Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Кононенко Даниил Сергеевич

Оценка параметров инвариантных преобразований в задачах прогнозирования временных рядов

511656 - Математические и информационные технологии

Магистерская диссертация

Научный руководитель:
н.с. ВЦ РАН, к.ф.-м.н.
Стрижов Вадим Викторович

Москва

2013

Содержание

Аннотация

В данной работе рассматривается локальный метод прогнозирования временных рядов. Для определения функции близости используются параметрические модели участков временного ряда. В качестве меры близости берется расстояние между апостериорными распределениями параметров моделей участков временного ряда. Для восстановления распределения параметров используется двухуровневый байесовский вывод. Предлагается метод кластеризации регрессионных выборок, который в контексте локального прогнозирования применяется для нахождения участков временного ряда, похожих на предысторию. Для кластеризации применяется метод k-медоид. Вводится понятие параметрического инвариантного преобразования. Эти преобразования моделируют локальные изменения отдельно взятых выборок относительно общей модели кластера. Для совместной настройки параметров инвариантных преобразований и восстановления распределения параметров модели предлагается подход, основанный на методе, известном в литературе как Self-Modeling Regression. Работа алгоритма кластеризации и основанном на нем алгоритме прогнозирования продемонстрирована на реальных и синтетических данных. Проведено сравнение с локальным методом прогнозирования временных рядов, использующим расстояние в пространстве данных.

1 Введение

1.1 Локальное прогнозирование

Одним из подходов к прогнозированию временных рядов является метод локального прогнозирования, рассмотренный в работе [14]. Пусть во временном ряде выделена предыстория — некоторая последняя часть временного ряда. На основании предыстории необходимо сделать прогноз на заданный момент времени вперед. Идея локального метода прогнозирования в том, что в ряде находятся участки, в некотором смысле похожие на предысторию — см. рис. 1. Прогноз строится на основании этих участков и значений временного ряда в моменты времени, соответствующие тому, который необходимо спрогнозировать.

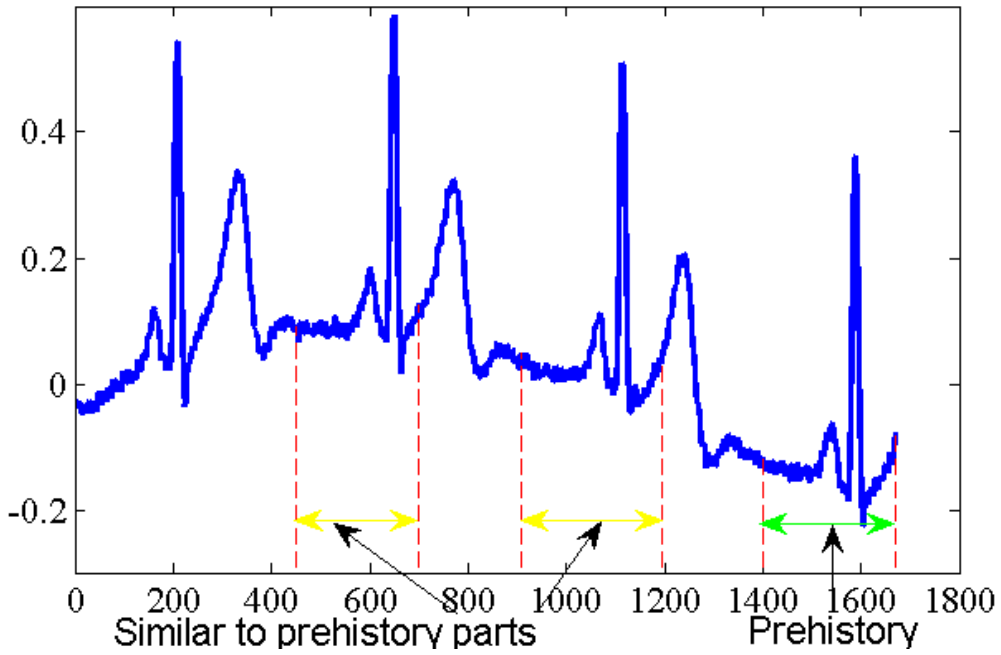

Рис. 1: Локальное прогнозирование временных рядов

В качестве функции близости участков временного ряда может браться различные классы функций [14], [2], [4], [5]. Если участки временного ряда имеют одинаковую длину, то такой функцией близости может быть, например, евклидово расстояние расстояние между участками — векторами. Функция близости может также принадлежать некоторому параметрическому семейству. Подбор параметров осуществляется по данным, минимизируя для каждых двух участков функции близости. Например, в работе [14] используется взвешенная евклидова метрика. В работах [4], [5]

рассматривается евклидова метрика с весами, убывающими как геометрическая прогрессия для более ранних по времени отсчетов. В работе [2] производится сравнение различных метрик: евклидовой, взвешенной евклидовой и метрики Минковского. Выбор функции близости определяется условиями конкретной прикладной задачи.

Данные реальных временных рядов имеют особенности, которые усложняют сравнение участков ряда между собой. Например, во временном ряде могут отсутствовать некоторые значения (пропуски в ряде) или среди известных значений могут попадаться сильно отличающиеся от остальных (выбросы). Один участок может быть растянут или сжат относительно другого во времени, иметь другую амплитуду — см. рис. 2. Поэтому в прикладных задачах редко случается полное совпадение участков временного ряда с предысторией. Вообще говоря, для хорошей работы метода локального прогнозирования необходимо выделять во временном ряде участки, похожие на предысторию в смысле выбранной функции близости, с точностью до некоторого преобразования, в общем случае нелинейного. В работах [4], [14], рассматривается проблема выбора параметров таких преобразований.

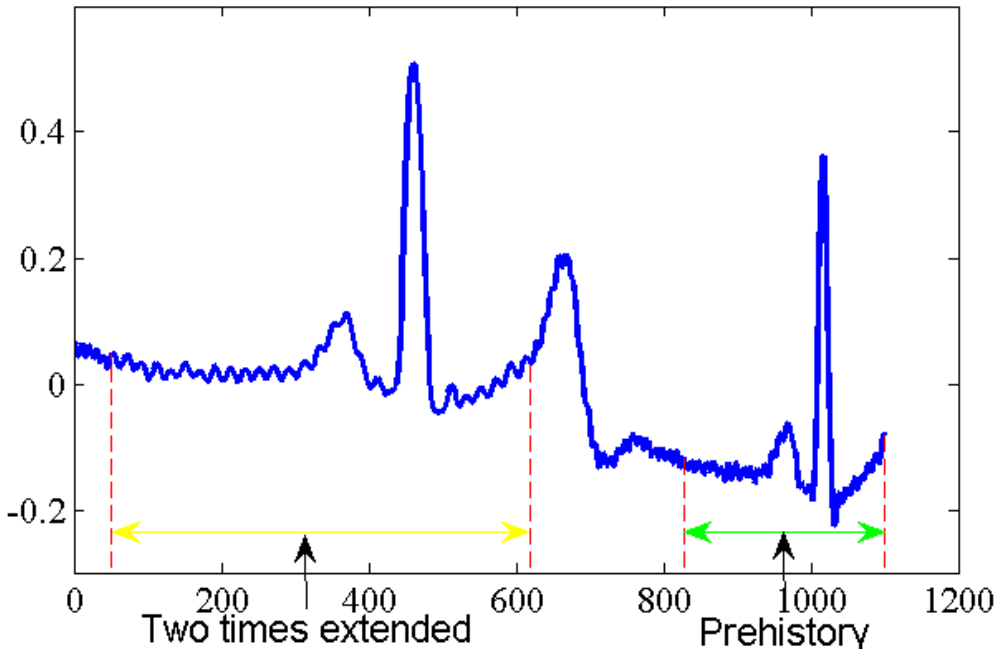

Рис. 2: Растяжение участка временного ряда

В данной работе предлагается использовать параметрические модели участков временного ряда для определения функции близости. Предполагается, что все близкие участки временного ряда порождены одной моделью. Таким образом, все выде-

ленные участки временного ряда могут быть кластеризованны: внутри кластера все участки порождены одной моделью. В качестве меры близости моделей предлагается брать близость распределения параметров модели. В качестве метрики на распределениях используется расстояние Йенсена-Шеннона [8]. Таким образом, участки временного ряда считаются близкими, если близки распределения их параметров.

Для кластеризации участков временного ряда предлагается использовать метод k-медоид [18]. Это вариант метода k-средних, когда центром каждого кластера обязательно является один из объектов.

Апостериорное распределение параметров модели оценивается по данным. Подробнее метод оценки описан в разделе 1.3.

В силу особенностей, описанных выше, участки временного ряда могут быть близки с точностью до некоторого преобразования. Поэтому рассматриваются параметрические преобразования участков временного ряда, в общем случае нелинейные. Эти преобразования называются инвариантными. Если после применения к участкам временных рядов инвариантных преобразований распределения их параметров близки, то данные участки считаются близкими.

1.2 SM-регрессия

Для совместной оценки параметров инвариантных преобразований и распределения параметров моделей участков временного ряда предлагается использовать подход, сходный с известным в литературе как Self-Modeling Regression (SEMOR).

Впервые термин Self-Modeling Regression встречается в работе [13]. В дальнейшем такие модели часто рассматриваются в литературе, например [6], [10]. Такой подход позволяет одновременно оценивать по данным и модель, и ее параметры. Модель оценивается с помощью непараметрических методов, в то время как ее параметры оцениваются с помощью параметрической нелинейной регрессии. Поэтому Self-Modeling Regression называют полупараметрическим методом. Обычно в литературе непараметрическую оценку модели и оценку параметров применяют итеративно, выбирая тем самым лучшую модель.

В этих работах рассматривается параметрическая модель семейства кривых

$$y_{ij} = f(t_{ij}, \boldsymbol{\theta}_i) + \varepsilon_{ij}; \tag{1.1}$$
$$i = 1 \dots m,$$
$$j = 1 \dots n_i.$$

Здесь m — число объектов измерения, для каждого из которых известны отсчеты y_{ij} в моменты времени t_{ij}, ε_{ij} — ошибка измерения.

Важным примером моделей, допускающих оценивание с помощью SEMOR, является shape invariant model (SIM). В этом случае индивидуальные модели объектов $f_i(t) = f(t, \boldsymbol{\theta}_i)$ получаются из общей функции формы φ с помощью параметрического преобразования. В самом простом случае линейного преобразования

$$f(t, \boldsymbol{\theta}_i) = f_i(t) = \theta_{i1}\varphi\left(\frac{t - \theta_{i2}}{\theta_{i3}}\right) + \theta_{i4}. \tag{1.2}$$

Это семейство всевозможных сдвигов и растяжений по обеим осям.

В работе [13] функция формы φ приближается сплайнами первого порядка. Вектор параметров $\boldsymbol{\theta}$ оценивается с помощью метода наименьших квадратов. Процедура повторяется итеративно, пока оценки изменяются больше, чем фиксированный порог.

В выражении (1.2) параметры $\boldsymbol{\theta}$ определены неоднозначно, т.е. для любого вектора параметров $\boldsymbol{\theta}$ существует бесконечно много отличных от него векторов параметров $\boldsymbol{\alpha} \neq \boldsymbol{\theta}$ таких, что получившиеся модели совпадают в любой момент времени:

$$\varphi_{\boldsymbol{\alpha}}(t) = \alpha_1\varphi\left(\frac{t - \alpha_2}{\alpha_3}\right) + \alpha_4, \tag{1.3}$$
$$\varphi_{\boldsymbol{\alpha}}(t) = \varphi_{\boldsymbol{\theta}}(t)\ \forall t \in \mathbb{R}, .$$

В работе [10] предлагается способ нормализации параметров $\boldsymbol{\theta}$, чтобы обеспечить единственность решения (выбор только одной функции из класса (1.3)). Функция формы φ оценивается с помощью ядерных методов (используется свертка с ядром). Процедура оценки функции формы и параметров повторяется итеративно. Предложенный алгоритм основывается на наличии начальных приближений для параметров $(\theta_{i2}, \theta_{i3})$. В работе [12] предлагается способ определения этих начальных приближений. В работах [10], [11] доказана асимптотическая сходимость оценок функции формы и параметров.

В работе [6] предлагается следующая смешанная модель:

$$\theta_{ik} = g(\mathbf{X}_i, \mathbf{Z}_i; \chi_k, \psi_k) + \eta_{ik}.$$

Здесь $\mathbf{X}$, $\mathbf{Z}$ — инвариантные по времени ковариаты, χ_k — неслучайные величины, связанные с $\mathbf{X}$, ψ_k — случайные величины, связанные с $\mathbf{Z}$. Рассматривается линейный случай:

$$\theta_{ik} = \mathbf{X}_i\varphi_k + \mathbf{Z}_i\psi_k + \eta_{ik}. \tag{1.4}$$

В данном случае φ — параметрическая функция, χ_k — неизвестные параметры, ψ_k, η_{ik}, ε_{ij} полагаются нормально распределенными. В модели (1.1), (1.2), (1.4) с указанными выше вероятностными предположениями для оценки функции φ используются регрессионные сплайны со штрафом, для оценок параметров — метод максимума правдоподобия.

В работе [9] данный подход обобщен. Рассматривается полупараметрическая нелинейная модель со смешанными случайными величинами:

$$\begin{aligned} &y_{ij} = f(\boldsymbol{\theta}_i, \varphi; t_{ij}) + \varepsilon_{ij}, \quad i = 1 \ldots m, \quad j = 1 \ldots n_i, \\ &\boldsymbol{\theta}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i, \quad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{D}). \end{aligned} \tag{1.5}$$

Здесь f — известная функция, φ — неизвестная функция, β — неизвестный вектор неслучайных величин, $\mathbf{b}_i$ — неизвестный вектор случайных величин, связанный с наблюдением i, $\mathbf{A}_i$, $\mathbf{B}_i$ — известные ковариационные матрицы, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Lambda}_i)$, амплитуда шума σ^2 известная, ковариационные матрицы $\mathbf{D}$, $\boldsymbol{\Lambda}_i$ известны. Для оценки параметров предлагается итерационная процедура. При фиксированных параметрах $\boldsymbol{\theta}$, функция f и вектора β, $\mathbf{b}_i$ оцениваются с помощью метода максимума правдоподобия (максимизируется взвешенная функция правдоподобия Хендерсона).

1.3 Анализ структурных параметров

В статье [3] предлагается метод оценки апостериорного распределения параметров в регрессионных моделях, который используется в данной работе. Рассматривается регрессионная выборка

$$D = \{\mathbf{x}_i, y_i\}_{i=1}^m = (\mathbf{X}, y)\,,$$

где $\mathbf{x}_i \in \mathbb{R}^n$, а $y_i \in \mathbb{R}$. Решается задача восстановления регрессии

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

где $\mathbf{f}(\mathbf{w}, \mathbf{X})$ — некоторая параметрическая вектор-функция. Предполагается, что многомерная случайная величина $\mathbf{y}$ имеет нормальное распределение:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_m),$$

где σ^2 — дисперсия распределения, $\mathbf{I}_m$ — единичная матрица размерности m. Требуется приблизить функцию $\mathbf{f}(\mathbf{w}, \mathbf{X})$ параметрической функцией $\hat{\mathbf{f}}(\mathbf{X}, \mathbf{w})$ из заданного класса. Отображение

$$\mathbf{f} : \mathbb{R}^{m \times n} \times \mathbb{W}^W \to \mathbb{R}^m$$

называется моделью.

Распределение зависимой переменной $\mathbf{y}$ можно представить в виде:

$$p(\mathbf{y}) = (2\pi\beta^{-1})^{-\frac{m}{2}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{f})^T \beta \mathbf{I} (\mathbf{y} - \mathbf{f})\right).$$

Функция правдоподобия данных имеет вид

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \beta, \mathbf{f}) = p(D|\mathbf{w}, \beta, \mathbf{f}) = \frac{\exp(-\beta E_D)}{Z_D(\beta)},$$

где функция ошибки

$$E_D = \frac{1}{2}(\mathbf{y} - \mathbf{f})^T (\mathbf{y} - \mathbf{f}),$$

нормирующий коэффициент

$$Z_D(\beta) = (2\pi\beta^{-1})^{\frac{m}{2}}.$$

Предполагается также, что $\mathbf{w} \in \mathbb{R}^W$ также является многомерной случайной величиной с нормальным распределением:

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

где $E_\mathbf{w} = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T \mathbf{A} (\mathbf{w} - \mathbf{w}_0)$, $Z_\mathbf{w}(\mathbf{A}) = (2\pi)^{\frac{n}{2}} |\mathbf{A}^{-1}|^{\frac{1}{2}}$.

Апостериорное распределение параметров находится по формуле Байеса

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

Обозначая функцию ошибки как

$$S(\mathbf{w}) = E_\mathbf{w} + \beta E_D = \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^T \mathbf{A} (\mathbf{w} - \mathbf{w}_0) + \beta \frac{1}{2}(\mathbf{y} - \mathbf{f})^T (\mathbf{y} - \mathbf{f}),$$

получаем

$$p(\mathbf{w}|D, \mathbf{A}, \beta, f) = \frac{\exp(-S(\mathbf{w}))}{Z_S(\mathbf{A}, \beta)}, \tag{1.6}$$

где $Z_S(\mathbf{A},\beta) = Z_D(\beta)Z_{\mathbf{w}}(\mathbf{A})\int p(D|\mathbf{w},\mathbf{B},f)p(\mathbf{w}|\mathbf{A},f)d\mathbf{w}$ — нормирующий коэффициент.

Оценка нормировочного коэффициента производится с помощью аппроксимации Лапласа. Требуется найти нормировочную константу Z ненормированного распределения $p(\mathbf{w})$:

$$Z = \int p(\mathbf{w})d\mathbf{w}.$$

Пусть $p(\mathbf{w})$ имеет максимум в точке $\mathbf{w}_0$, т.е.

$$\left.\frac{dp(\mathbf{w})}{d\mathbf{w}}\right|_{\mathbf{w}=\mathbf{w}_0} = 0.$$

Тогда $p(\mathbf{w})$ можно разложить по Тейлору в окрестности точки $\mathbf{w}_0$:

$$\ln p(\mathbf{w}) = p(\mathbf{w}_0) - \frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T\mathbf{C}(\mathbf{w}-\mathbf{w}_0) + \ldots,$$

где $\mathbf{C}$ — матрица Гессе:

$$c_{ij} = -\left.\frac{\partial \ln p(\mathbf{w})}{\partial w_i \partial w_j}\right|_{\mathbf{w}=\mathbf{w}_0}.$$

Отбрасывая все члены выше квадратичного, получаем, что ненормированное распределение приближено нормированным нормальным:

$$\hat{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}_0, \mathbf{C}^{-1}).$$

Применяя этот метод для аппроксимации апостериорного распределения (1.6), получаем

$$p(\mathbf{w}|D,\mathbf{A},\beta,f) \approx \mathcal{N}(\mathbf{w}_0,\mathbf{H}^{-1}) = \frac{|H|^{\frac{1}{2}}\exp(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^T\mathbf{H}(\mathbf{w}-\mathbf{w}_0)}{(2\pi)^{\frac{n}{2}}},$$

где $\mathbf{H}$ — гессиан функции ошибки:

$$h_{ij} = -\left.\frac{\partial S(\mathbf{w})}{\partial w_i \partial w_j}\right|_{\mathbf{w}=\mathbf{w}_0}.$$

Для правдоподобия получаем приближение

$$\ln p(D|\mathbf{A},\beta,f) = -\frac{1}{2}\ln|\mathbf{A}^{-1}| - \frac{m}{2}\ln(2\pi) + \frac{m}{2}\ln\beta^{-1} - S(\mathbf{w}_0) - \frac{1}{2}|\mathbf{H}|. \tag{1.7}$$

Гиперпараметры A и β находятся максимизацией правдоподобия, для чего поочередно приравниваются к нулю производные выражения (1.7).

2 Постановка задачи

2.1 Выделение участков временного ряда

Будем рассматривать одномерные временные ряды — ряды, в которых каждому моменту времени сопоставляется вещественное число.

$$\left[t_1, t_2, \ldots, t_N\right]^T \to \left[y_1, y_2, \ldots, y_N\right]^T$$

Требуется предсказать следующее значение последовательности y_{N+1}, которое будет определяться значениями предыстории $\left[y_{N-L+1}, y_{N-L+2}, \ldots, y_N\right]^T$ длины L.

Для прогнозирования временного ряда в нем выделяются участки различной длины — кандидаты на то, чтобы быть похожими на предысторию:

$$\begin{gathered}\left[y_{i_1^1}, \ldots, y_{i_{n_1}^1}\right]^T, \\ \left[y_{i_1^2}, \ldots, y_{i_{n_2}^2}\right]^T, \\ \ldots, \\ \left[y_{i_1^{K-1}}, \ldots, y_{i_{n_{K-1}}^{K-1}}\right]^T.\end{gathered}$$

Вместе с предысторией получаем, что во временном ряде выделено K участков. Здесь $n_k \in \{1, \ldots, N-1\}$, $k = 1, \ldots, K$.

2.2 Модель семейства регрессионных подвыборок

Далее будем рассматривать выделенные участки как регрессионные подвыборки $(\mathbf{X}_1, \mathbf{y}_1), \ldots, (\mathbf{X}_K, \mathbf{y}_K)$, где $\mathbf{y}_k = (y_{i_1^k}, \ldots, y_{i_{n_k}^k})$. В рассматриваемой задаче прогнозирования временного ряда матрицы $\mathbf{X}_k$ — это вектора моментов времени: $\mathbf{X}_k = \left[t_{i_1^k}, \ldots, t_{i_{n_k}^k}\right]^T$. Однако все дальнейшие рассуждения будут верны и в более общем случае, когда задан набор подвыборок

$$\begin{gathered}D_1, \ldots, D_K : \\ D_k = (\mathbf{X}_k, \mathbf{y}_k) = \{(\mathbf{x}_{kj}, y_{kj})\}_{j=1}^{n_k},\end{gathered}$$

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

Смысл данных такой же, как в разделе 1.2 в выражении (1.1): имеется K объектов измерения, для каждого из которых есть подвыборка D_k из n_k объектов. В задачах SEMOR для всех объектов строилась одна регрессионная кривая, но для

каждого объекта она подвергалась некоторому преобразованию. В данной работе мы рассмотрим случай нескольких регрессионных кривых, каждая из которых описывает часть объектов.

Формально, пусть задан набор меток кластеров $\{1, \ldots, C\}$ и каждой метке поставлена в соответствие некоторая модель μ_c. Рассматривается параметрическое семейство прогностических моделей:

$$z = f_{\mathbf{w}}(\mathbf{x}) = f(\mathbf{x}, \mathbf{w}), \tag{2.1}$$

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

Каждой модели μ_c соответствует вектор параметров $\mathbf{w}_c$, т.е. $\mu_c = f(\mathbf{X}, \mathbf{w}_c)$. Каждая подвыборка принадлежит некоторому кластеру $c \in \{1, \ldots, C\}$ и соответствует некоторой модели $\mu_c(X, \mathbf{w})$. Метки кластеров выборок обозначим $c_1, \ldots, c_K$.

Как мы отмечали ранее в разделе 1.1, каждому объекту может соответствовать некоторое нелинейное преобразование.

Определение 2.1. *Инвариантным преобразованием назовем функционал, действующий в пространстве моделей:*

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

Т.к. модель $f(\cdot)$ действует из $\mathbb{R}^n$ в $\mathbb{R}$, то инвариантное преобразование действует из множества пар $(\mathbb{R}^n, \mathbb{R})$ в такое же множество пар. Инвариантное преобразование ставит в соответствие модели $f(\mathbf{x})$ некоторую новую модель $g(\mathbf{x}, f(\mathbf{x}))$. Например, преобразование $g(\mathbf{x}, f(\mathbf{x})) = (\mathbf{x}, f(\mathbf{x} - \mathbf{s}))$ сдвигает входное пространство на вектор $\mathbf{s}$.

Определение 2.2. *Введем параметрическое семейство инвариантных преобразований*

$$\mathcal{G} = \{g_{\boldsymbol{\theta}}(\mathbf{x}, f(\mathbf{x})) = g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) | \boldsymbol{\theta} \in \mathbb{R}^p\}. \tag{2.2}$$

Частным случаем такого семейства являются Shape Invariant Models (1.2), рассмотренные в разделе 1.2:

$$g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}) = \left(\mathbf{x}, \theta_1 f\left(\frac{\mathbf{x} - \boldsymbol{\theta}_2}{\theta_3}\right) + \theta_4\right). \tag{2.3}$$

Пусть каждой подвыборке D_k соответствует некоторый набор параметров $\boldsymbol{\theta}_k$. Таким образом, если подвыборка D_k принадлежит кластеру c и ей соответствует

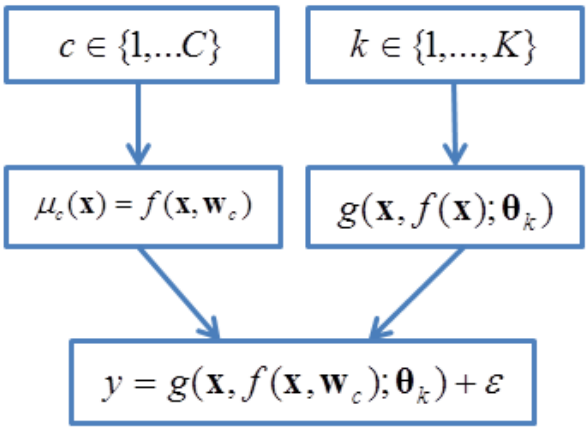

Рис. 3: Схема порождения данных

набор параметров $\boldsymbol{\theta}_k$ инвариантного преобразования, то она порождена моделью $\mu_c(\mathbf{x}, \mathbf{w}) = f(\mathbf{x}, \mathbf{w}_c)$, измененной инвариантным преобразованием $g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}_k)$:

$$\mathbf{y}_k = \mathbf{g}(\mathbf{X}_k, \boldsymbol{\mu}_c(\mathbf{X}_k, \mathbf{w}); \boldsymbol{\theta}_k),$$
$$\mathbf{X}_k \in \mathbb{R}^{n_k \times n}, \mathbf{y}_k \in \mathbb{R}^{n_k}.$$

Общая модель порождения данных имеет вид

$$y_j = g(\mathbf{x}_j, f(\mathbf{x}_j, \mathbf{w}_c); \boldsymbol{\theta}_k) + \varepsilon; \tag{2.4}$$
$$c \in \{1, \ldots, C\},\ k \in \{1, \ldots, K\}, j \in \{1, \ldots n_k\};$$
$$\mathbf{x}_j \in \mathbb{R}^n,\ \mathbf{w}_c \in \mathbb{R}^n,\ \boldsymbol{\theta}_k \in \mathbb{R}^p.$$

Здесь c — номер кластера, k — номер подвыборки, j — номер объекта в подвыборке.

Тогда для подвыборки $(\mathbf{X}, \mathbf{y})$:

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

где шум $\boldsymbol{\varepsilon}$ — многомерная нормальная случайная величина с нулевым математическим ожиданием и матрицей ковариации $\mathbf{B}^{-1}$:

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

Схема порождения данных изображена на рисунке 3.

2.3 Прогнозирование временного ряда

Задача оценки параметров и кластеризации регрессионных выборок состоит в том, чтобы по заданным выборкам $D_1, \ldots, D_K$, параметрическим семействам $f(\mathbf{x}, \mathbf{w})$,

$g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta})$ и числу кластеров C восстановить метки кластеров подвыборок $c_1, \ldots, c_K$ и оценить параметры инвариантных преобразований $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$ и параметры моделей кластеров $\mathbf{w}_1, \ldots, \mathbf{w}_C$.

После оценки параметров получаем модель предыстории в виде

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

Здесь c — номер кластера предыстории, k — номер выборки, которая соответствует предыстории. Эта модель может быть использована для прогнозирования временного ряда по предыстории. Напомним, что в задаче прогнозирования выборки — это участки временного ряда, и независимая переменная $\mathbf{x}$ — это время t. Тогда прогноз в будущий момент времени t

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

3 Алгоритм прогнозирования временного ряда

3.1 Оценка параметров инвариантных преобразований

Сначала покажем, как настраивать параметры инвариантного преобразования (2.2). Рассмотрим подвыборку D_k, $k \in 1, \ldots, K$. Пусть в общей модели (2.4) фиксирован номер кластера этой подвыборки $c \in \{1, \ldots C\}$ и параметры $\mathbf{w}_c$ модели $f(\mathbf{x}, \mathbf{w}_c)$. Необходимо оценить параметры $\boldsymbol{\theta}_k$ преобразования из семейства $g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta})$. Введем квадратичную функцию ошибки на подвыборке D_k

$$S_{LS}(\boldsymbol{\theta}_k) = \sum_{i=1}^{n_k} (y_{ki} - \hat{y}_{ki}(\boldsymbol{\theta}_k))^2, \tag{3.1}$$

где $\hat{y}_{ki}(\boldsymbol{\theta}_k) = g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}_c); \boldsymbol{\theta}_k)$. Для нахождения оптимальных параметров предлагается минимизировать эту квадратичную ошибку:

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

Оптимизацию предлагается проводить с помощью алгоритма Левенберга-Марквардта [15]. В данной работе использовалась реализация Matlab, функция nlinfit.

3.2 Оценка апостериорного распределения параметров модели

Теперь опишем алгоритм оценки распределения параметров модели (2.1). Рассмотрим подвыборку D_k, $k \in 1, \ldots, K$. Пусть в общей модели (2.4) фиксирован номер кластера этой подвыборки $c \in \{1, \ldots C\}$ и некоторое инвариантное преобразование $g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta}_k)$. Необходимо оценить апостериорное распределение параметров модели $f(\mathbf{x}, \mathbf{w})$, и найти оценку параметров $\mathbf{w}_k$ как максимум этого распределения.

Обозначим $h(\mathbf{x}, \mathbf{w}) = g(f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta}_k)$. Модель порождения выборки D_k

$$\mathbf{y}_k = h(\mathbf{X}_k, \mathbf{w}) + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{B}^{-1})$. Метод оценки апостериорного распределения $p(\mathbf{w}_k | D_k, \hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k, \boldsymbol{\theta}_k, f)$ в такой модели с помощью оптимизации правдоподобия описан в разделе 1.3 — используется аппроксимация Лапласа.

Пусть рассматривается семейство SIM инвариантных преобразований (2.3). Фиксируем конкретное инвариантное преобразование с параметрами $\boldsymbol{\theta}_k$. Тогда модель порождения выборки D_k имеет вид

$$y = \theta_{k1} f\left(\frac{\mathbf{x} - \boldsymbol{\theta}_{k2}}{\theta_{k3}}, \mathbf{w}_c\right) + \theta_{k4}.$$

После преобразований получаем

$$\frac{y - \theta_{k4}}{\theta_{k1}} = f\left(\frac{\mathbf{x} - \boldsymbol{\theta}_{k2}}{\theta_{k3}}, \mathbf{w}_c\right).$$

Обозначим выборку $D'_k = (\mathbf{X}'_k, \mathbf{y}'_k)$:

$$\mathbf{x}'_{ki} = \frac{\mathbf{x}_{ki} - \boldsymbol{\theta}_{k2}}{\theta_{k3}},$$
$$y'_{ki} = \frac{y_{ki} - \theta_{k4}}{\theta_{k1}}.$$

Получаем, что выборка D'_k порождена моделью

$$\mathbf{y}'_k = f(\mathbf{X}'_k, \mathbf{w}_c) + \boldsymbol{\varepsilon},$$

где шум $\boldsymbol{\varepsilon}$ — многомерная нормальная случайная величина с нулевым математическим ожиданием и матрицей ковариации $\mathbf{B}^{-1}$:

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

Метод оценки апостериорного распределения в такой модели описан в разделе 1.3.

Пусть теперь необходимо оценить распределение параметров $\mathbf{w}_c$ кластера. Предполагаем, что фиксированы выборки $D_{i_{c1}}, D_{i_{c2}}, \ldots, D_{i_{cl_c}}$ и инвариантные преобразования с параметрами $\boldsymbol{\theta}_{i_{c1}}, \boldsymbol{\theta}_{i_{c2}}, \ldots, \boldsymbol{\theta}_{i_{cl_c}}$. Обозначим

$$\begin{aligned} h_{i_{c1}}(\mathbf{x}, \mathbf{w}) &= g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta}_{i_{c1}}); \\ h_{i_{c2}}(\mathbf{x}, \mathbf{w}) &= g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta}_{i_{c2}}); \\ &\ldots \\ h_{i_{cl_c}}(\mathbf{x}, \mathbf{w}) &= g(\mathbf{x}, f(\mathbf{x}, \mathbf{w}); \boldsymbol{\theta}_{i_{cl_c}}); \end{aligned}$$

Объединяем подвыборки, лежащие в кластере c, в одну выборку $D_c = (\mathbf{X}_c, \mathbf{y}_c)$. Каждая подвыборка входит в объединение со своим инвариантным преобразованием. Обозначаем общую модель

$$\mathbf{h}_c(\mathbf{X}, \mathbf{w}) = [\mathbf{h}_{i_{c1}}(\mathbf{X}, \mathbf{w}), \mathbf{h}_{i_{c2}}(\mathbf{X}, \mathbf{w}), \ldots, \mathbf{h}_{i_{cl_c}}(\mathbf{X}, \mathbf{w})]^T.$$

Модель порождения для выборки D_c имеет вид

$$\mathbf{y}_c = \mathbf{h}(\mathbf{X}_c, \mathbf{w}) + \boldsymbol{\varepsilon},$$

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

Метод оценки апостериорного распределения $p(\mathbf{w}_c | D_{i_{c1}}, \ldots, D_{i_{cl_c}}, \hat{\mathbf{A}}_c, \hat{\mathbf{B}}_c, \boldsymbol{\theta}_{i_{c1}}, \ldots, \boldsymbol{\theta}_{i_{cl_c}}, f)$ в такой модели с помощью оптимизации правдоподобия описан в разделе 1.3.

3.3 Кластеризация регрессионных подвыборок

Наконец, рассмотрим алгоритм кластеризации подвыборок. Пусть в общей модели (2.4) для каждой подвыборки D_k фиксировано распределение параметров $p(\mathbf{w}_k)$ модели $f(\mathbf{x}, \mathbf{w}_k)$ и некоторое инвариантное преобразование $g(\mathbf{x}, f(x); \boldsymbol{\theta}_k)$. Пусть также фиксировано число кластеров $C \leq K$, на которые необходимо разделить подвыборки $D_1, \ldots, D_K$.

В качестве расстояния между распределениями предлагается брать метрику Йенсена-Шеннона [8]:

$$JSD^2(p_1(\mathbf{x}), p_2(\mathbf{x})) = \frac{1}{2} D_{KL}\left(p_1(\mathbf{x}), p(\mathbf{x})\right) + \frac{1}{2} D_{KL}\left(p_2(\mathbf{x}), p(\mathbf{x})\right).$$

Здесь $p(\mathbf{x}) = \frac{1}{2}p_1(\mathbf{x}) + \frac{1}{2}p_2(\mathbf{x})$, $D_{KL}\left(p(\mathbf{x}), q(\mathbf{x})\right)$ — расстояние Кульбака-Лейблера [7], [8]:

$$D_{KL}\left(p(\mathbf{x}), q(\mathbf{x})\right) = \int p(\mathbf{x}) \ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x}.$$

Расстояния Йенсена-Шеннона между двумя нормальными распределениями не выражается аналитически. Поэтому использовался приближенный метод [16]. Использовалась реализация из MVN Toolbox [17].

Таким образом, получаем задачу кластеризации K подвыборок, между которыми задано попарное расстояние $\rho(D_i, D_j) = JSD(p(\mathbf{w}_i), p(\mathbf{w}_j)$. Для кластеризации предлагается использовать метод k-медоид [18]. Этот метод является вариантом метода k-средних и отличается тем, что центром каждого кластера обязательно является один из объектов.

В данном методе минимизируется суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$J = \sum_{c=1}^{C} \sum_{k \in R_c} \rho(D_k, D_{l_c}). \tag{3.2}$$

Здесь D_{l_c} — центры кластеров, l_c — номер объекта, являющегося центром кластера c, R_c — множество объектов, лежащих в кластере c.

Если объектов и кластеров немного, то оптимизировать функционал (3.2) можно полным перебором. В случае, если полный перебор занимает слишком много времени, можно воспользоваться приближенным алгоритмом: итеративно перевычисляются центры кластеров, полученных на предыдущем шаге, затем объекты разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе — см. алгоритм 3.1.

3.4 Алгоритм оценки параметров и кластеризации подвыборок

Объединим теперь описанные выше методы оценки параметров инвариантных преобразований, оценки апостериорного распределения параметров моделей и кластеризации регрессионных подвыборок в общую итеративную процедуру.

В данной работе предлагается два алгоритма. Алгоритм 3.2 основан на методе оценки параметров, используемом в статьях по SEMOR [13], [6], [10]. Параметры моделей и параметры инвариантных преобразований для каждой подвыборки оцениваются по очереди: оптимизируются одни при фиксированных других. Процесс

Алгоритм 3.1. Метод k-медоид

Вход: C, $\rho(D_i, D_j)$, $i, j = 1, \ldots, K$;

Выход: l_c, R_c, $c = 1, \ldots, C$;

1: Инициализируем $l_1, \ldots, l_C$ случайно;
2: **пока** $l_1, \ldots, l_C$ изменяются
3: **для** c = 1, . . . , C
4: $R_c := \emptyset$;
5: **для** k = 1, . . . , K
6: $c_k = \arg \min_{c=1,\ldots,C} \rho(D_k, D_c)$.
7: $R_{c_k} := R_{c_k} \cup k$;
8: **для** c = 1,. . . , C
9: $l_c := \arg \min_{k_c \in R_c} \sum_{k \in R_c} \rho(D_{k_c}, D_k)$;

повторяется итеративно до выполнения некоторого критерия: например, фиксированное число итераций или до стабилизации параметров. После этого подвыборки кластеризуются и оцениваются параметры моделей в каждом кластере при фиксированных инвариантных преобразованиях подвыборок.

Алгоритм 3.3 учитывает отличия предложенной в разделе 2.2 модели (2.4) от SEMOR, а именно наличие нескольких кластеров подвыборок с общей моделью внутри кластера. В итерации (строки 2-9 алгоритма 3.3) кроме оценки параметров моделей и параметров инвариантных преобразований подвыборок включается кластеризация и оценка параметров моделей кластеров.

4 Вычислительные эксперименты

4.1 Эксперименты на модельных данных по кластеризации подвыборок

Предложенные алгоритмы кластеризации 3.2, 3.3 были протестированы на модельных данных. Рассматривались линейные и квадратичные модели

Алгоритм 3.2. Итеративная оценка параметров и финальная кластеризация подвыборок

Вход: Подвыборки $D_k = (\mathbf{X}_k, \mathbf{y}_k),\ k = 1, \ldots, K$, семейство $f(\mathbf{x}, \mathbf{w})$, семейство $g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta})$, количество кластеров C;

Выход: Кластеры R_c, параметры инвариантных преобразований $\boldsymbol{\theta}_k$, распределения параметров моделей кластеров $p(\mathbf{w}_c),\ k = 1, \ldots, K,\ c = 1, \ldots, C$;

1: **для** k = 1,. . . , K

2: Инициализация параметров $\boldsymbol{\theta}_k$ и $p(\mathbf{w}_k)$;

3: **пока** не выполнен критерий остановки

4: При фиксированном $\boldsymbol{\theta}_k$ оценить распределения $p(\mathbf{w}_k | D_k, \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\theta}_k, f)$ с помощью максимизации правдоподобия:

$$[\hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k] = \arg \max_{\mathbf{A}_k, \mathbf{B}_k} p(D_k | \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\theta}_k, f).$$

Оценка параметров есть максимум этого распределения:

$$\hat{\mathbf{w}}_k = \arg \max_{\mathbf{w}_k} p(\mathbf{w}_k | D_k, \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\theta}_k, f).$$

5: При фиксированных значениях $\hat{\mathbf{w}}_k$ оценить параметры инвариантного преобразования $\boldsymbol{\theta}_k^* = \arg \min_{\boldsymbol{\theta}_k \in \mathbb{R}^p} S_{LS}(\boldsymbol{\theta}_k)$ (см. выражение (3.1));

6: Кластеризовать подвыборки $D_1, \ldots, D_k$ с помощью полного перебора или алгоритма k-медоид 3.1.

7: **для** c = 1,. . . , C

8: Оценить апостериорные распределения параметров моделей кластеров $p(\mathbf{w}_c | D_{i_{c1}}, \ldots, D_{i_{cl_c}}, \hat{\mathbf{A}}_c, \hat{\mathbf{B}}_c, \boldsymbol{\theta}_{i_{c1}}, \ldots, \boldsymbol{\theta}_{i_{cl_c}}, f)$ с помощью метода максимума правдоподобия.

$$y = wx,$$
$$y = (wx)^2,$$
$$x, y \in \mathbb{R}.$$

В качестве инвариантного преобразования брался сдвиг по оси x:

$$g(f(x), \theta) = f(x - \theta).$$

Алгоритм 3.3. Итеративная кластеризация подвыборок и оценка параметров

Вход: Подвыборки $D_k = (\mathbf{X}_k, \mathbf{y}_k)$, $k = 1, \ldots, K$, семейство $f(\mathbf{x}, \mathbf{w})$, семейство $g(\mathbf{x}, f(\mathbf{x}); \boldsymbol{\theta})$, количество кластеров C;

Выход: Кластеры R_c, параметры инвариантных преобразований $\boldsymbol{\theta}_k$, распределения параметров моделей кластеров $p(\mathbf{w}_c)$, $k = 1, \ldots, K$, $c = 1, \ldots, C$;

1: Инициализация параметров $\boldsymbol{\theta}_k$;

2: **пока** не выполнен критерий остановки

3: **для** k = 1,. . . , K

4: При фиксированном $\boldsymbol{\theta}_k$ оценить распределения $p(\mathbf{w}_k | D_k, \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\theta}_k, f)$ с помощью максимизации правдоподобия:

$$[\hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k] = \arg\max_{\mathbf{A}_k, \mathbf{B}_k} p(D_k | \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\theta}_k, f).$$

Оценка параметров есть максимум этого распределения:

$$\hat{\mathbf{w}}_k = \arg\max_{\mathbf{w}_k} p(\mathbf{w}_k | D_k, \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\theta}_k, f).$$

5: Кластеризовать подвыборки $D_1, \ldots, D_k$ с помощью полного перебора или алгоритма k-медоид 3.1;

6: **для** c = 1,. . . , C

7: Оценить апостериорные распределения параметров моделей кластеров $p(\mathbf{w}_c | D_{i_{c1}}, \ldots, D_{i_{cl_c}}, \hat{\mathbf{A}}_c, \hat{\mathbf{B}}_c, \boldsymbol{\theta}_{i_{c1}}, \ldots, \boldsymbol{\theta}_{i_{cl_c}}, f)$ с помощью метода максимума правдоподобия;

8: **для** k = 1,. . . , K

9: При фиксированных значениях $\hat{\mathbf{w}}_c$ (c —номер кластера, которому принадлежит выборка D_k) оценить параметры инвариантного преобразования $\boldsymbol{\theta}_k^* = \arg\min_{\boldsymbol{\theta}_k \in \mathbb{R}^p} S_{LS}(\boldsymbol{\theta}_k)$ (см. выражение (3.1)).

Генерировались подвыборки двух кластеров со значениями параметров $w_1 = 1$ и $w_2 = 5$. В каждом кластере генерировалось пять подвыборок с разными значениями сдвига θ. К данным добавлялся случайный нормальный шум с дисперсией $\sigma^2 = 0.1$. Сгенерированные подвыборки: линейная модель — рис. 4, квадратичная модель — рис. 5.

В таблице 1 представлены результаты эксперимента. В ней указаны значения параметров, с которыми генерировалась подвыборка, и значения, восстановленные

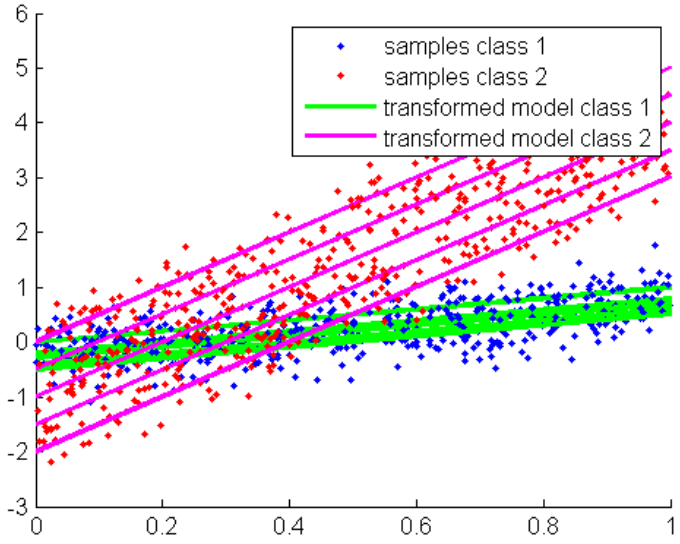

Рис. 4: Синтетические данные, линейная модель

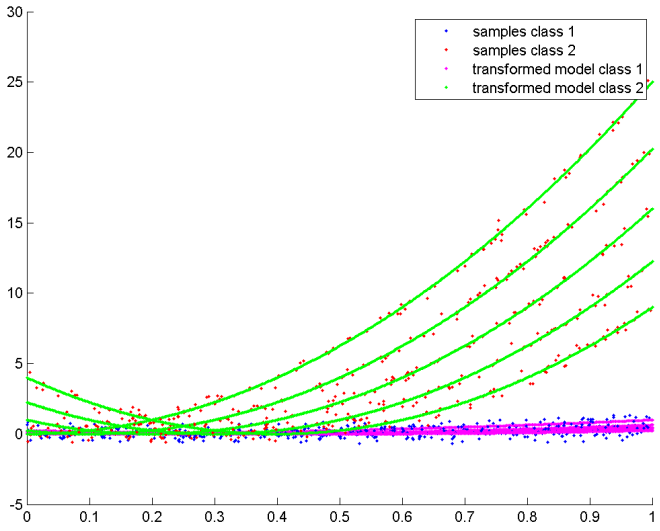

Рис. 5: Синтетические данные, квадратичная модель

алгоритмами 3.2 и 3.3. Указано, к какому кластеру принадлежат генерируемые 10 подвыборок, значения параметров w этих кластеров и значения сдвига θ для каждой подвыборки. Алгоритм 3.3 показал лучшие результаты в эксперименте с квадратичной моделью — лучше восстановились значения параметров подвыборок первого кластера.

Таблица 1: Результаты эксперимента на синтетических данных

	Сгенерированные значения	Алгоритм 3.2	Алгоритм 3.3
Линейная модель			
Метки кластеров подвыборок	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)
Параметры моделей кластеров w	(1, 5)	(0.97, 5.02)	(0.97, 5.02)
Параметры инвариантных преобразований θ	(0, 0.2, 0.3, 0.4, 0.5, 0, 0.1, 0.2, 0.3, 0.4)	(-0.0092, 0.12, 0.29, 0.37, 0.56, 0.00040, 0.11, 0.19, 0.30, 0.40)	(-0.034, 0.17, 0.27, 0.36, 0.56, 0.00093, 0.11, 0.20, 0.30, 0.40)
Квадратичная модель			
Метки кластеров подвыборок	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)	(1, 1, 1, 1, 1, 2, 2, 2, 2, 2)
Параметры моделей кластеров w	(1, 5)	(0.0076, 5.02)	(0.90, 4.96)
Параметры инвариантных преобразований θ	(0, 0.2, 0.3, 0.4, 0.5, 0, 0.1, 0.2, 0.3, 0.4)	(-0.034, 0.22, -0.16, -2.03, -34.43, -0.016, 0.083, 0.18, 0.29, 0.40)	(-0.024 0.20, 0.28, 0.45, 0.54, -0.0056, 0.094, 0.19, 0.30, 0.40)

4.2 Эксперименты на реальных данных по прогнозированию временных рядов

Цель эксперимента на реальных данных — продемонстрировать работу предложенного алгоритма прогнозирования. Для эксперимента был взят временной ряд ЭКГ [1]. Длина ряда 4170 отсчетов. В качестве тестовой выборки взяты последние 20 отсчетов. Прогноз делается на один отсчет вперед. В качестве предыстории для каждого элемента тестовой выборки брались предыдущие 60 отсчетов.

Была зафиксирована прогностическая модель

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

и семейство инвариантных преобразований SIM

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

Это семейство соответствует всевозможным сдвигам и растяжениям по обеим осям. Применялся алгоритм итеративной кластеризации подвыборок и оценки параметров 3.3. Было взято $C = 2$, т.е. подвыборки разделялись на 2 кластера (участки временного ряда, похожие и не похожие на предысторию).

Результаты прогнозирования — см. рис 6. Зеленый цвет — последняя часть временного ряда, красный — предыстория, фиолетовый — прогноз, синий — реальные значения, черный — прогноз методом локального прогнозирования, использующим расстояние в пространстве данных. Этот метод взят из работ [4], [5]. Метод прогнозирования с помощью инвариантных преобразований дал несколько меньшую ошибку: $\mathrm{rMSE}_{\mathrm{invariant}} = 0.0108$, $\mathrm{rMSE}_{\mathrm{localAlg}} = 0.0122$. Данная ошибка усреднялась по точкам тестовой выборки. На рис. 7 отдельно показан прогноз.

На рисунке 8 визуализировано попарное расстояние между подвыборками на последней итерации алгоритма, отсортированное по расстоянию до предыстории. Видно, что выделяется кластер участков, похожих на предысторию. Также выделяются другие кластера участков, похожих между собой.

На рисунке 9 изображены апостериорные распределения параметров для разных кластеров на последней итерации алгоритма. На рисунках сверху — распределения из кластера, в который попала предыстория, а снизу — распределения из другого кластера. Видно, что внутри кластера ковариационные матрицы близки.

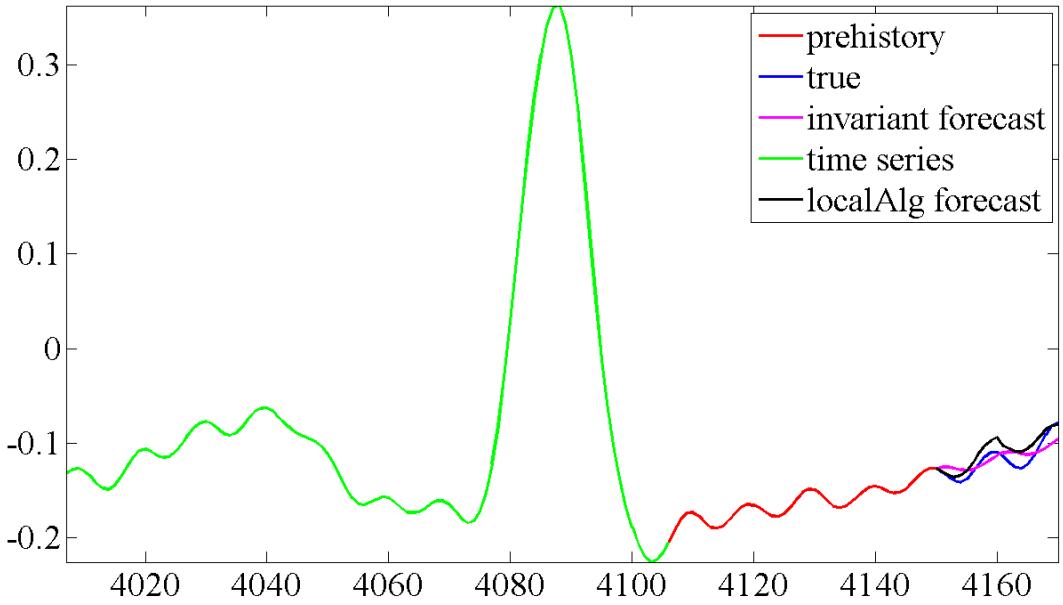

Рис. 6: Результаты прогнозирования

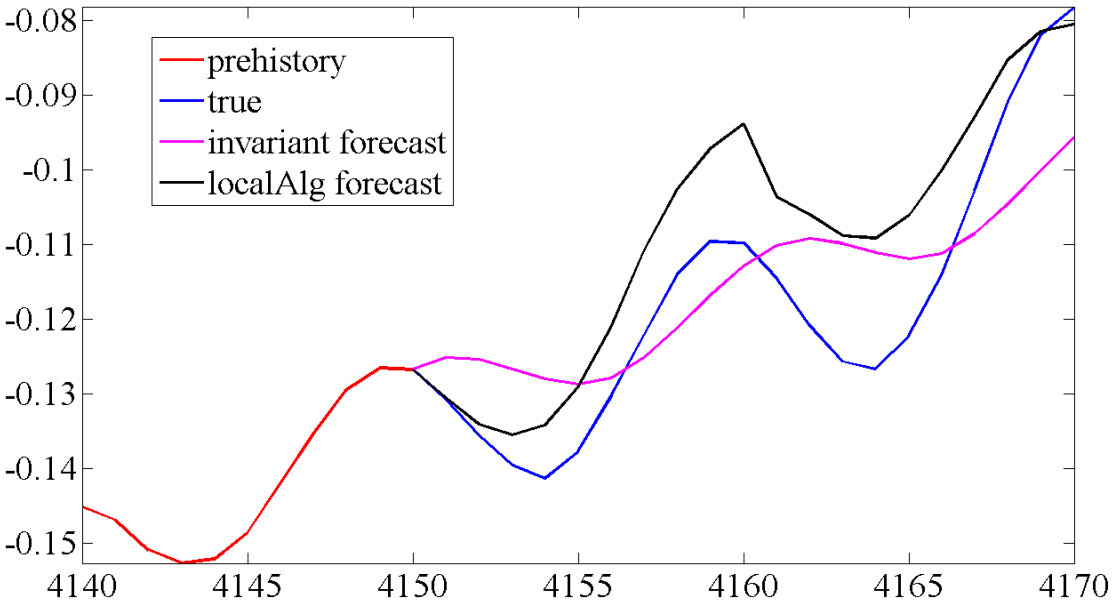

Рис. 7: Прогноз

5 Заключение

В данной работе рассматривается проблема оценки параметров инвариантных преобразований в задачах локального прогнозирования временных рядов. Предлага-

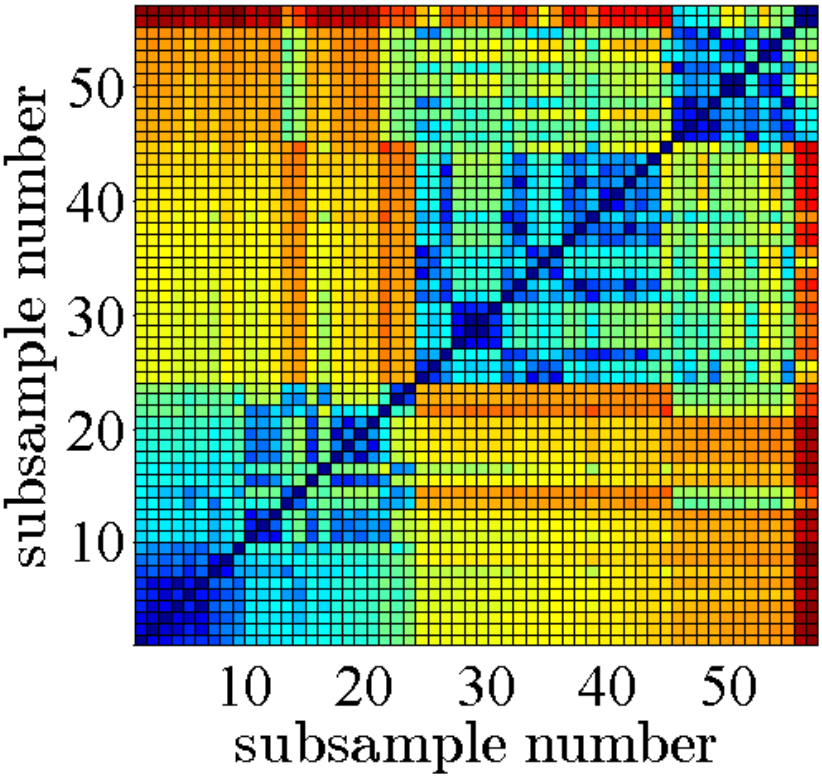

Рис. 8: Расстояние между подвыборками

ется в качестве функции расстояния между участками временного ряда использовать расстояние между моделями этих участков. Рассматривается параметрические модели участков, и в качестве расстояния берется расстояние Йенсена-Шеннона между апостериорными распределениями параметров моделей.

Для описания ситуация, когда сравниваемый с предысторией участок может быть изменен, вводится понятие инвариантного преобразования как преобразования модели этого участка. С помощью инвариантных преобразований и параметрический моделей предлагается способ описания семейства регрессионных подвыборок, применимый не только к задачам прогнозированиям временных рядов. В данном способе каждая подвыборка принадлежит некоторому кластеру и одновременно ей соответствует индивидуальное инвариантное преобразования. Модель порождения этой подвыборки — прогностическая модель кластера, измененная инвариантным преобразованием.

В работе предложен метод совместной оценки параметров моделей и инвариантных пребразований и кластеризации подвыборок. Данный метод применим к широкому классу задач, в которых есть несколько объектов измерения и необходимо

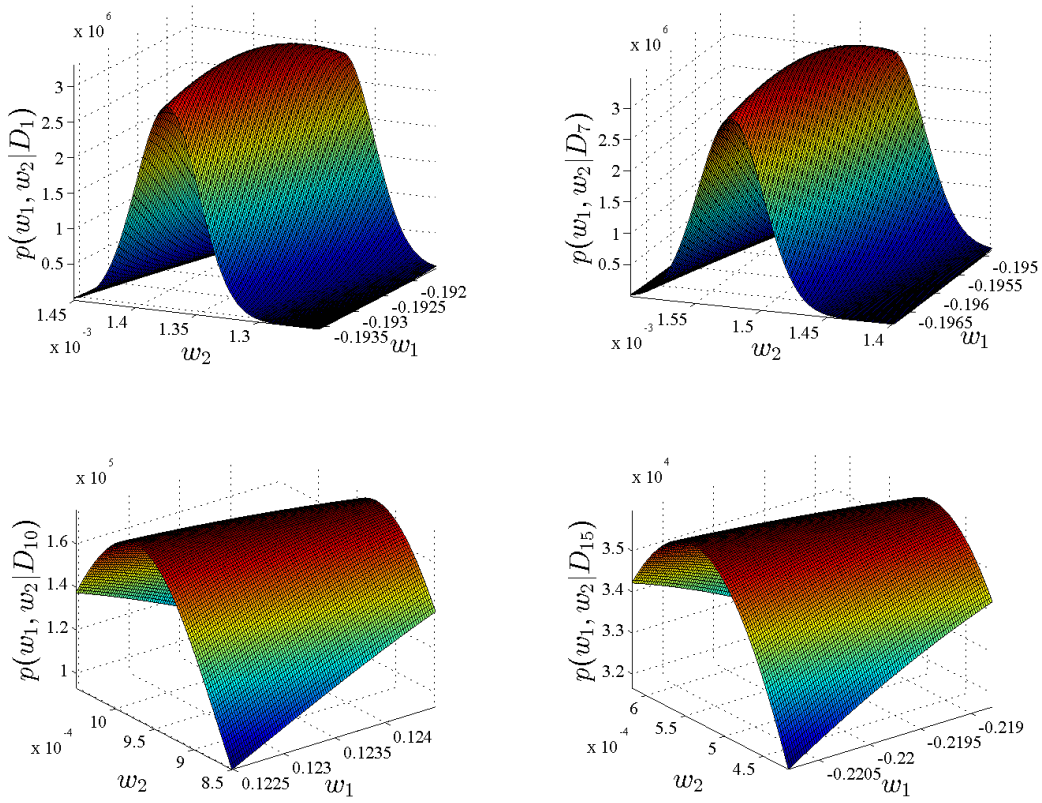

Рис. 9: Совместная плотность распределения первых двух параметров на подвыборках 1, 7, 10 и 15

устанавливать сходство между ними и строить общие модели. Предложенный метод продемонстрирован в вычислительных экспериментах на модельных и реальных данных в среде MATLAB.

Итак, кратко перечислим основные результаты.

- 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.

Список литературы

[1] http://www.clear.rice.edu/elec301/projects02/empiricalmode/code.html.

[2] А.А. Варфоломеева. Локальные методы прогнозирования с выбором метрики. *Машинное обучение и анализ данных*, 1(3):367–375, 2012.

[3] А. А. Токмакова and В. В. Стрижов. Оценивание гиперпараметров линейных регрессионных моделей при отборе шумовых и коррелирующих признаков. *Информатика и её применения*, 6(4):66–75, 2012.

[4] В.П. Федорова. Локальные методы прогнозирования временных рядов. Master's thesis, МГУ, 2009.

[5] С.В. Цыганова. Локальные методы прогнозирования с выбором преобразования. *Машинное обучение и анализ данных*, 1(3):311–317, 2012.

[6] Naomi Altman and Julio Villarreal. Self-modelling regression for longitudinal data with time-invariant covariates. *Canadian Journal of Statistics*, 32(3):251–268, 2004.

[7] C.M. Bishop et al. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[8] Dominik M. Endres and Johannes E. Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860, 2003.

[9] Chunlei Ke and Yuedong Wang. Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*, 96(456):1272–1298, 2001.

[10] Alois Kneip and Joachim Engel. Model estimation in nonlinear regression under shape invariance. *The Annals of Statistics*, 23(2):551–570, 1995.

[11] Alois Kneip and Theo Gasser. Convergence and consistency results for self-modeling nonlinear regression. *The Annals of Statistics*, pages 82–112, 1988.

[12] Alois Kneip and Theo Gasser. Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics*, 20(3):1266–1305, 1992.

[13] WH Lawton, EA Sylvestre, and MS Maggio. Self modeling nonlinear regression. *Technometrics*, 14(3):513–532, 1972.

[14] James McNames. *Innovations in local modeling for time series prediction*. PhD thesis, Stanford University, 1999.

[15] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer Science+ Business Media, 2006.

[16] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer. On rhythm and general music similarity. 2009.

[17] Dominik Schnitzer. Multivariate normals (mvn) toolbox, http://www.ofai.at/ dominik.schnitzer/mvn/.

[18] T Velmurugan and T Santhanam. Computational complexity between k-means and k-medoids clustering algorithms for normal and uniform distributions of data points. *Journal of Computer Science*, 6(3):363, 2010.