

На правах рукописи



Мотренко Анастасия Петровна

ВЫБОР МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ
МУЛЬТИКОРРЕЛИРУЮЩИХ ВРЕМЕННЫХ РЯДОВ

05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2019

Работа выполнена на кафедре интеллектуальных систем федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский институт)».

Научный руководитель:

Стрижов Вадим Викторович

доктор физико-математических наук, федеральный исследовательский центр «Информатика и управление» Российской академии наук, отдел интеллектуальных систем, ведущий научный сотрудник.

Официальные оппоненты:

Хачай Михаил Юрьевич

доктор физико-математических наук, профессор РАН, федеральное государственное бюджетное учреждение науки Институт математики и механики им. Н.Н. Красовского УрО РАН, заведующий отделом математического программирования

Панов Максим Евгеньевич

кандидат физико-математических наук, автономная некоммерческая образовательная организация высшего образования «Сколковский институт науки и технологий», старший преподаватель

Ведущая организация:

Федеральное государственное бюджетное образовательное учреждение высшего образования «Тульский государственный университет».

Защита состоится «12» сентября 2019 года в 13:00 на заседании диссертационного совета Д 002.073.05 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук» и на сайте <http://www.frccsc.ru/>

Автореферат разослан «__» ____ 2019 года.

Ученый секретарь

диссертационного совета Д 002.073.05
д.ф.-м.н., профессор



В.В.Рязанов

Общая характеристика работы

Актуальность темы. Исследование направлено на решение проблемы построения моделей прогнозирования и классификации временных рядов в условиях избыточного признакового пространства. Работа посвящена разработке методов анализа наборов мультикоррелирующих временных рядов: классификации, прогнозирования и декодирования временного ряда. Следствием избыточности — наличия мультикоррелирующих признаков — в задачах регрессии является неустойчивость оценок параметров и рост их дисперсии, снижение обобщающей способности построенной регрессионной модели (Стрижов: 2014).

Задача *декодирования* временного ряда состоит в восстановлении значений ряда по истории наблюдений набора временных рядов. Отличие задачи декодирования от задачи прогнозирования временного ряда заключается в том, что исторические значения целевого временного ряда не используются для решения задачи. Отличие от задачи регрессии в том, что признаковое описание не определено, и его выделение является частью решения задачи декодирования.

Для признакового описания временного ряда в задаче декодирования временных рядов характерно наличие корреляций сложной многоиндексной (пространственной, частотной и временной) структуры. Для описания объектов — отсчетов времени — в таком признаковом пространстве используются многоиндексные матрицы. Применение методов регрессии и классификации, разработанных для двухиндексного матричного представления данных — шаговой регрессии (Bougrain: 2009), регуляризации (Bundy: 2016), линейного дискриминантного анализа (Sellers: 2010), метод частичных наименьших квадратов (Shimoda: 2011) — к многоиндексным данным требует либо модификации методов, либо уплотнения данных.

В работе исследуется фильтрационный подход к выбору признаков. Предложена многоиндексная модификация фильтрационного метода QPFS выбора признаков (Katrutsa: 2017). Результат ранжирования признаков с помощью фильтрационного метода не привязан к конкретной модели и в этом смысле является универсальным. Преимуществом фильтрационных методов отбора признаков перед методами снижения размерности, основанными на преобразованиях признаков является интерпретируемость и разреженность получаемых результатов. Выбор наиболее значимых признаков предоставляет ценную экспертную информацию. На текущий момент методы выбора признаков, учитывающие многоиндексную структуру признаков (Zhao: 2013, Cao: 2014, Kim: 2007, Eliseyev: 2011) относятся либо встроенным, либо к оберточным методам. Таким образом, разработка фильтрационного метода отбора признаков, обладающих многоиндексной структурой, является актуальной задачей.

Для анализа временных рядов в метрических пространствах в работе предложен метод снижения размерности в задачах классификации, основанный на вложении многообразий (Fefferman: 2013). Предполагается что объекты содержатся вблизи нелинейного многообразия: геодезически близкие в исходном пространстве объекты метрически далеки друг от друга. Линейные методы, основанные на анализе дисперсии — латентно-семантический анализ (Brooke: 2012), анализ главных компонент (Горбань: 2008) — нарушают локальную структуру

близости между объектами при нелинейных многообразий. Метод стохастических вложений, сохраняющий вероятностную структуру близости между объектами, не предусматривает вложения новых данных в результирующее пространство без повторного вложения обучающих данных. Предложенная в работе модификация метода стохастического вложения ближайшими соседями (van der Maaten: 2008) учитывает целевую переменную, что позволяет повысить качество классификации в пространстве сниженной размерности. Предложенный метод расширяет границы применимости оригинального метода стохастических вложений.

Исследуются методы построения выборки оптимальным образом с точки зрения баланса между размерностью выборки и ее информативностью. При высокой частоте сэмплирования значения сигнала в соседние моменты времени сильно скоррелированы между собой и могут быть частично опущены из рассмотрения. Статистические методы (Self: 1988, Self: 1992, Shieh: 2000, Demidenko: 2007) имеют ряд ограничений, связанных с применением на практике. Для оценки объема выборки необходимо оценить дисперсию оценки параметров, или в более общем случае, иметь оценку параметра нецентральности в распределении используемой статистики при истинности альтернативной гипотезы. Статистические методы оценки объема выборки не отвечают на вопрос о том, как получить эти значения. Кроме того, дисперсия оценок и параметр нецентральности так же будут получены с некоторой дисперсией, влияние которой на результат оценки объема выборки не учитывается. Исследуются байесовские методы оценки достаточного объема выборки.

Цели работы.

1. Предложить метод выбора модели прогнозирования мультикоррелирующих временных рядов для решения задачи декодирования временных рядов.
2. Предложить метод выбора признаков, учитывающий многоиндексную структуру данных. Метод должен быть вычислительно простым и учитывать связь между признаками.
3. Предложить метод снижения размерности, сохраняющий метрическую структуру данных. Предполагается, что данные располагаются вблизи нелинейного многообразия, вложенного в пространство большей размерности.
4. Предложить критерий информативности объекта для построения выборки оптимального состава. Использовать построенный критерий для оценки объема выборки и выбора оптимальной модели.
5. Разработать алгоритм построения модели декодирования и провести вычислительный эксперимент для сравнения различных методов выбора признаков при решении задачи декодирования временных рядов.

Основные положения, выносимые на защиту.

1. Предложен метод выбора признаков, учитывающий многоиндексное мультикоррелирующее представление данных. Метод учитывает связи между

признаками независимо от модели прогнозирования, не перебирая все комбинации признаков. Исследованы свойства предложенного метода. Сформулированы условия локальной сходимости предложенного алгоритма. Доказана приближенная оптимальность низкорангового релаксированного решения.

2. Разработан метод снижения признакового пространства, сохраняющий локальную структуру близости между объектами. Исследована сходимость предложенного алгоритма стохастического вложения.
3. Разработан метод выделения сегментов временных рядов для интерпретируемого сегментирования квазипериодических сигналов.
4. Приведены рекомендации по практическому использованию разработанных алгоритмов. Рассмотрены практические задачи декодирования временных рядов, выделения внутреннего плагиата, задачи агрегирования и прогнозирования временных рядов грузовых железнодорожных перевозок.

Методы исследования. Для решения задачи декодирования временных рядов используется многоиндексное представление данных (Leuthardt: 2006, Shimoda: 2011, Elisseyev: 2011). Для формулировки оптимизационной задачи выбора признаков в многоиндексной постановке используются матричные разложения высшего порядка (Cichocki: 2013). Для построения признакового пространства в задаче декодирования применяются спектрально-временные преобразования. Для получения оптимального признакового описания используются различные методы выбора признаков и снижения размерности (Katrutsa: 2017, Zhao: 2013, Wang: 2005).

Для снижения размерности признакового пространства в метрических пространствах используются методы обучения вложений (van der Maaten: 2008). Критерий близости выборок при оценке оптимального объема выборки формулируется с помощью дивергенции Кульбака-Лейблера (Kullback: 1959, Gibbs: 2002) между апостериорными распределениями параметров распределения подвыборок различного объема. Обоснование предложенного метода опирается на известные (Ali: 1966) свойства распределения Кульбака-Лейблера, а также полученные (Motrenko: 2014) на их основе результаты. Предлагаемый метод относится к классу байесовских методов оценки объема выборки (Lindley: 1997, Rubin: 1998, Wang: 2002), в котором оценка объема выборки определяется исходя из максимизации ожидаемой функции полезности (Lindley: 1997). Функция полезности может включать в явном виде функции распределения параметров и штрафы за увеличение объема выборки.

Научная новизна.

1. Предложен метод выбора признаков, учитывающий многоиндексную структуру данных. Метод основан на решении задачи квадратичного программирования.
2. Предложен метод снижения размерности признакового пространства в задаче классификации, учитывающий как локальную структуру близости между объектами, так и целевую переменную.

3. Предложен метод сегментирования квазипериодических временных рядов с помощью анализа фазовой траектории главных компонент временного ряда.
4. Предложен метод оценки достаточного объема выборки, основанный на сравнении оценок функции распределения данных, полученных на различных подвыборках исследуемой выборки. Оценка достаточного объема выборки зависит от модели классификации или регрессии.

Теоретическая значимость. Работа в целом носит теоретический характер. Предложен метод выбора признаков, учитывающий многоиндексную структуру данных. Предложен приближенный алгоритм решения, основанный на низкоранговом разложении тензора индикаторных переменных. Получены условия сходимости решения задачи оптимизации, лежащей в основе предложенного метода. Доказана приближенная оптимальность решения. Предложен метод снижения размерности с сохранением локальной структуры в задачах обучения с учителем. Продемонстрирована устойчивость алгоритма стохастических вложений к параметрам разбиения выборки. Исследованы теоретические свойства предложенных методов. Предложен критерий информативности добавляемого в выборку объекта, исследованы его статистические свойства.

Практическая значимость. Предложенный метод выбора признаков предназначен для построения адекватных и вычислительно простых моделей реконструкции траектории движений по кортикальной активности с целью создания НКИ. Разработка методов для НКИ имеет важное значение для развития вспомогательных технологий и реабилитационных устройств. Разработанные методы направлены на повышение точности и устойчивости результатов прогнозирования движений пациентов, в частности за счет снижения сложности моделей.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах прогнозирования и классификации временных рядов; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях.

1. Sample size estimation in classification problems // International Digital Processing Conference, Barcelona, 2016 (Motrenko: 2016).
2. Оценка объема выборки в задачах классификации // Математические методы распознавания образов, г. Светлогорск, Калининградская область, 2015 (Motrenko: 2015).
3. Bayesian Sample Size Estimation for Patient Classification Survey // 20th Conference of the International Federation of Operational Research Societies, Barcelona, 2014 (Motrenko: 2014a).
4. Small CVD sample set classification: generative versus discriminative // 26th European conference on operation research, Rome, 2013 (Motrenko: 2013).

5. Multiclass classification of cardio-vascular disease patients with sample size estimation // 25th European conference on operation research, Vilnius, 2012 (Motrenko: 2012).

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в 15 печатных изданиях, 9 из которых изданы в журналах, рекомендованных ВАК.

Личный вклад. Результаты получены самостоятельно под научным руководством д.ф.-м.н. В. В. Стрижова.

Структура и объем работы. Диссертация состоит из оглавления, введения, пяти разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из 170 наименований. Текст работы занимает 110 страниц.

Основное содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулированы цели и методы исследования, поставлены основные задачи, обоснована научная новизна, теоретическая и практическая значимость полученных результатов.

В **главе 1** приводится формальная постановка задач декодирования временных рядов и задачи выбора признаков. Вводятся основные определения и обозначения, функции качества модели декодирования временных рядов. Описаны этапы построения выборки: сегментирование временного ряда, извлечение признаков.

Определение 1. *Временным рядом* называется функция дискретного аргумента $\mathbf{s}(t)$, сопоставляющая отсчетам времени $t_i \in \mathcal{T}$ вектор значений измеряемых переменных $\mathbf{s}(t_i) = \mathbf{s}_i \in \mathbb{R}^d$.

Определение 2. Пусть заданы набор временных рядов $\{\mathbf{s}_n(t)\}_{n=1}^{N_s}$ и набор целевых временных рядов $\{\mathbf{y}_n(t)\}_{n=1}^{N_y}$. Задача восстановления значений $\{\mathbf{y}_n(t)\}_{n=1}^{N_y}$ по предыдущим значениям $\{\mathbf{s}_n(t)\}_{n=1}^{N_s}$ называется *задачей декодирования временных рядов* $\{\mathbf{y}_n(t)\}_{n=1}^{N_y}$.

Отличие задачи декодирования от задачи прогнозирования временного ряда заключается в том, что исторические значения $\{\mathbf{y}_n(t)\}_{n=1}^{N_y}$ не используются для построения признаковового описания при решении задачи. Отличие от задачи регрессии в том, что признаковое описание не определено, и его построение является частью решения задачи декодирования.

Задача линейной регрессии для многоиндексных матриц. Выборка $(\underline{\mathbf{D}}, \mathbf{Y})$ состоит из временных рядов $\mathbf{s}(t)$, $\mathbf{y}(t)$:

$$\underline{\mathbf{D}} \in \mathbb{R}^{M \times n_1 \times \dots \times n_D}, \quad D_{(m, :, :, :)} = \underline{\mathbf{X}}_m, \quad \mathbf{Y} = [\mathbf{y}_1^\top, \dots, \mathbf{y}_M^\top]^\top, \quad (1)$$

где $\mathbf{y}_m = \mathbf{y}(t_m)$, а $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \dots \times n_D}$ – многоиндексная матрица. Требуется восстановить значения $\mathbf{y}_m$ при заданной матрице $\underline{\mathbf{X}}_m$, $m = 1, \dots, M$. Приближим

целевой временной ряд с помощью линейной модели

$$\hat{\mathbf{y}}_m = \text{vec}(\mathbf{X}_m)^\top \hat{\mathbf{W}}, \quad (2)$$

где вектор весов $\hat{\mathbf{W}} \in \mathbb{R}^{n_1 \cdots n_D \times N_y}$ минимизирует сумму остатков:

$$\hat{\mathbf{W}} = \arg \min_{\mathbf{W}} \frac{1}{M \cdot N_y} \|\hat{\mathbf{Y}} - \mathbf{Y}\|_F^2. \quad (3)$$

Для внешней оценки качества прогнозирования рассматриваются ошибка среднеквадратичного отклонения sMSE (scaled mean square error)

$$\text{sMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^M \|\hat{\mathbf{y}}_m - \mathbf{y}_m\|_2}{\sum_{m=1}^M \|\bar{\mathbf{y}} - \mathbf{y}_m\|_2}, \quad \bar{\mathbf{y}} = \frac{1}{M} \sum_{i=1}^M \mathbf{y}_m, \quad (4)$$

и коэффициент корреляции $\hat{\mathbf{Y}}$ между спрогнозированными и реальными значениями целевых временных рядов $\mathbf{Y}$

$$\text{corr}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{N_y} \sum_{n=1}^{N_y} \frac{\text{cov}(\hat{\mathbf{y}}_n, \mathbf{y}_n)}{\sqrt{\text{cov}(\hat{\mathbf{y}}_n, \hat{\mathbf{y}}_n) \text{cov}(\mathbf{y}_n, \mathbf{y}_n)}}, \quad (5)$$

$$\text{cov}(\hat{\mathbf{y}}_n, \mathbf{y}_n) = \frac{1}{M} \sum_{m=1}^M (\hat{y}_{mn} - \bar{\hat{y}}_n)(y_{mn} - \bar{y}_n),$$

где $\mathbf{y}_n$ – вектор-столбец матрицы $\mathbf{Y}$.

Определение 3. Признак χ_i называется *релевантным* в задаче (2), если для некоторого достаточно малого $\delta > 0$ и $\mathbf{y}_j$ выполнено

$$\text{corr}(\chi_i, \mathbf{y}_j) \geq 1 - \delta.$$

Определение 4. Признаки χ_i, χ_j называется *коррелирующими*, если для некоторого достаточно малого $\delta > 0$ выполнено

$$\text{corr}(\chi_i, \chi_j) \geq 1 - \delta.$$

Определение 5. Множество признаков $\chi_1, \dots, \chi_K$ называется *мульти-коррелирующими*, если любая пара признаков $\chi_i, \chi_j, i, j \in \{1, \dots, K\}$ из этого множества коррелируют.

Определение 6. Признаковое описание в задаче (2) называется *избыточным*, если существует такое преобразование признаков $\mathbf{X} = [\chi_1, \dots, \chi_N] \mapsto \mathbf{Z} = [\chi'_1, \dots, \chi'_K]$, такое что $N < K$, и

$$\mathbb{E} \|\mathbf{y} - f(\mathbf{X})\|^2 \geq \mathbb{E} \|\mathbf{y} - g(\mathbf{Z})\|^2,$$

где f и g – оптимальные модели (2), построенные для признаковых описаний $\mathbf{X}$ и $\mathbf{Z}$.

Мультикоррелирующие признаки содержат избыточную информацию. В работе предложены методы отбора признаков (3) для снижения вычислительных затрат и повышения устойчивости решения.

В данной работе развивается фильтрационный метод QPFS (Katrutsa: 2017). Фильтрационные методы ранжируют признаки вне зависимости от модели, поэтому они вычислительно более эффективны чем методы, требующие настройки модели. Метод QPFS позволяет сбалансировать сходство между признаками и релевантность выбранных признаков задаче регрессии, не перебирая всех возможных комбинаций признаков. *Сходство* между i -м и j -м признаками задается как

$$q_{ij} = |\text{corr}(\mathbf{x}_i, \mathbf{x}_j)|, \quad (6)$$

где $\mathbf{x}_i, \mathbf{x}_j$ – столбцы матрицы $\mathbf{X}$. Согласно определению 3, *релевантность* признака $\mathbf{x}_i$ для прогнозирования целевой переменной $\mathbf{Y}$ задается как

$$b_i = \frac{1}{N_y} \sum_{n=1}^{N_y} |\text{corr}(\mathbf{x}_i, \mathbf{y}_n)|. \quad (7)$$

Задача выбора признаков формулируется в виде задачи квадратичного программирования

$$\mathbf{a} = \arg \min_{\mathbf{a} \in \{0,1\}^N} (\mathbf{a}^T \mathbf{Q} \mathbf{a} - \mathbf{b}^T \mathbf{a}), \quad (8)$$

где $\mathbf{Q} \in \mathbb{R}^{N \times N}$ — матрица попарного сходства между признаками, $\mathbf{b}$ — вектор релевантностей признаков. Результат выбора признаков зависит от выбранного способа измерять сходство $\mathbf{Q}$ между признаками и их релевантность $\mathbf{b}$. Зафиксируем такой способ определения $\mathbf{Q}$, при котором значения элементов матрицы $\mathbf{Q}$ лежат на отрезке $[0, 1]$, а все диагональные элементы равны 1.

Оптимизационный критерий в задаче (8) уравнивает сходство между признаками из выбранного набора и их предсказательную способность, варьируя значения индикаторной переменной $\mathbf{a} \in [0, 1]^N$, определяющей набор выбранных признаков:

$$\mathbf{X}_{\mathbf{a}} = [\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_K}], \text{ где } a_{i_k} = 1, k = 1, \dots, K.$$

Задача (8) сформулирована для случая двухиндексных матриц. При решении практических задач в данной работе рассматриваются многоиндексные признаковые описания $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \dots \times n_D}$. Предлагается тензорная модификация метода QPFS, использующая многоиндексную структуру признаков $\underline{\mathbf{X}}_m$. В частности, для каждого направления вводится отдельная матрица сходства $\mathbf{Q}_d \in \mathbb{R}^{n_d \times n_d}$, $d = 1, \dots, D$. Многоиндексная матрица релевантности $\underline{\mathbf{B}}$ имеет ту же размерность, что и $\underline{\mathbf{X}}_m$.

Определение 7. Пусть функция

$$\varphi : \{1, \dots, n_1\} \times \dots \times \{1, \dots, n_D\} \mapsto \{1, \dots, (n_1 \cdot \dots \cdot n_D)\}$$

задает некоторое перечисление наборов $(i_1, \dots, i_D)$ индексов многоиндексной матрицы $\underline{\mathbf{X}}$. Операцию, ставящую в соответствие многоиндексной матрице $\underline{\mathbf{X}}$ вектор

$$\mathbf{x} = \text{vec}(\underline{\mathbf{X}}), \text{ где } x_{\varphi(i_1, \dots, i_D)} = x_{i_1 \dots i_D}, \mathbf{x} \in \mathbb{R}^{n_1 \cdot \dots \cdot n_D}, \quad (9)$$

называется *векторизацией*.

Определение 8. Рассмотрим многоиндексную матрицу плана $\underline{\mathbf{D}} \in \mathbb{R}^{M \times n_1 \times \dots \times n_d}$, где m -ый объект описывается матрицей $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \dots \times n_d}$. Операцию, ставящую в соответствие многоиндексной матрице $\underline{\mathbf{D}}$ плоскую матрицу

$$\mathbf{X} = [\mathbf{x}_1^T, \dots, \mathbf{x}_M^T]^T \in \mathbb{R}^{M \times (n_1 \dots n_d)}, \text{ где } \mathbf{x}_m = \text{vec}(\underline{\mathbf{X}}_m), \quad (10)$$

называется *уплощением* матрицы $\underline{\mathbf{D}}$.

Обозначим через $\mathbf{X} \in \mathbb{R}^{M \times n_1 \dots n_d}$ уплощение матрицы $\underline{\mathbf{D}} \in \mathbb{R}^{n_1 \dots n_d \times M}$.

Определим индикаторную переменную $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times \dots \times n_d}$, кодирующую вхождения признаков $\chi_{i_1 \dots i_d}$ в модель (2). Соответствующую двухиндексную матрицу обозначим $\mathbf{X}_{\underline{\mathbf{A}}}$

$$\mathbf{X}_{\underline{\mathbf{A}}} = [\dots, \chi_{i_1 \dots i_d}, \dots], \text{ где } a_{i_1 \dots i_d} = 1.$$

Матрица $\mathbf{X}_{\underline{\mathbf{A}}}$ составлена из столбцов-признаков $\chi_{i_1 \dots i_d}$, для которых индикаторная переменная $a_{i_1 \dots i_d}$ принимает ненулевое значение.

В принятых обозначениях задача выбора признаков формулируется следующим образом:

$$\underline{\mathbf{A}} = \arg \min_{\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times \dots \times n_d}} \mathcal{L}(\mathbf{X}_{\underline{\mathbf{A}}} \mathbf{W}_{\underline{\mathbf{A}}}, \mathbf{Y}),$$

где $\mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y})$ — функция потерь, а $\mathbf{W}_{\underline{\mathbf{A}}}$ минимизирует квадратичные потери (3) для $\mathbf{X}_{\underline{\mathbf{A}}}$.

Введем обозначения для операций с многоиндексными матрицами. Через $\mathbf{a} \circ \mathbf{b}$ обозначим тензорное произведение двух векторов $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$

$$\mathbf{a} \circ \mathbf{b} \in \mathbb{R}^{n_1 \times n_2}, \quad [\mathbf{a} \circ \mathbf{b}]_{ij} = a_i b_j, \quad \mathbf{a} \in \mathbb{R}^{n_1}, \mathbf{b} \in \mathbb{R}^{n_2}.$$

Через $\underline{\mathbf{A}} \times_d \mathbf{B}$ обозначим произведение многоиндексной матрицы $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times \dots \times n_d}$ и матрицы $\mathbf{B} \in \mathbb{R}^{m \times n_d}$ вдоль направления d

$$[\underline{\mathbf{A}} \times_d \mathbf{B}]_{i_1 \dots i_d} = \sum_{i'=1}^{n_d} a_{i_1 \dots d-1 i' i_{d+1}} b_{ii'},$$

а через $\underline{\mathbf{A}} * \mathbf{B}$ обозначим поэлементное умножение

$$[\underline{\mathbf{A}} * \mathbf{B}]_{i_1 \dots i_d} = a_{i_1 \dots i_d} b_{i_1 \dots i_d}.$$

Пусть заданы матрицы сходства $\mathbf{Q}_d \in \mathbb{R}^{n_d \times n_d}$ для каждого направления $d = 1, \dots, D$ многоиндексной матрицы $\underline{\mathbf{X}}$ и задана матрица релевантности $\underline{\mathbf{B}}$. Переформулируем задачу (8) в виде

$$\underline{\mathbf{A}} = \arg \min_{\underline{\mathbf{A}} \in \{0,1\}^{n_1 \times \dots \times n_d}} \left(\sum_{d=1}^D (\underline{\mathbf{A}} \times_d \mathbf{Q}_d) * \underline{\mathbf{A}} - \underline{\mathbf{B}} * \underline{\mathbf{A}} \right) \times_1 \mathbf{1}_{n_1} \dots \times_D \mathbf{1}_{n_d}, \quad (11)$$

Операция $\underline{\mathbf{A}} \times_1 \mathbf{1}_{n_1} \times \dots \times_D \mathbf{1}_{n_d}$ эквивалентна суммированию всех элементов $\underline{\mathbf{A}}$:

$$\underline{\mathbf{A}} \times_1 \mathbf{1}_{n_1} \times \dots \times_D \mathbf{1}_{n_d} = \sum_{i_1=1}^{n_1} \dots \sum_{i_D=1}^{n_D} a_{i_1 \dots i_D}.$$

Предложено приближенное решение (11), основанное на низкоранговом приближении $\underline{\mathbf{A}}$, которое позволяет решать задачу (11) итеративно, таким образом что на каждом шаге решается задача квадратичного программирования. Предложенное решение опирается на существование точного низкорангового разложения для бинарной многоиндексной матрицы:

$$\underline{\mathbf{A}} = \sum_{r=1}^R \mathbf{a}_1^{(r)} \circ \dots \circ \mathbf{a}_D^{(r)}, \quad \mathbf{a}_d^{(r)} \in \mathbb{R}^{n_d}, \quad d = 1, \dots, D. \quad (12)$$

Перепишем функцию потерь (11) с учетом (12) в виде

$$F(\boldsymbol{\alpha}_1^{(k)}, \dots, \boldsymbol{\alpha}_d, \dots, \boldsymbol{\alpha}_D^{(k-1)}) = \boldsymbol{\alpha}_d^\top \left(\tilde{\mathbf{Q}}_d^{(k-1)} + \tilde{\mathbf{I}}_d^{(k-1)} \right) \boldsymbol{\alpha}_d - \tilde{\mathbf{B}}_d^{(k-1)} \boldsymbol{\alpha}_d, \quad (13)$$

где

$$\boldsymbol{\alpha}_i = [\mathbf{a}_i^{(1)\top}, \dots, \mathbf{a}_i^{(R)\top}]^\top \in \mathbb{R}^{nR}, \quad i = 1, \dots, D, \quad (14)$$

$\tilde{\mathbf{Q}}_d^{(k)}$ и $\tilde{\mathbf{I}}_d^{(k)}$ — блочно-диагональные матрицы с r -ми блоками $\tilde{\mathbf{Q}}_d^{(k,r)}$ и $\tilde{\mathbf{I}}_d^{(k-1)}$ вида

$$\tilde{\mathbf{Q}}_d^{(k,r)} = \mathbf{Q}_d \prod_{\substack{j=1 \\ j \neq d}}^D \|\mathbf{a}^{(k,r)}\|_j^2, \quad (15)$$

$$\tilde{\mathbf{I}}_d^{(k)} = \sum_{\substack{j=1 \\ j \neq d}}^D \left(\prod_{\substack{j'=1 \\ j' \neq d, j' \neq j}}^D \|\mathbf{a}^{(k,r)}\|_{j'}^2 \cdot \mathbf{a}_j^{(k,r)\top} \mathbf{Q}_j \mathbf{a}_j^{(k,r)} \right) \mathbf{I}_{n_d}, \quad (16)$$

и

$$\tilde{\mathbf{B}}_d^{(k)} = [\tilde{\mathbf{B}}_d^{(k,1)\top}, \dots, \tilde{\mathbf{B}}_d^{(k,R)\top}]^\top, \quad \tilde{\mathbf{B}}_d^{(k,r)} = \mathbf{B} \prod_{\substack{j=1 \\ j \neq d}}^D \times_j \mathbf{a}_j^{(k,r)}. \quad (17)$$

Задача целочисленной оптимизации (11) не является выпуклой. Используется релаксация невыпуклых ограничений $\underline{\mathbf{A}} \in \{0, 1\}^{n_1 \times \dots \times n_D}$

$$\boldsymbol{\alpha}_d^{(k)} = \arg \min_{\boldsymbol{\alpha} \in [0,1]^{nR}} (\boldsymbol{\alpha}_1^{(k)}, \dots, \boldsymbol{\alpha}_d, \dots, \boldsymbol{\alpha}_D^{(k-1)}), \quad (18)$$

После получения субоптимального решения $\hat{\underline{\mathbf{A}}}$ как решения релаксированной задачи, значения переменной $\hat{\underline{\mathbf{A}}}$ бинаризируются путем отсечения с некоторым порогом ϵ .

Бинаризованные переменные $\underline{\mathbf{A}}(\epsilon)$ определяют набор выбранных признаков $\mathbf{X}_{\underline{\mathbf{A}}}$. Выбор значения порога отсечения ϵ влияет на результат $\mathbf{X}_{\underline{\mathbf{A}}(\epsilon)}$ выбора признаков. Решение $\hat{\underline{\mathbf{A}}}$ релаксированной задачи задает на множестве признаков линейный порядок

$$\mathbf{x}_{i_1 \dots i_D} \preceq \mathbf{x}_{i'_1 \dots i'_D} \Leftrightarrow \hat{a}_{i_1 \dots i_D} \leq \hat{a}_{i'_1 \dots i'_D}. \quad (19)$$

Предложенный алгоритм (1) совершает несколько итераций, на каждой из которых решается задача (18). Задача (18), решаемая относительно d -го направления, требует вычисления n_d^2 значений, что существенно меньше чем $(n_1 \dots n_D)^2$ в случае уплощенных матриц.

Algorithm 1 Многоиндексный алгоритм выбора признаков NQPFS.

Дано: начальное приближение $\xi^{(0)} = [\alpha_1^{(0)}, \dots, \alpha_D^{(0)}]$, матрицы $Q_1, \dots, Q_D$, $\underline{B}$, параметры $R, \epsilon, \varepsilon_F$.

Результат: многоиндексная матрица индикаторных переменных $\underline{A} \in \{0, 1\}^{n_1 \times \dots \times n_D}$.

```

 $k = 0$ 
while  $|F(\xi^{(k)}) - F(\xi^{(k+1)})| > \varepsilon_F$  do
  for  $1 \leq d \leq D$  do пересчитать матрицы  $Q_d^{(k-1)}, \tilde{I}_d^{(k-1)}, B_d^{(k-1)}$  соглас-
  но (15), (16), (17), получить новое приближение  $\alpha_d^{(k)}$ 
     $\alpha_d^{(k)} = \arg \min_{\alpha_d \in [0,1]^{n_d R}} F(\alpha_1^{(k)}, \dots, \alpha_{d-1}^{(k)}, \alpha_d^{(k)}, \alpha_{d+1}^{(k-1)}, \dots, \alpha_D^{(k-1)})$ 
  end for
   $k \leftarrow k + 1$ 
end while
 $\hat{\underline{A}} = \sum_{r=1}^R \mathbf{a}_1^{(r)} \times \dots \times \mathbf{a}_D^{(r)}$ 
 $\underline{A} = \underline{A}(\epsilon)$ 
  
```

Определение 9. Последовательность $\{\xi_k\}$ сходится к ξ^* линейно, если найдутся такие $n_0 > 0, \rho \in [0, 1)$, что для любого $k > n_0$ выполнено

$$\|\xi_{k+1} - \xi^*\| \leq \rho \|\xi_k - \xi^*\|.$$

Теорема 1. Для любого $k > 0$ выполнено

$$F(A^{(k+1)}) \leq F(A^{(k)}),$$

где $\{\alpha^{(k)}\}$ — последовательность решений, сгенерированных алгоритмом, $\{\underline{A}^{(k)}\}$ — соответствующая (12), (14) последовательность индикаторных переменных.

Обозначим векторизацию индикаторной переменной $\underline{A}$ через $\xi = \text{vec}(\underline{A}) = [\alpha_1^T, \dots, \alpha_D^T]^T \in \mathbb{R}^{(n_1 + \dots + n_D)R}$. Обозначим $\xi = [\mathbf{a}_1^T, \dots, \mathbf{a}_D^T]^T \in \mathbb{R}^{(n_1 + \dots + n_D)R}$.

Теорема 2. Пусть $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$ — решение, полученное с помощью алгоритма (1). Пусть $Q_1, \dots, Q_D$ — положительно определенные матрицы с положительными элементами. Пусть точка $\xi_* \in [0, 1]^{(n_1 + \dots + n_D)R}$ такова что

1) выполнены необходимые условия локального минимума $F(\xi)$:

$$\nabla_{\xi} F|_{\xi_*} = \mathbf{0}_{n_1 + n_2 + n_3}, \quad (20)$$

2) существует $\delta > 0$, такая что для всех $d = 1, \dots, D$ выполнено

$$\alpha_d^* = \arg \min_{\|\alpha_d - \alpha_d^*\| < \delta} F(\xi_{*\setminus d}), \quad \text{где } \xi_{*\setminus d} = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_{d-1}^* \\ \alpha_d \\ \alpha_{d+1}^* \\ \vdots \\ \alpha_D^* \end{bmatrix}, \quad (21)$$

3) для всех $r = 1, \dots, R$ выполнено

$$\underline{\mathbf{B}} \times_1 \left(\mathbf{a}_1^{(r)} \right)^* \cdots \times_D \left(\mathbf{a}_D^{(r)} \right)^* > 0. \quad (22)$$

Тогда существует ε -окрестность точки ξ^* , в которой функция $F(\xi)$ — выпуклая.

В главе 2 предложен метод снижения размерности признакового пространства, сохраняющий локальную структуру близости между объектами в исходном пространстве $\mathbb{R}^n$. Предполагается, что существует гладкое отображение $f: \mathbb{R}^d \rightarrow \mathbb{R}^n$, $d < n$, такое что $\mathbf{X} = f(\mathbf{Z}) + \varepsilon$.

Определение 10. Представление $\mathbf{Z} \subset \mathbb{R}^k$ объектов выборки $\mathbf{X}$ в результирующем пространстве назовем *вложением* $\mathbf{X}$ в $\mathbb{R}^k$.

Определение 11. Локальной структурой близости между объектами $\mathbf{X}$ называется распределение вероятности отношения ближайшего соседства

$$p_{ij} = P(\mathbf{x}_i \text{ и } \mathbf{x}_j \text{ — ближайшие соседи}).$$

Структура близости $p_{ij} = P(\mathbf{x}_i, \mathbf{x}_j)$ в исходном и $q_{ij} = Q(\mathbf{z}_i, \mathbf{z}_j)$ в результирующем пространстве моделируется как

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2m}, \quad p_{i|j} = \frac{\exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_k\|^2}{2\sigma_i^2}\right)},$$

$$q_{ij} = \frac{(1 + \|\mathbf{z}_i - \mathbf{z}_j\|^2)^{-1}}{\sum_{k \neq i} (1 + \|\mathbf{z}_i - \mathbf{z}_k\|^2)^{-1}}, \quad q_{ii} = 0, \quad i, j \in \{1, \dots, m\}.$$

Координаты $\mathbf{Z}$ объектов $\mathbf{X}$ в результирующем пространстве $\mathbb{R}^k$

$$\mathbf{Z} = [\mathbf{z}_1 \ \dots \ \mathbf{z}_m]^\top \subset \mathbb{R}^k \quad (23)$$

находятся путем минимизации дивергенции Кульбака-Лейблера

$$\mathbf{Z}_{\min} = \operatorname{argmin}_{\mathbf{Z} \in \mathbb{R}^{m \times k}} D_{\text{KL}}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \quad (24)$$

Для учета разметки обучающей выборки признаковая матрица $\mathbf{X}$ расширяется дополнительным столбцом признаков,

$$\tilde{\mathbf{X}} = (\mathbf{X} \mid \mu \mathbf{y}),$$

где μ — вес меток как признаков.

Рассмотрим задачу классификации с обучающей выборкой $\mathbf{X}_A$ и тестовой выборкой $\mathbf{X}_B$. Так как на этапе обучения данные $\mathbf{X}'$ могут быть недоступны, непараметрический метод стохастических вложений не применим для снижения размерности в задачах классификации. Назовем это *проблемой непросмотренных объектов*. Для ее решения предлагается минимизировать (24) независимо по различным подмножествам объектов. Пусть $\mathbf{X}_A \mapsto \mathbf{Z}_A$. При вложении $\mathbf{X}_B$ решается

$$Z_{BA} = \arg \min_{Z \in \mathbb{R}^{k \times m'}} \text{KL}(P_{BA} \parallel Q_{BA}) = \sum_{i=m+1}^{m+m'} \sum_{j=1}^m p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

несимметричные матрицы P_{BA} и Q_{BA} описывают близость объектов $\mathbf{X}_B$ к $\mathbf{X}_A$ и $\mathbf{Z}_B$ к $\mathbf{Z}_A$

$$P_{BA} = P(\mathbf{X}_B, \mathbf{X}_A) = [p_{ij}], \quad i \in A, \quad j \in B, \quad Q_{BA} = Q(\mathbf{Z}_B, \mathbf{Z}_A), \\ A = \{1, \dots, m\}, \quad B = \{m+1, \dots, m+m'\}.$$

В главе 3 предложен метод разбиения квазипериодического временного ряда на интерпретируемые сегменты. Квазипериодический временной ряд $\mathbf{s} = \{s(i)\}_{i=1}^m$ состоит из повторяющихся сегментов, представляющих собой трансформированный базовый сегмент. Трансформация включает изменение формы сегмента и масштабирование времени.

Определение 12. Квазипериодический временной ряд $\mathbf{s}$ определяется набором $\langle \mathbf{s}^*, a(i, s), f(i) \rangle$, так что

$$s(i) = a(i, s_{[f(i)]}), \quad (25)$$

где вектор $\mathbf{s}^* = [s_1, \dots, s_T]^\top$ описывает базовый сегмент, функция $a(i, s), i \in \{1, \dots, m\}$ модифицирует форму базового сегмента, а кусочно-заданная функция $f(i) \mapsto \{1, \dots, T\}$ задает масштабирование по времени. Функция f имеет конечное множество точек разрыва, на каждой из областей непрерывности f монотонно возрастает от 1 до T .

Точки $\mathcal{I}_S^* = \{i_1^*, \dots, i_S^*\} = \{i_t^* \mid f(i_t + 1) < f(i_t)\}$ считаются концами периодов, которые алгоритму сегментирования необходимо найти. Результатом решения задачи сегментирования квазипериодического ряда является разбиение $\mathcal{I}_S = \{i_1, \dots, i_S\}$ временного ряда $\mathbf{s}$ на последовательность $\mathcal{S} = \{S_1, \dots, S_T\}$ сегментов $S_t = [s(i_t), \dots, s(i_{t+1} - 1)]^\top$, соответствующих преобразованному базовому периоду (25).

При оценке результатов процедуры выделения периодов, сегментирование временного ряда рассматривается как задачу двухклассовой классификации набора точек $\{s(1), \dots, s(m)\}$. Для фиксированного набора истинных точек сегментации $\mathcal{I}_S^* = \{i_1^*, \dots, i_S^*\}$ класс каждого объекта $x(i)$ определяется как

$$x(i) \in \begin{cases} 1, & \text{если } i \in \mathcal{I}_S, \\ 0 & \text{иначе.} \end{cases}$$

Вводится параметр ϵ , в рамках которого отклонение точек $i_t \in \mathcal{I}_S$ от $\mathcal{I}_S^*$ не считается ошибкой классификации. Для оценки качества используется коэффициент F_1

$$F_1 = \frac{2TP}{2TP + FN + FP}. \quad (26)$$

Пусть временной ряд $\mathbf{s}$ представим в виде декомпозиции

$$\mathbf{s} = \hat{\mathbf{s}} + \tilde{\mathbf{s}} + \boldsymbol{\epsilon}, \quad (27)$$

где $\hat{\mathbf{s}}$ — нестационарный тренд, $\tilde{\mathbf{s}}$ — периодическая компонента временного ряда, а $\boldsymbol{\epsilon}$ — шум. Каждая составляющая декомпозиции (27) может быть приближена комбинацией главных компонент траекторной матрицы

$$\mathbf{H} = \begin{pmatrix} s(1) & s(2) & \dots & s(N) \\ s(2) & s(3) & \dots & s(N+1) \\ \vdots & \vdots & \ddots & \vdots \\ s(m-N+1) & s(m-N+2) & \dots & s(m) \end{pmatrix}^T$$

центрированного временного ряда $\mathbf{s}$

$$s(i) \mapsto s(i) - \frac{1}{m} \sum_{i=1}^m s(i), \quad i = 1, \dots, m.$$

Параметр N , определяющий длину наибольшего периода, детектируемого методом главных компонент называется длиной окна.

Траекторная матрица $\mathbf{H}$ используется для восстановления временного ряда X или отдельных его компонент через антидиагональное усреднение. Для аппроксимации временного ряда $\tilde{\mathbf{s}}$ с помощью сингулярного разложения $\mathbf{H}$

$$\frac{1}{N} \mathbf{H}^T \mathbf{H} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^T, \quad \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_N) \quad (28)$$

вычисляются главные компоненты $\mathbf{y}_j = \mathbf{H} \mathbf{v}_j$, соответствующие положительным значениям $\mathbf{H}^T \mathbf{H}$. Каждая главная компонента $\mathbf{y}_j$ отвечает некоторой составляющей траекторной матрицы $\mathbf{H}$

$$\mathbf{H} = \mathbf{H}_1 + \dots + \mathbf{H}_d, \quad \mathbf{H}_j = \sqrt{\lambda_j} \mathbf{v}_j \mathbf{y}_j^T.$$

Для извлечения периодической составляющей необходимо выделить соответствующие искомым периодам главные компоненты. Ниже будет показано, что соответствуют две главные компоненты с последовательными номерами j и $j+1$.

Теорема 3. Пусть временной ряд

$$\mathbf{s} = \{s(1), \dots, s(i), \dots, s(m)\}$$

имеет вид

$$s(i) = A_i \cos(2\pi w i + \phi) \quad (29)$$

с параметрами $w \in (0, 1/2)$, $\phi \in [0, 2\pi)$, $m \cdot w \in \mathbb{N}$, где коэффициенты A_i равномерно абсолютно ограничены сверху некоторой константой, $|A_i| < C$. Тогда главные компоненты $\mathbf{y}_1$ и $\mathbf{y}_2$ могут быть представлены в виде

$$y_1(l) = B_1(l) \cos(2\pi w l + \phi_1),$$

$$y_2(l) = B_2(l) \cos(2\pi w l + \phi_2),$$

$$\phi_1, \phi_2 \in [0, 2\pi), l = 1, \dots, m - N + 1$$

где разница между $|\phi_1 - \phi_2| \rightarrow \pi/2$.

Из теоремы следует, что каждая периодическая компонента с $T > 2$ ($w < 1/2$) может быть аппроксимирована с помощью двух главных компонент, синусной и косинусной функций с той же частотой что и искомая периодика.

Для выделения периода нет необходимости восстанавливать саму периодическую составляющую, достаточно выбрать соответствующую ей пару компонент $\mathbf{y}_j$ и $\mathbf{y}_{j+1}$ и рассмотреть ее траекторию $(\mathbf{y}_j, \mathbf{y}_{j+1})$. При рассеении траектории некоторой осью формируются полупериоды. Каждое из множеств $\mathcal{I}_S^+$ и $\mathcal{I}_S^-$ конечных точек этих полупериодов определяет сегментацию временного ряда. В качестве $\mathcal{I}_S$ выбирается множество $\mathcal{I}_S^\bullet$ с меньшим расстоянием между $(y_j(i_t^\bullet), y_{j+1}^\bullet)$

$$\mathcal{I}_S = \arg \min_{\bullet \in \{+, -\}} \sum_{i_t^\bullet, i_q^\bullet \in \mathcal{I}_S^\bullet} \|(y_j(i_t^\bullet), y_{j+1}(i_t^\bullet)) - (y_j(i_q^\bullet), y_{j+1}(i_q^\bullet))\|.$$

3. Оценка объема выборки

В главе 4 предложен байесовский критерий добавления объектов в выборку. Оценка объема выборки в байесовской постановке включает оценку апостериорного распределения $p(D|\mathbf{w})$ параметров модели. При отсутствии наблюдаемых данных, апостериорное распределение $p(\mathbf{w}|D) = p(D|\mathbf{w})p(\mathbf{w})/p(D)$ совпадает с априорным $p(\mathbf{w})$, как в классических методах оценки объема выборки. Разница между вторым и третьим случаями заключается только в способе оценки распределения $p(\mathbf{w}|D)$ — на основе сэмплированных, либо реально наблюдаемых данных. В остальном процедура оценки объема выборки не меняется.

Обозначим через $T = \langle L, m, \mathbb{P}, \xi \rangle$ критерий вида

$$\int L(\mathbf{y}, \mathbf{X}) \prod_{i=1}^m \mathbb{P}(y_i, \mathbf{x}_i) \prod_{i=1}^m dy_i \dots d\mathbf{x}_i \leq \xi \quad (30)$$

Так как истинное порождающее распределение $\mathbb{P}(y_i, \mathbf{x}_i)$ неизвестно, используется выборочная оценка математического ожидания 31.

$$\hat{T}(m) = \frac{1}{K} \sum_{k=1}^K L(\mathbf{y}_k, \mathbf{X}_k) \leq \xi, \quad (31)$$

где $(\mathbf{y}_k, \mathbf{X}_k)$ — выборки объема m , полученные из наблюдаемой выборки $D_{m'}$, $m \leq m'$.

Определение 13. Назовем объем m^* выборки из неизвестного распределения $\mathbb{P}(y, \mathbf{x})$ *достаточным* согласно критерию T , если T выполняется для всех $m \geq m^*$.

Предлагаемый в данной работе метод основан на наблюдении за апостериорным распределением параметров модели. Численным критерием для оценки объема выборки является устойчивость апостериорного распределения параметров, измеряемая как дивергенция Кульбака-Лейблера между апостериорными распределениями, полученными на близких подвыборках.

Дивергенция Кульбака-Лейблера $D_{\text{KL}}(Q||P) = \sum Q \ln \left(\frac{Q}{P} \right)$ — частный случай f -дивергенции

$$D_f(Q||P) = \sum P \cdot f \left(\frac{Q}{P} \right) \text{ для } f(t) = t \ln t.$$

Теорема 4. При $m \rightarrow \infty$ выполнено

$$\frac{2m}{f''(1)} \cdot D_f(\hat{P}_m||P) \rightarrow \chi_N^2,$$

где $\hat{P}_m$ — гистограмма, построенная по выборке X_m из распределения P .

Теорема 5. Пусть выполнены следующие условия:

- 1) функция f строго выпукла и дважды дифференцируема в единице,
- 2) распределение $P : \mathbb{R} \rightarrow [0, 1]$ таково, что для всех $x \in \mathbb{R}$ $P(x) \neq 0$,
- 3) выборки Z и Z' выбираются независимо из распределения P , причем $\frac{1}{\rho} \leq m'/m \leq \rho$ при всех m, m' для некоторого $0 < \rho \leq 1$.

Тогда случайная величина $\frac{2m}{f''(1)} \cdot D_f(\hat{P}_m||\hat{P}_{m'})$ в пределе ограничена сверху и снизу случайными величинами из распределения χ_N^2

$$C_1 \chi_N^2 \leq \frac{2m}{f''(1)} \cdot D_f(\hat{P}_m||P) \leq C_2 \chi_N^2, \text{ при } m \rightarrow \infty$$

$$C_1 \chi_N^2 \leq \frac{2m}{f''(1)} \cdot D_f(\hat{P}_m||\hat{P}_{m'}) \leq C_2 \chi_N^2, \text{ при } m, m' \rightarrow \infty.$$

Для оценки объема выборки решается задача двух выборок: $H_0 : P(x) \neq P'(x)$ при альтернативе $H_1 : P(x) = P'(x)$. При H_1 выполнено $C'_1 \chi_N^2(\Delta) \leq D_{\text{KL}}(\hat{P}_m, \hat{P}'_m) \leq C'_2 \chi_N^2(\Delta)$, где $\chi^2(\Delta)$ — нецентральное χ^2 -квадрат распределение с параметром нецентральности $\Delta = \frac{1}{\mathbb{D}p_i} \sum_{i=1}^N [\mathbb{E}(p_i - p'_i)]^2$. При H_1 с.в. $C_1 \chi_N^2 \leq D_{\text{KL}}(\hat{P}_m, \hat{P}'_m) \leq C_2 \chi_N^2$.

4. Прикладные задачи

В главе 5 продемонстрировано применение предложенных методов к прикладным задачам сегментирования квазипериодических временных рядов, декодирования временных рядов, выделения внутреннего плагиата.

Таблица 1. Коэффициент корреляции между восстановленной и истинной траекториями.

Эксперимент	Метод	$N = 10$	$N = 25$	$N = 200$	$N = 500$
А, 20090116	QPFS	0.333 ± 0.007	0.34 ± 0.008	0.414 ± 0.041	0.505 ± 0.015
	NQPFS	0.271 ± 0.028	0.307 ± 0.009	0.459 ± 0.014	0.511 ± 0.020
	PLS	0.306 ± 0.007	0.376 ± 0.012	0.476 ± 0.017	0.51 ± 0.010
	NPLS	0.0367 ± 0.006	0.125 ± 0.016	0.329 ± 0.021	0.358 ± 0.054
К, 20090602	QPFS	0.357 ± 0.017	0.382 ± 0.014	0.424 ± 0.014	0.468 ± 0.010
	NQPFS	0.29 ± 0.007	0.306 ± 0.003	0.381 ± 0.006	0.474 ± 0.017
	PLS	0.341 ± 0.012	0.369 ± 0.010	0.442 ± 0.008	0.416 ± 0.015
	NPLS	0.111 ± 0.009	0.167 ± 0.007	0.306 ± 0.012	0.344 ± 0.026

Задача декодирования временных рядов состоит в том, чтобы предсказать траектории конечности по временным рядам кортикальной активности. Признаковое описание объекта времени описывается трехиндексной матрицей, которая имеет пространственную, временную и частотную компоненты. Эти компоненты описывают изменение спектральных характеристик сигналов, снимаемых с различных электродов. В таблице 1 приведены значения коэффициента корреляции (5) между восстановленной и истинной траекториями запястья для методов QPFS, его тензорной модификации NQPFS, и наиболее часто применяемых для решения задачи декодирования методов PLS и NPLS. Приведены результаты для $N = 10, 25, 200, 500$ признаков, выбранных каждым методом. Для каждого из восьми экспериментов и каждого значения N лучший результат выделен жирным шрифтом.

В заключении представлены основные результаты диссертационной работы.

1. Поставлена задача декодирования временных рядов. Исследованы основные этапы построения моделей: сегментирование временного ряда, построение признакового описания, снижение размерности. Предложен алгоритм автоматической сегментации квазипериодического временного ряда на интерпретируемые сегменты.
2. Исследована проблема выбора признаков, обладающих многоиндексной структурой. Предложен метод выбора признаков, учитывающий многоиндексную структуру данных.
3. Предложен метод снижения размерности, сохраняющий локальную структуру близости в задачах обучения с учителем. Метод учитывает целевую переменную. Предложенный метод расширяет границы применимости метода стохастических вложений ближайшими соседями. В отличие от исходного метода стохастических вложений, разработанная модификация предусматривает вложение новых данных в результирующее пространство без повторного вложения обучающих данных, учитывает разметку обучающих данных и имеет параллельную реализацию.
4. Предложен метод оценки достаточного объема выборки. Предлагается считать объем выборки достаточным, если расстояние между распределениями, оцененными на подвыборках данного объема, достаточно мало. Кроме того, оценку можно производить как при наличии предположений о

распределении данных, так и в их отсутствие. Проанализированы способы последовательного добавления объектов в выборку.

5. Проведен анализ свойств предложенных методов. Проведено сравнение предложенных алгоритмов с существующими методами.

Публикации соискателя по теме диссертации

Публикации в журналах из списка ВАК.

1. Motrenko, A. Multi-way Feature Selection for ECoG-based Brain-Computer Interface / A. Motrenko, V. Strijov. // Expert Systems with Applications, 2018.
 2. Aduenko, A. Object selection in credit scoring using covariance matrix of parameters estimations / A. Aduenko, A. Motrenko, V. Strijov // Annals of Operations Research. – 2018. № 1. – P. 3–21.
 3. Rudakov, K.V. Selecting an Optimal Model for Forecasting the Volumes of Railway Goods Transportation / K.V. Rudakov, V.V. Strizhov, D.O. Kashirin, M.P. Kuznetsov, A.P. Motrenko, M.M. Stenina // Automation and Remote Control. – 2017. – V. 78. – №. 1. P. 74–87.
 4. Motrenko, A. Combining endogenous and exogenous variables in a special case of non-parametric time series forecasting model / A. Motrenko, K. Rudakov, V. Strijov // Moscow University Computational Mathematics and Cybernetics. – 2016. V. 36. – №. 2. – P. 71–78.
 5. Motrenko, A. Extracting fundamental periods to segment biomedical signals / A. Motrenko, V. Strijov // Journal of Biomedical and Health Informatics. – 2016. V. 20. – № 6. – P. 1466–1476.
 6. Молибог, И.О., Мотренко А.П. Методы снижения размерности в задаче обнаружения внутреннего плагиата / И. О. Молибог, А. П. Мотренко // Информатика и ее применения. – 2017, Т. 11. – № 3. – С. 59–71.
 7. Мотренко, А. П. Построение агрегированных прогнозов объемов железнодорожных грузоперевозок / А. П. Мотренко, В. В. Стрижов // Информатика и ее применения. – 2014. – Т. 8. – № 2. – С. 86–97.
 8. Motrenko, A. Bayesian sample size estimation for logistic regression / A. Motrenko, V. Strijov, G.-W. Weber // Journal of Computational and Applied Mathematics. – 2014. – № 255. – P. 743–752.
 9. Мотренко, А.П. Многоклассовая логистическая регрессия / А. П. Мотренко, В. В. Стрижов // Известия ТулГУ. – 2012. № 1. – С. 153–162.
- Остальные публикации.
10. Гасанов, Э.Э. Построение аппроксимирующего описания скалограммы в задаче прогнозирования движений по электрокортикограмме / Э.Э. Гасанов, А.П. Мотренко // Машинное обучение и анализ данных. – 2017. – Т. 3. – № 2. – С. 160–169.
 11. Kuznetsov, M. Methods for Intrinsic Plagiarism Detection and Author Diarization / M. Kuznetsov, A. Motrenko, R. Kuznetsova, V. Strijov // CEUR Workshop Proceedings. CLEF2016 Working Notes. – 2016. – V. 1609. – P. 912–919.

12. Вальков, А.С. Построение кросс-корреляционных зависимостей при прогнозе загруженности железнодорожного узла / А.С. Вальков, Е.М. Кожанов, А.П. Мотренко, Ф.И. Хусаинов // Машинное обучение и анализ данных. – 2013. – № 5. – С. 503–517.
13. Мотренко, А.П. Оценка плотности совместного распределения / А. П. Мотренко // Машинное обучение и анализ данных. – 2013. № 4. – С. 428–435.
14. Мотренко, А.П. Оценка необходимого объема выборки пациентов при прогнозировании сердечно-сосудистых заболеваний / А. П. Мотренко // Машинное обучение и анализ данных. – 2012. № 3. – С. 354–366.
15. Мотренко, А.П. Использование теста Грейнджера при прогнозировании временных рядов // Машинное обучение и анализ данных. – 2011. – № 1. – С. 51–60.