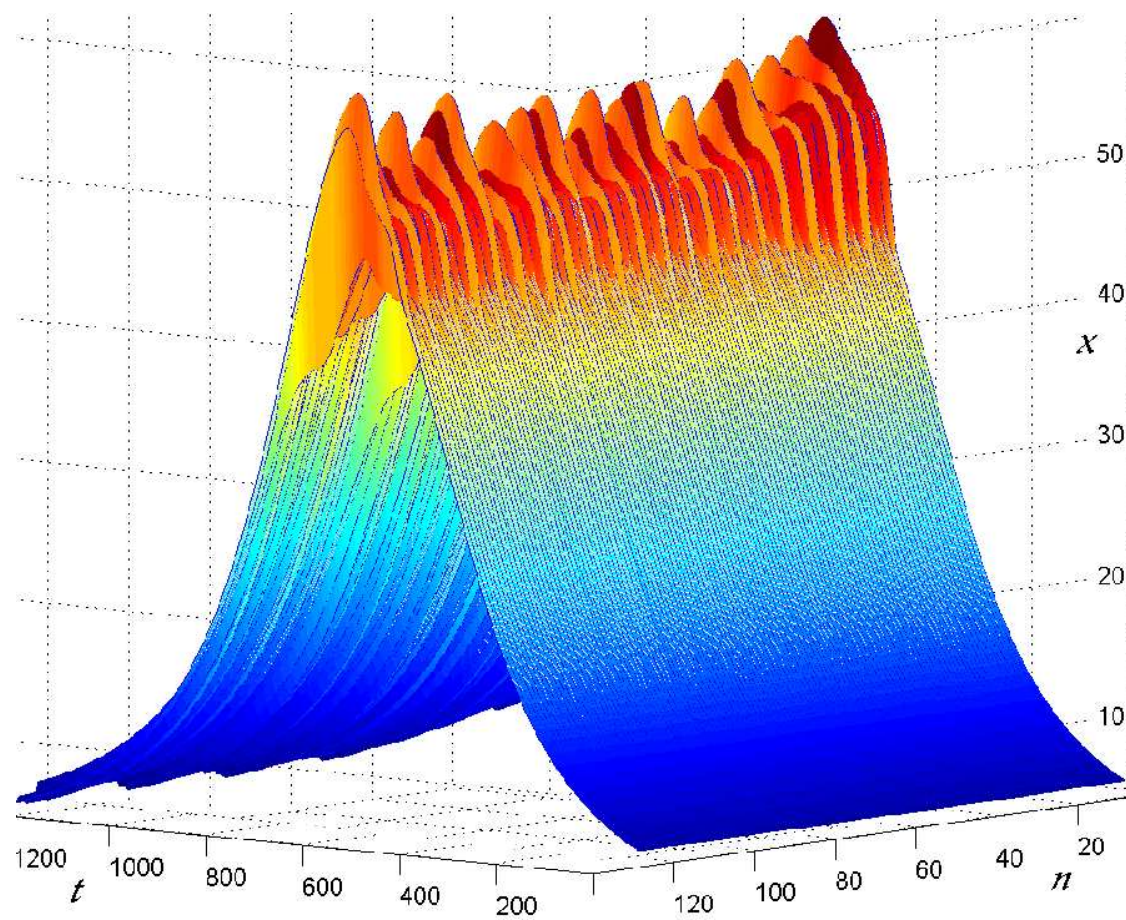


## Выбор функции $f$ из $\mathcal{F}$ и кластеризация путем проверки гипотезы $\mathcal{H}$

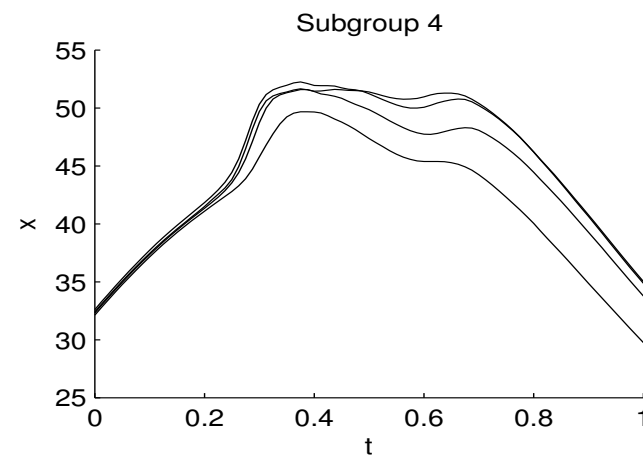
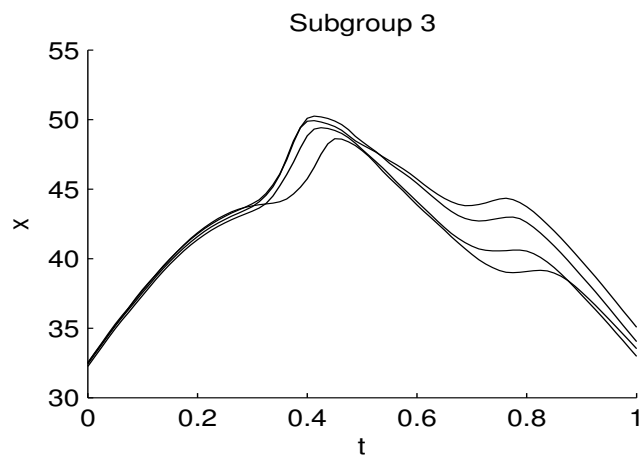
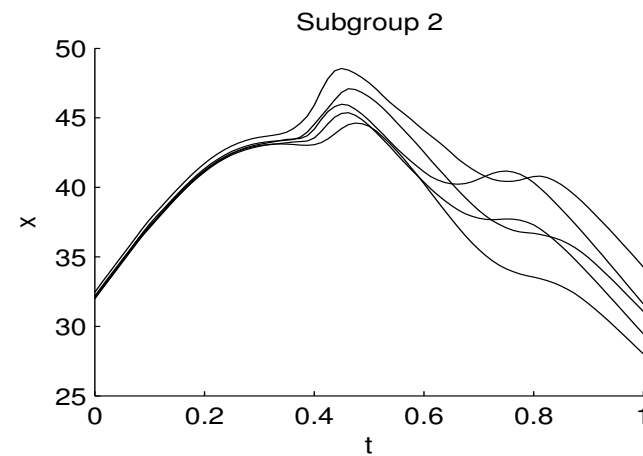
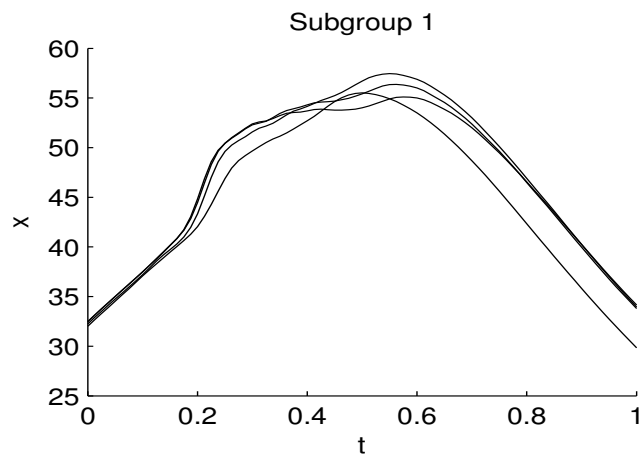
Выберем такую функцию  $f$ , доставляющую суммарную минимальную ошибку  $S$ , при которой параметры, полученные в результате идентификации на всех рядах, удовлетворяющих гипотезе  $\mathcal{H}$ , попадают в заданный класс эквивалентности (тривиальная корректность).

Для каждого нового временного ряда  $x_k$  находим  $\{w_{ik}\}$  для  $i \in \{1, \dots, N\}$ . Если полученные векторы  $w_{ik} \in [w]$ , 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.