Оценка параметров инвариантных преобразований в задачах прогнозирования временных рядов

Д. С. Кононенко

Научный руководитель
к.ф.-м.н. В. В. Стрижов

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

20 июня
Москва

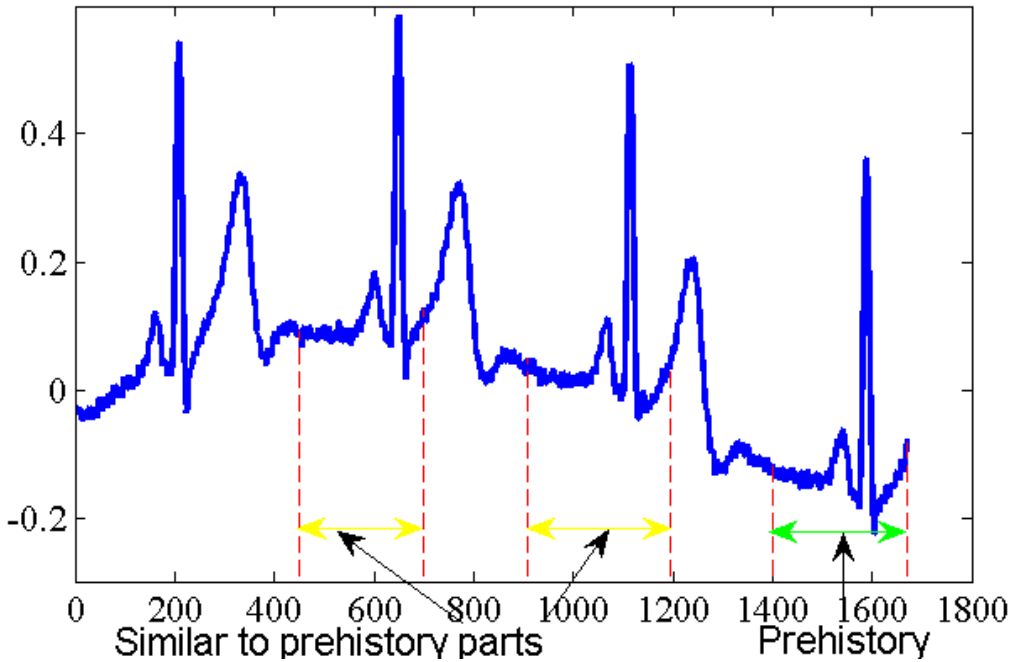
Цели работы

- Предложить метод локального прогнозирования временных рядов, в котором функция расстояния между участками временного ряда основана на параметрических моделях этих участков и инвариантна относительно определенного класса преобразований.
- Предложить метод совместной оценки параметров инвариантных преобразований и параметров моделей.

Основная литература:

1. J. McNames. Innovations in local modeling for time series prediction. PhD thesis, Stanford University, 1999.
2. A. Kneip and J. Engel. Model estimation in nonlinear regression under shape invariance. The Annals of Statistics, 23(2):551–570, 1995.
3. А. А. Токмакова, В. В. Стрижов. Оценивание гиперпараметров линейных регрессионных моделей при отборе шумовых и коррелирующих признаков. Информатика и её применения, 6(4):66–75, 2012.

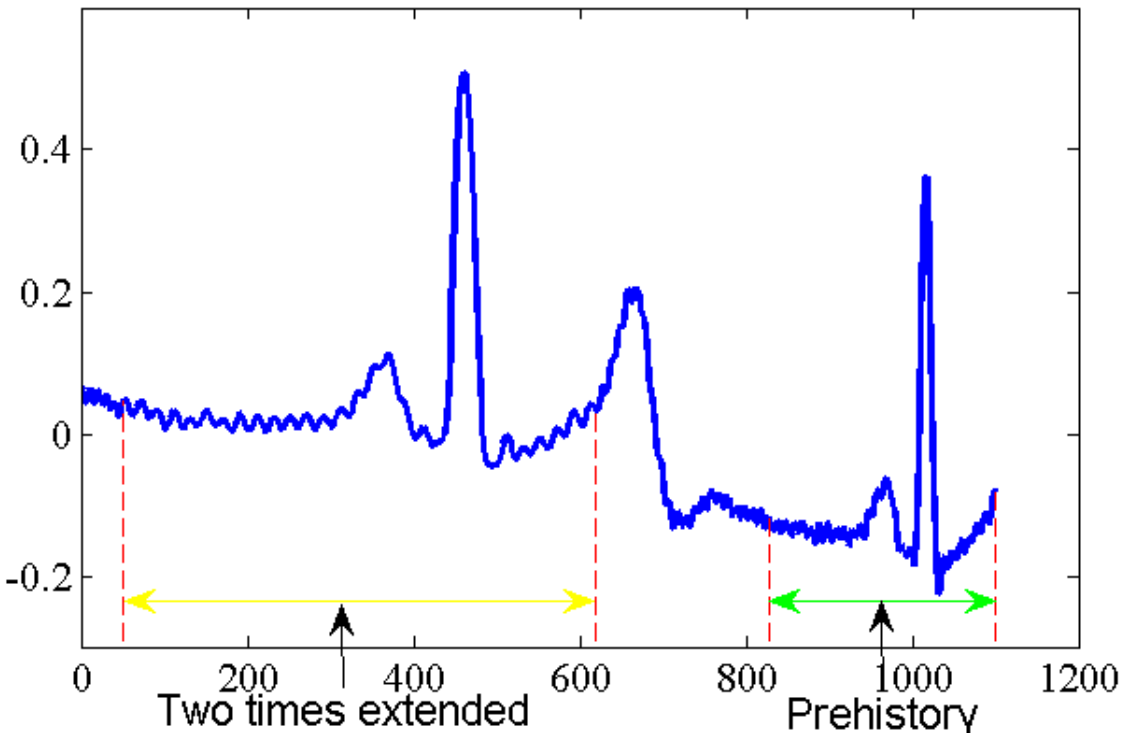
Локальное прогнозирование



Временной ряд: $[t_1, t_2, \ldots, t_N]^T \to [y_1, y_2, \ldots, y_N]^T$.
Предыстория: $[y_{N-L+1}, y_{N-L+2}, \ldots, y_N]^T$.
Функция расстояния:

$$D([x_1, \ldots, x_l]^T, [y_1, \ldots, y_l]^T) = \sqrt{\sum_{i=1}^{l} \lambda_i (x_i - y_i)^2}.$$

Преобразования участков временных рядов



Некоторый класс преобразований $f([x_1, \ldots, x_l]^T)$.
Выбирается оптимальное преобразование из класса $D([y_{N-L+1}, y_{N-L+2}, \ldots, y_N]^T, f([x_1, \ldots, x_l]^T))$.

Предположения о данных

Дано:

- регрессионные подвыборки $D_1, \ldots, D_K$:

$$D_k = (X_k, \mathbf{y}_k) = \{(\mathbf{x}_{ki}, y_{ki})\}_{i=1}^{n_k},$$

где $\mathbf{x}_{ki} \in \mathbb{R}^n$, $y_{ki} \in \mathbb{R}$, $k \in \{1, \ldots, K\}$;

- подвыборки делятся на C кластеров;
- прогностическая модель

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

где $\mathbf{w} \in \mathbb{R}^W$, $f : \mathbb{R}^n \to \mathbb{R}$;

- модель кластера

$$\mu_c(\mathbf{x}) = f(\mathbf{x}, \mathbf{w}_c).$$

Инвариантные преобразования

- Инвариантное преобразование — функционал, преобразующий модель:

$$g : (\mathbb{R}^n, \mathbb{R}) \longrightarrow (\mathbb{R}^n, \mathbb{R}).$$

- Параметрическое семейство инвариантных преобразований:

$$\mathcal{G} = \{g_{\boldsymbol{\theta}}(\mathbf{x}, f(\mathbf{x})) = g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) | \boldsymbol{\theta} \in \mathbb{R}^p\}.$$

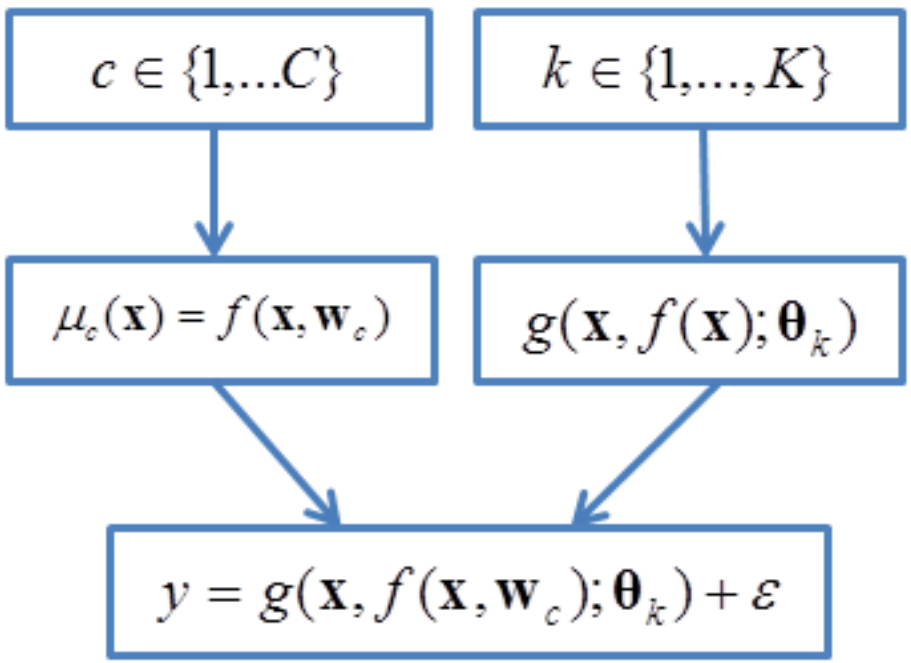
- Пример: семейство $g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) = (\mathbf{x}, f(\mathbf{x} - \boldsymbol{\theta}))$, $\boldsymbol{\theta} \in \mathbb{R}^n$ — всевозможные сдвиги.

Модель порождения данных

- Каждой подвыборке D_k поставим в соответствие инвариантное преобразование с параметрами $\boldsymbol{\theta}_k$.
- Модель порождения данных

$$\mathbf{y} = \mathbf{g}(\mathbf{X}, \boldsymbol{\mu}_c(\mathbf{X}, \mathbf{w}); \boldsymbol{\theta}) + \boldsymbol{\varepsilon},$$

$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{B}^{-1})$.



Прогнозирование временного ряда: задача оценки параметров и кластеризации подвыборок

Дано:

- $D_1, \ldots, D_K$ — подвыборки;
- $f(\mathbf{x}, \mathbf{w})$ — прогностическая модель;
- $g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta})$ — семейство инвариантных преобразований;
- C — число кластеров.

Требуется найти:

- $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ — параметры инвариантных преобразований;
- $\mathbf{w}_1, \ldots, \mathbf{w}_C$ — параметры моделей кластеров;
- $c_1, \ldots, c_K$ — метки кластеров подвыборок.

Получаем модель предыстории и прогноз в будущий момент времени t:

$$y = \mathbf{g}(\mathbf{x}, f(\mathbf{x}, \mathbf{w}_c); \boldsymbol{\theta}_k), \quad y_t = \mathbf{g}(t, f(t, \mathbf{w}_c); \boldsymbol{\theta}_k),$$

c — номер кластера предыстории, k — номер подвыборки предыстории.

Оценка параметров инвариантных преобразований

- Подвыборка $D_k = (\mathbf{X}_k, \mathbf{y}_k)$, $\mathbf{y}_k = [y_{k1}, y_{k2}, \ldots, y_{kn_k}]^T$.
- Фиксированы параметры $\mathbf{w}$ модели $f(\mathbf{x}, \mathbf{w})$.
- Обозначим $\hat{y}_{ki}(\boldsymbol{\theta}_k) = g(f(\mathbf{x}, \mathbf{w}_c); \boldsymbol{\theta}_k)$.
- Оптимизируем квадратичную ошибку

$$S_{LS}(\boldsymbol{\theta}_k) = \sum_{i=1}^{n_k} (y_{ki} - \hat{y}_{ki}(\boldsymbol{\theta}_k))^2.$$

- Параметры инвариантного преобразования k-ой подвыборки

$$\boldsymbol{\theta}_k^* = \arg \min_{\boldsymbol{\theta}_k \in \mathbb{R}^p} S_{LS}(\boldsymbol{\theta}_k).$$

Функция расстояния в пространстве параметров

- $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|D, \mathbf{A}, \mathbf{B}, \boldsymbol{\theta}, f)$.
- $[\hat{\mathbf{A}}, \hat{\mathbf{B}}] = \arg\max_{\mathbf{A},\mathbf{B}} p(D|\mathbf{A}, \mathbf{B}, \boldsymbol{\theta}, f)$.
- Расстояние между подвыборками — расстояние между апостериорными распределениями параметров $p(\mathbf{w}_i|D_i, \hat{\mathbf{A}}_i, \hat{\mathbf{B}}_i, \boldsymbol{\theta}_i, f)$ и $p(\mathbf{w}_j|D_j, \hat{\mathbf{A}}_j, \hat{\mathbf{B}}_j, \boldsymbol{\theta}_j, f)$. Используется расстояние Йенсена-Шеннона.
- Зная попарное расстояние, можно кластеризовать подвыборки.
- Беря несколько подвыборок из одного кластера, можно оценить распределение параметров модели кластера $p(\mathbf{w}_c|D_{i_1}, \ldots, D_{i_c}, \mathbf{A}_c, \mathbf{B}_c, \boldsymbol{\theta}_{i_1}, \ldots, \boldsymbol{\theta}_{i_c}, f)$.

Алгоритм кластеризации подвыборок и оценки параметров

Итеративно повторяются шаги:

1. для каждой подвыборки D_k при фиксированном инвариантном преобразовании $g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta}_k)$ оцениваются распределения параметров модели $p(\mathbf{w}_k)$;
2. подвыборки $D_1, \ldots, D_K$ кластеризуются;
3. при фиксированных инвариантных преобразованиях $g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta}_k)$ оцениваются распределения параметров моделей кластеров $p(\mathbf{w}_c)$;
4. для каждой подвыборки D_k при фиксированных значениях параметров моделей кластеров $\mathbf{w}_c$, найденных в предыдущем пункте, оцениваются параметры $\boldsymbol{\theta}_k$ инвариантных преобразований.

Описание эксперимента

- Цель эксперимента: продемонстрировать работу предложенного алгоритма прогнозирования.
- Временной ряд — ЭКГ, длина 4170 отсчетов. В качестве предыстории взяты последние 60 отсчетов. В качестве тестовой выборки взяты последние 20 отсчетов.
- Задана модель

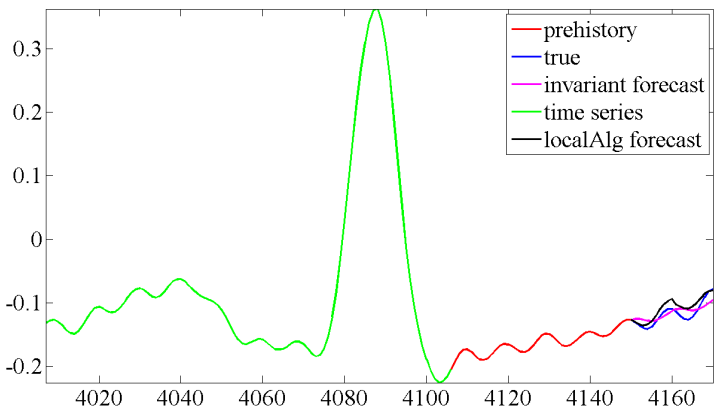
$$y = f(t, \mathbf{w}) = w_1 + w_2 t + w_3 \sin(w_4 t) + w_5 \sin(w_6 t),$$

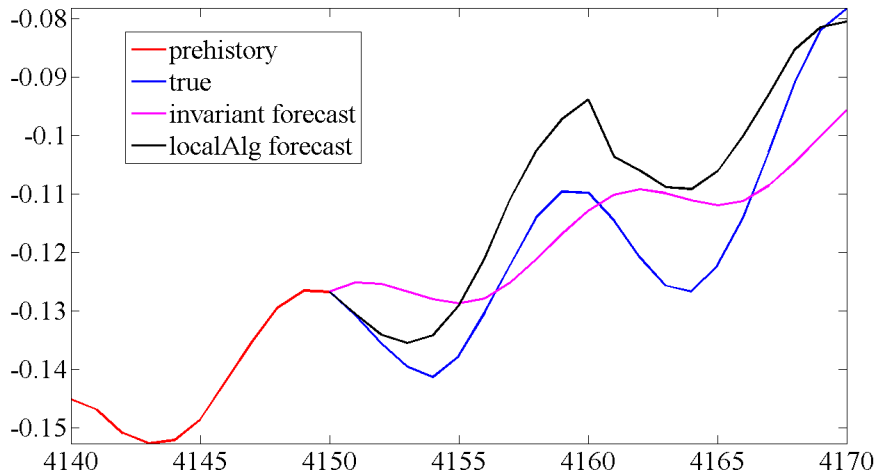
$w_i \in \mathbb{R}$.

- Задано семейство инвариантных преобразований SIM:

$$g(t, f(t, \mathbf{w}); \boldsymbol{\theta}) = \left(t, \theta_1 f\left(\frac{t - \theta_2}{\theta_3}, \mathbf{w} \right) + \theta_4 \right).$$

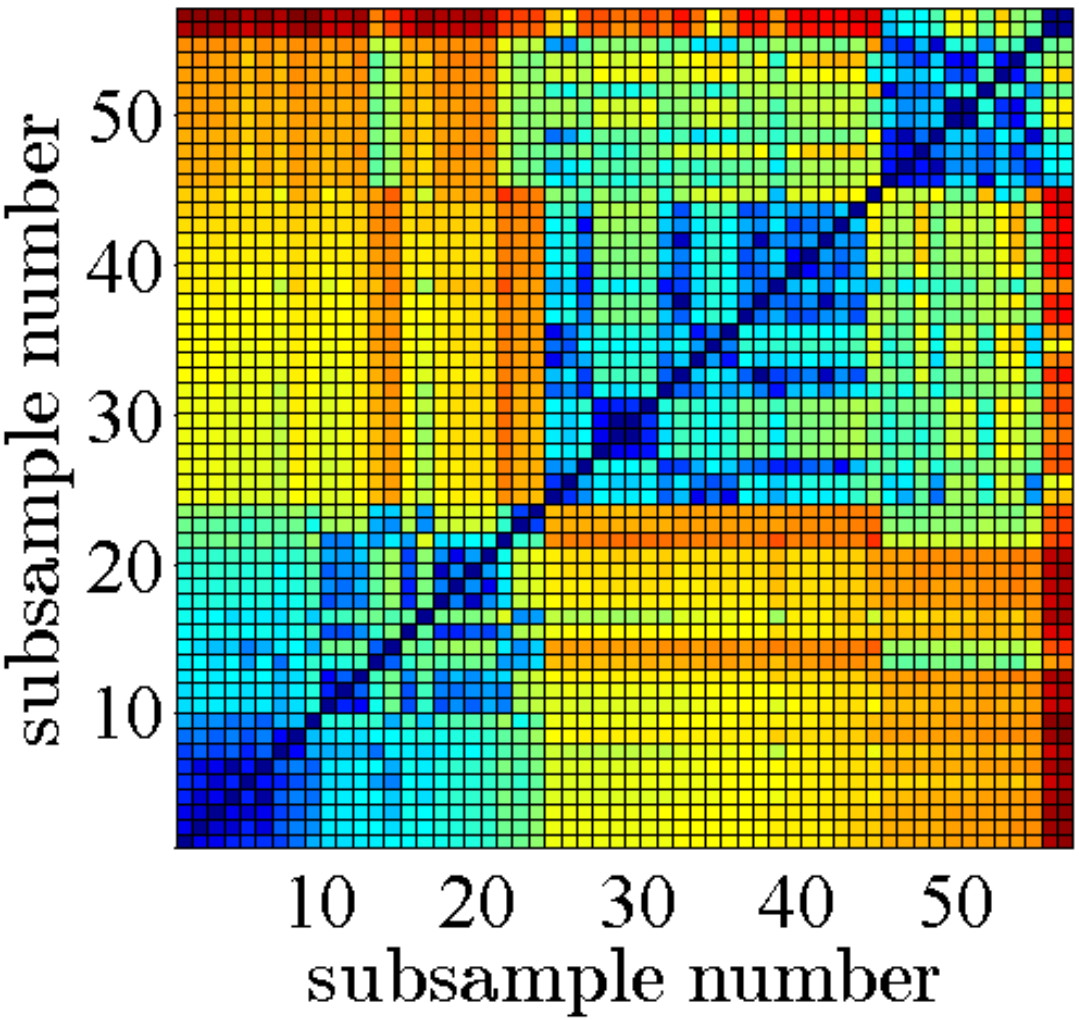
Результаты прогнозирования



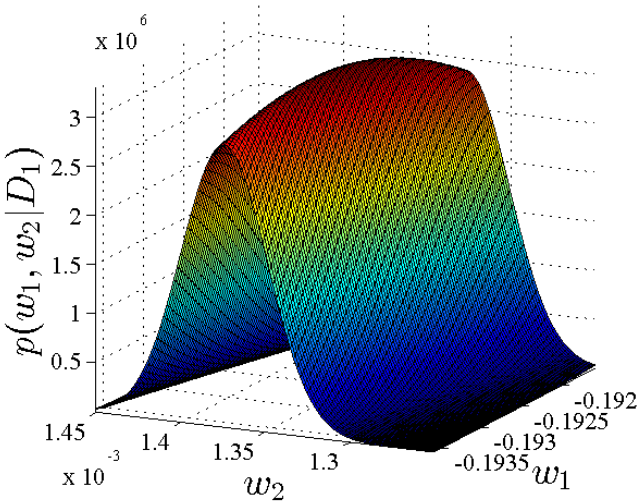


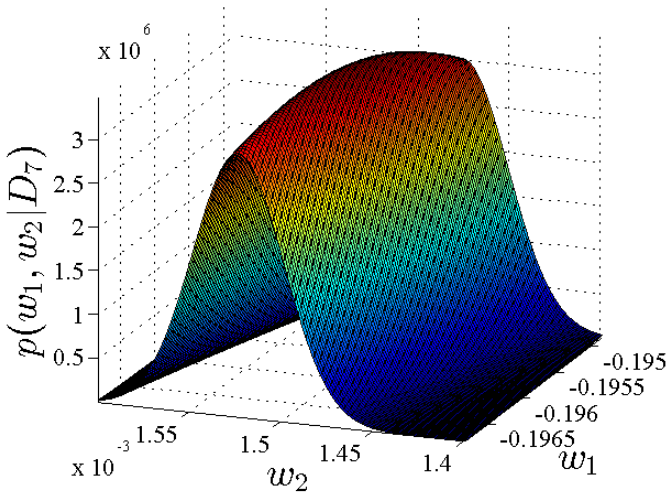
- Для каждого отсчета из тестовой выборки предыстория — все предыдущие отсчеты от начала prehistory.
- Прогноз делается на один отсчет вперед.
- $\mathrm{rMSE}_{\mathrm{invariant}} = 0.0108$, $\mathrm{rMSE}_{\mathrm{localAlg}} = 0.0122$.

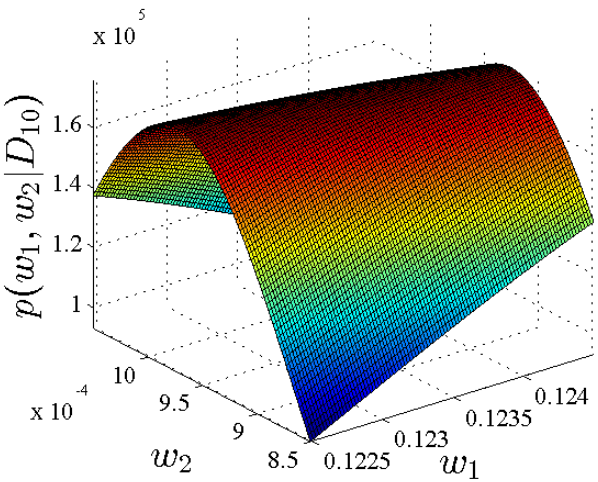
Попарное расстояние между подвыборками

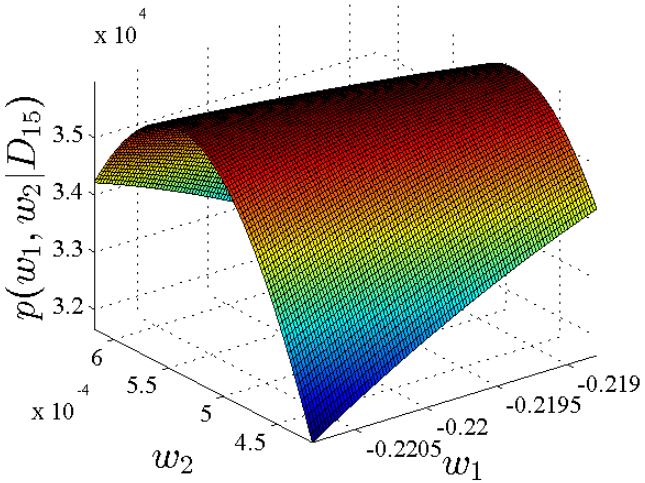


Распределения параметров









Выводы

- Предложен метод описания семейства регрессионных подвыборок, в котором:
  - кластеру подвыборок соответствует общая прогностическая модель;
  - каждой подвыборке соответствует инвариантное преобразование из экспертно заданного множества.
- Предложена функция расстояния между моделями, описывающими регрессионные подвыборки, основанная на распределении параметров моделей и инвариантная относительно определенного класса преобразований.
- Предложен алгоритм оценки параметров и кластеризации регрессионных подвыборок.
- Предложен метод локального прогнозирования временных рядов с введенным расстоянием между инвариантными участками временного ряда.