Московский физико-технический институт
(национальный исследовательский институт)

На правах рукописи

Мотренко Анастасия Петровна

ВЫБОР МОДЕЛЕЙ ПРОГНОЗИРОВАНИЯ МУЛЬТИКОРРЕЛИРУЮЩИХ ВРЕМЕННЫХ РЯДОВ

05.13.17 — Теоретические основы информатики

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
д.ф.-м.н. В. В. Стрижов

Москва — 2019

Оглавление

Введение

Актуальность темы. Исследование направлено на решение проблемы построения моделей прогнозирования и классификации в условиях избыточного признакового пространства. Следствием избыточности — наличия мультикоррелирующих признаков — в задачах регрессии является неустойчивость оценок параметров и рост их дисперсии, снижение обобщающей способности построенной регрессионной модели [1].

Для уменьшения размерности применяются методы преобразования и отбора признаков [2, 3]. При отборе признаков вводится функция качества, оценивающая релевантность различных поднаборов признаков решению задачи регрессии. Дополнительные ограничения на выбираемый набор признаков накладываются с помощью регуляризаторов. Например, в линейных моделях выбираются признаки с наибольшими значениями параметров модели. Чтобы ограничить значения параметров в методах Lasso [4] и LARS [5] в качестве регуляризатора используется взвешенная сумма модулей параметров модели, в гребневой регрессии [6] — взвешенная евклидова норма вектора параметров. В методе Elastic net [7] регуляризатором является линейная комбинация регуляризаторов Lasso и LARS. В шаговой регрессии [8] используется пошаговая стратегия выбора оптимального поднабора признаков. На каждом шагу алгоритма проверяется статистическая значимость добавляемого или удаляемого признака.

Работа посвящена разработке методов анализа наборов мультикоррелирующих временных рядов: классификации, прогнозирования и *декодирования* временного ряда. Задача декодирования временного ряда состоит в восстановлении значений ряда по истории наблюдений набора временных рядов. Число измеряемых временных рядов в наборе может превышать размерность целевого ряда. Измеряемые ряды могут отличаться от целевого частотой сэмплирования.

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. Признаковое описание в задаче декодирования сигналов электрокортикограмм имеет пространственную, частотную и временную размерности. Для описания объектов – отсчетов времени – в таком признаковом пространстве используются многоиндексные матрицы. Так как регистрируемые сигналы коррелирует для близко расположенных электродов и некоторых частот, необходимо снижать размерность признакового пространства либо отбирать признаки. Для двумерного представления данных в задачах декодирования нейросигналов используются шаговая регрессия [9], регуляризация [10], линейный дискриминантный анализ [11], метод частичных наименьших квадратов [12]. Применение методов регрессии и классификации разработанных для матричного представления данных, к многоиндексным данным требует либо модификации методов, либо уплощения данных.

В работе исследуется фильтрационный подход к выбору признаков. Предложена многоиндексная модификация фильтрационного метода QPFS выбора признаков [13]. Методы выбора признаков разделяются на фильтрационные, оберточные и встроенные [14]. В рамках оберточных и встроенных методов обучение модели является частью процесса отбора признаков. В фильтрационных методах обучение модели и выбор признаков независимы друг от друга. Благодаря этому фильтрационные методы вычислительно более эффективны и применимы для отбора признаков в пространствах более высокой размерности [15]. Другим преимуществом фильтрационных методов является то, что результат ранжирования признаков с помощью фильтрационного метода не привязан к конкретной модели и в этом смысле является универсальным. Преимуществом фильтрационных методов отбора признаков с методами снижения размерности, основанными на преобразованиях признаков (например, с методами главных компонент или частичных наименьших квадратов) является интерпретируемость получаемых результатов. Выбор наиболее значимых признаков предоставляет ценную экспертную информацию. Так же ценным свойством методов выбора признаков является разреженность, так как разреженность позволяет снизить стоимость измерений путем исключения нерелевантных признаков. На текущий момент методы выбора признаков, учитывающие многоиндексную структуру признаков относятся либо встроенным, либо к оберточным методам. В таблице 1 приведены ключевые работы по данной теме. Все приведенные методы зависят от выбранной модели прогнозирования и не применимы для других регрессионных моделей. Разреженность признаков в этих работах либо отсутствует (это касается преобразований признаков), либо вызывается с помощью специальных ограничений, накладываемых на функцию оптимизации. Таким образом, разработка фильтрационного метода отбора признаков, обладающих многоиндексной структурой, является актуальной задачей.

Метод QPFS [13], в отличие от пошаговых методов, отбирает подмножество признаков, решая задачу оптимизации по вектору индикаторных переменных. Каждый элемент вектора соответствует одному признаку и определяет его вхождение в набор выбранных признаков. Оптимизируемый функционал представляется в виде квадратичной формы. Квадратичное по индикаторной переменной слагаемое отвечает за попарное сходство между объектами, линейное – за индивидуальную релевантность каждого признака. Количественная мера попарного сходства признаков и их релевантности зависит от задачи. От способа определения мер сходств и релевантности зависит результат отбора признаков. Рассматривается коэффициент корреляции между парами признаков как мера сходства, и между признаком и целевой переменной — как мера релевантности. Такой выбор гарантирует положительную полуопределенность квадратичной формы в оптимизационной задаче. Так как поставленная задача целочисленного программирования не является выпуклой, при выборе признаков применяется релаксация индикаторных переменных. После релаксации целочисленных ограничений задача становится выпуклой и эффективно решается. Для пере-

Таблица 1. Многоиндексные методы выбора признаков.

Метод	Тип
[16]: выбор признаков основан на анализе весов многоиндексного метода опорных векторов	оберточный
[17]: анализе весов многоиндексного метода опорных векторов с регуляризацией, поощряющей разреженность признакового пространства	оберточный
[18]: регуляризированная логистическая регрессия	оберточный
[55]: ядерный метод наименьших квадратов	встроенный
[19]: канонический корреляционный анализ	встроенный / преобразование признаков
[20]: итерационный многоиндексный метод наименьших квадратов	встроенный / преобразование признаков
[21]: многоиндексный метод наименьших квадратов с дополнительной регуляризацией	встроенный / преобразование признаков

хода от релаксированных решения к бинарным значениям индикаторной переменной необходимо ввести дополнительный параметр, определяющий количество выбранных признаков. В случае, если выбранная мера попарного сходства признаков не гарантирует положительной полуопределенности квадратичной формы, задача не является выпуклой, и необходимо применение приближенных методов оптимизации даже для релаксированной задачи.

В работе также рассмотрен метод снижения размерности в задачах классификации, основанный на вложении многообразий. Предполагается что объекты содержатся вблизи многообразия много меньшей размерности, чем размерность исходного пространства. Назовем это предположение гипотезой многообразия [22]. Гипотеза многообразия используется при решении ряда практических задач, включая задачи анализа генома, анализа текста и распознавания изображений не противоречат этой гипотезе [23]. Формальное определение гипотезы многообразия, а так же некоторое способы ее проверки предлагаются в [24].

Для снижения размерности требуется построить гладкое отображение множества признаков в пространстве исходных данных в некоторое множество в пространстве меньшей размерности. В задачах снижения размерности применяются методы, основанные на анализе дисперсии: латентно-семантический анализ [25, 26], анализ главных компонент [27]. Такие методы не применимы в случае, когда исходные данные образуют нелинейные многообразия: геодезически близкие в исходном пространстве объекты метрически далеки друг от

друга, и линейные методы [25, 26, 27] нарушают локальную структуру близости между объектами при снижении размерности. Для выполнения вложений из нелинейных многообразий были разработаны алгоритмы, использующие изометрические отображения. Алгоритмы ISOMAP [28] и Laplacian Eigenmap [29] приближают геодезическое расстояние с помощью графа k ближайших соседей. Алгоритмы Local Linear Embedding [30] и Hessian-based Locally Linear Embedding [31] основаны на предположении, что многообразие аппроксимируется кусочно-линейной функцией. Для каждого объекта исходного пространства строится его линейная аппроксимация через соседние объекты, после чего по этим аппроксимациям строятся образы в результирующем пространстве. Метод [31] использует для описания объектов квадратичную форму, что гарантирует асимптотическую оптимальность метода даже в случае невыпуклых множеств. Алгоритм Local Tangent Space Alignment Algorithm [32] также использует кусочно-линейную аппроксимацию. Многообразие приближается гиперплоскостью в окрестности каждой точки, после чего полученные приближения сглаживаются между собой. Метод Semidefinite Embedding [33] в точности сохраняет расстояния между ближайшими объектами. Для этого метод максимизирует след матрицы Грама для образов при ограничениях, накладываемых отношением соседства объектов исходного пространства и их матрицей Грама. Все перечисленные методы нацелены на наиболее точное сохранение расстояний между объектами при снижении размерности. Это может привести к неустойчивости решения, связанной с тем что, изменения расстояния между далекими и близкими объектами штрафуются одинаково. Кроме того, они не приспособлены для решения задачи классификации, поскольку не учитывают разметку при выполнении вложения, хотя существуют их модификации, обладающие этим свойством, например, метод TRIMAP [34], основанный на методе ISOMAP.

В данной работе предложена модификация метода t-SNE (t-distributed Stochastic Neighbor Embedding) [23], учитывающая целевую переменную. Метод t-SNE локализует изолированные плотные пространственные структуры в исходном пространстве. Под изолированной плотной структурой подразумевается множество точек, имеющих близких соседей из той же структуры, но сравнительно удаленных от всех остальных точек. Такой эффект достигается тем, что сохранению расстояния между близкими объектами и между далекими объектами назначаются различные приоритеты. В работе [35] описан альтернативный вариант параметрической модификации t-SNE. Дополнительным ограничением применимости метода t-SNE является высокая, по сравнению с другими методами вложений, вычислительная сложность. Хотя в [36] предлагается два способа вычисления градиента, при использовании которых сложность непараметрического t-SNE составляет $O(k \cdot m \log(m))$, где m — размер выборки, а k — размерность результирующего пространства, этого ускорения не достаточно для обеспечения работы даже с выборками объемом порядка 10^3. Предложенный метод расширяет границы применимости оригинального метода t-SNE. В

отличие от исходного метода t-SNE, разработанная модификация предусматривает вложение новых данных в результирующее пространство без повторного вложения обучающих данных, учитывает разметку обучающих данных и имеет параллельную реализацию.

В работе также рассмотрена проблема оптимального построения выборки. Исходные объекты — сигналы электрокортикограмм — содержат заведомо избыточную информацию: измерения напряженности электрического поля, снимаемые с соседних электродов, сильно коррелируют. При дальнейшем формировании признакового пространства количество коррелирующих признаков может увеличиваться. Так же при высокой чистоте сэмплирования значения сигнала с соседние моменты времени сильно скоррелированы между собой и могут быть частично опущены из рассмотрения. Таким образом, построение выборки по сигналам электрокортикограмм включает этапы выбора объектов, то есть отсчетов времени, и построения оптимального признакового пространства. Требуется построить выборку оптимальным образом с точки зрения баланса между размерностью выборки и ее информативностью.

Таблица 2. Способы оценки объема выборки.

Метод	Формула	Ссылка
Тест на равенство средних: $Z = \frac{\hat{p}-p_0}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0:\ p = p_0$ против $H_1:\ p \neq p_0$	$m = \frac{(z_{\mathrm{Pow}}+z_{\alpha/2})^2 p(1-p)}{(\hat{p}-p_0)^2}$	[40]
Тест на сравнение средних: $Z = \frac{\hat{p}-p_0-\delta}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0:\ p - p_0 \leq \delta$ против $H_1:\ p - p_0 > \delta$	$m = \frac{(z_{\mathrm{Pow}}+z_{\alpha/2})^2 p(1-p)}{(\hat{p}-p_0-\delta)^2}$	[40]
Тест на эквивалентность: $Z^- = \frac{\hat{p}-p_0-\delta}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{m}$, $Z^+ = \frac{\hat{p}-p_0+\delta}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{m}$ при $H_0:\ \lvert p-p_0\rvert \leq \delta$ против $H_1:\ \lvert p-p_0\rvert > \delta$	$m = \frac{(z_{\mathrm{Pow}}+z_{\alpha/2})^2 p(1-p)}{(\lvert\hat{p}-p_0\rvert-\delta)^2}$	[40]
Тест множителей Лагранжа: $\gamma_m:\ \chi^2_{p,1-\mathrm{Pow}}(\gamma_m) = \chi^2_{p,\alpha}$		[41]
Тест отношения правдоподобия: $\gamma_m:\ \chi^2_{p,1-\mathrm{Pow}}(\gamma_m) = \chi^2_{p,\alpha}$	$m = \frac{\gamma_m}{\Delta^*}$, $\Delta^* = \mathrm{E}_X\left[\frac{-X(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}{1+e^{-X\boldsymbol{\beta}}} - \log\left(\frac{1+e^{-X\boldsymbol{\beta}}}{1+e^{-X\boldsymbol{\beta}^*}}\right)\right]$	[42, 43]
Статистика Вальда: $Z = \frac{\hat{\beta}-\beta^0}{\sqrt{\hat{V}_0}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0:\ \beta = \beta^0$	$\hat{m} = \frac{\left(\sqrt{V_1}z_{\mathrm{Pow}}+\sqrt{V_0}z_{\alpha/2}\right)^2}{(\beta^1-\beta^0)^2}$	[45]
$Z* = \frac{\hat{\beta}-\beta^0}{\sqrt{V_1}}\sqrt{m} \to \mathcal{N}(0,1)$ при $H_0:\ \beta = \beta^0$	$\hat{m} = \frac{V_1\left(z_{\mathrm{Pow}}+z_{\alpha/2}\right)^2}{(\beta^1-\beta^0)^2}$	[44]
$\hat{\beta}_j = t_{1-\alpha/2}(m-n-1)\sqrt{\frac{1-\mathbf{R}^2}{(1-\mathbf{R}_j^2)(m-n-1)}}$ $\mathbf{R} = \boldsymbol{\rho}'_{yx}\mathbf{R}_{xx}^{-1}\boldsymbol{\rho}_{yx}$	$m = \frac{z^2_{\alpha/2}}{\delta^2}\left(\frac{1-\mathbf{R}^2}{1-\mathbf{R}_j^2}\right) + n + 1$ $m^* = \frac{z^2_{\alpha/2}}{\delta^2}\left(\frac{1-\mathbf{R}^2}{1-\mathbf{R}_j^2}\right)\left(\frac{\chi^2_{1-\gamma}(m-1)}{m-n-1}\right) + n + 1$	[48]
Бутстреп: проверка завышенности оценки m для получения заданного уровня $1-\alpha$.	$m = \left(\frac{z_{\alpha/2}\sigma}{\bar{X}-\mu}\right)^2$ и $m = \frac{z^2_{\alpha/2}}{(\bar{X}-\mu)^2}\left(\frac{1-R^2}{1-R_j^2}\right) + n$	[49]

Задача оценки объема выборки связана с планированием эксперимента. При планировании эксперимента требуется оценить минимальный объем выборки — количество производимых измерений некоторого параметра или набора параметров, необходимый для выполнения некоторых ранее сформулированных

условий. В работах [40, 41, 42, 43, 44, 45], эти условия формулируются в виде статистического критерия, и связанный с данным критерием метод оценки объема выборки гарантирует достижение фиксированной статистической мощности с величиной ошибки первого рода, не превышающей заданного значения. Выбор метода оценки объема выборки зависит от решаемой задачи, определяющей формулировку статистической гипотезы и статистику для ее проверки [40]. В таблице 2 приведен ряд методов оценки объема выборки. Проверяются гипотезы следующего вида: проверка равенства измеряемого значения ожидаемому значения, сравнение его с пороговым значением; сравнение частоты наблюдаемого явления с ожидаемым значением; сравнение параметров, измеряемых по двум выборкам. Необходимо отметить, что при исследовании одного и того же параметра формулы для оценки объема выборки при различных проверяемых гипотезах могут отличаться даже при использовании одной и той же статистики.

В работе [41] предложены процедуры оценки мощности критерия множителей Лагранжа для коэффициентов обобщенно-линейной регрессии необходимой оценки объема выборки. Недостатком предложенного метода является то, что в случае, когда альтернативная гипотеза сильно отличается от нулевой, оценки максимального правдоподобия для параметров модели и ковариационной матрицы, используемые при вычислении мощности, не являются асимптотически состоятельными при альтернативной гипотезе. Тем не менее, для случая логистической модели и конечного набора реализаций независимых переменных предложенная аппроксимация позволяет оценить мощность теста с высокой точностью. Позднее авторами [41] был предложен [42] подход к оценке мощности и связанного с ней объема выборки на основе теста отношения максимального правдоподобия. Этот подход оказался более точным для ряда распределений независимых переменных. В работе [43] подход [42], в котором непрерывные величины дискретизируются путем разбиения множества значений на интервалы, распространен на случай непрерывных величин без категоризации. Категоризация случайных величин ведет к потере информации о распределениях рассматриваемых величин и ухудшению точности оценок мощности и объема выборки. Кроме того, не ясно, как именно разбивать непрерывные величины на интервалы: в работе [42] рекомендаций по способу разбиения не дается; предлагается сравнивать оценки, полученные при различных способах. В [42] предлагается способ оценки мощности теста отношения правдоподобия, не требующий категоризации переменных. В [42] показано, что предложенный метод аппроксимации мощности дает более точные оценки мощности теста отношения правдоподобия для ряда непрерывных распределений в случае, когда эти распределения известны. Метод, предложенный в [42] более устойчив по отношению к неточности в предположениях о распределении независимых переменных. Методы, основанные на максимизации правдоподобия асимптотически эквивалентны с точки зрения ошибки первого рода. Более того, они локально эквивалентны, то есть их функции мощности близки, когда альтернативная гипотеза близка к ну-

левой [46, 47]. Однако глобально эти тесты различны и дают различные оценки объема выборки. В работе [44] в случае логистической регрессии предлагается использовать оценку объема выборки, основанную на статистике Вальда. Авторы модифицируют полученную оценку, в соответствии с тем видом статистики Вальда [45], который реально используется на практике и демонстрируют преимущество модифицированной оценки по сравнению в классической в случае одной бинарной независимой переменной.

В работе [48] предлагает при оценке объема выборки в задачах регрессии опираться не только на мощность тестов, но и учитывать точность оцениваемых параметров — ширину доверительного интервала. При таком подходе оценка объема выборки сводится к нахождению компромисса между двумя максимизацией мощности и точностью оцениваемых параметров.

Статистические методы 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 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.

Статистические методы позволяют оценить объем выборки, исходя из предположений о распределении данных и информации о соответствии наблюдаемых величин предположениям нулевой гипотезы. В случае, если объем исследуемой выборки достаточен или избыточен, возможно применение методов, основанных на наблюдении за изменением некоторой характеристики процедуры построения модели при увеличении объема выборки. В частности, наблюдая за отношением качества прогнозирования на контрольной выборке и обучающей выборке, определим достаточный объем выборки как соответствующий началу переобучения. В работе [49] для оценки достаточного объема выборки используется метод бутстреп. Проводится проверка избыточности текущего объема выборки на основе анализа доверительных интервалов оцениваемого параметра. Ширина доверительного интервала при различных значениях объема выборки оценивается с помощью метода бутстреп. Для этого заданное количество раз ресемплируются выборки меньшего объема и по полученным реализациям выборок вычисляется доверительный интервал для ошибки при оценке некоторого параметра θ. Объем выборки считается достаточным, если ширина доверительного интервала не превышает некоторого значения.

Аналогичным образом производится оценка объема выборки в рамках предлагаемого метода: предлагается считать объем выборки достаточным, если расстояние между распределениями, оцененными на подвыборках данного объема, достаточно мало. Такой подход не требует дополнительного обобщения на случай многих переменных. Кроме того, оценку можно производить как при наличии предположений о распределении данных, так и в их отсутствие. Недо-

статком данного подхода является то что количественные оценки возможно получить лишь в случае, когда объем выборки избыточен. В противном случае метод позволяет лишь определить, является ли текущий объем выборки достаточным.

Предлагаемый метод относится к классу байесовских методов оценки объема выборки. Перечисленные выше ограничения статистических методов оценки объема выборки подробно рассматриваются в байесовской постановке [50, 51, 52], в которой оценка объема выборки определяется исходя из максимизации ожидаемой функции полезности [50]. Функция полезности может включать в явном виде функции распределения параметров и штрафы за увеличение объема выборки. Альтернатива подходам, основанным на функции полезности — оценка объема выборки путем ограничения некоторого критерия качества оценки параметров модели. Примеры таких критериев — критерий средней апостериорной дисперсии APVC, критерий среднего покрытия ACC, критерий средней длины ALC, критерий эффективного объема выборки ESC — перечислены в [53, 54]. Для каждого из перечисленных критериев оценка объема выборки определяется как минимальное значение объема выборки, для которого ожидаемое значение выбранного критерия не превосходит некоторого фиксированного порога. Одним из направлений исследований в байесовской оценке объема выборки является разработка новых функций полезности и изучение их свойств. Для некоторых функций полезности и априорных предположений о распределении параметров, формулу для оценки объема выборки можно получить в явном виде. В остальных случаях решение находится методами имитационного моделирования [52].

В данной работе подхода к оценке объема выборки связан с выбором модели. Предполагается, что на этапе сбора данных зафиксировано несколько возможных моделей. Исследования в данном направлении сосредоточены на разработке методов оценки объема выборки, при котором использование выбранного методы выбора модели приведет к выбору оптимальной модели с достаточно высокой вероятностью. В такой постановке исследователь отвечает на вопрос о том, при каком объеме выборки возможно статистическое различение моделей. Например, если исследуемые модели имеют одинаковые функции правдоподобия, а отличаются только априорными функциями распределения, при увеличении объема выборки апостериорные распределения станут неразличимы. С этой точки зрения методы, основанные на отношении правдоподобий, неприменимы [52], так как при увеличении объема выборки невозможно гарантировать, что более простая модель будет выбрана в случае ее истинности со сколь угодно большой вероятностью.

Цели работы. Целью работы является разработка методов выбора моделей прогнозирования и классификации временных рядов, обладающий корреляцией сложной пространственно-временной структуры.

1. Предложить метод выбора модели прогнозирования мультикоррелирую-

щих временных рядов для решения задачи декодирования временного ряда.

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 и провести вычислительный эксперимент для сравнения различных методов выбора признаков при решении задачи декодирования электрокортикограмм.

Методы исследования.

Для выбора оптимальной модели используются методы выбора признаков и снижения размерности [55, 56]. В работе исследуются методы анализа многоиндексных признаковых описаний. Многоиндексное представление активно используется при анализе биомедицинских и химических данных в связи с наличием многоиндексной структуры данных в этих областях. Преобразование многоиндексных данных в плоские матрицы может привести потере зависимостей, присутствующих в многоиндексных данных. Напротив, многоиндексные подходы сохраняют структуру данных, и в ряде случаев позволяют повысить качество регрессии. Использование многоиндексных представлений позволяет повысить качество регрессии в ряде приложений [55]. Метод PLS и его многоиндексные модификации [21, 55] используются для решения задачи декодирования электрокортикограмм [55, 20, 38]. По аналогии с классическим методом PLS, опирающимся на сингулярное разложение, многоиндексный PLS использует многоиндексные разложения, такие как разложение Таккера или PARAFAC [57]. Для повышения устойчивости результатов PLS-регрессии в условиях большого числа коррелирующих признаков, используются регуляризаторы и их многоиндексные модификации [21].

Исследуется альтернативный метод выбора признаков для многоиндексных данных. Используется метод QPFS [13], позволяющий учитывать отношения между признаками, не перебирая всех комбинаций признаков. Для решения задачи выбора признаков в методе QPFS решается задача квадратичного программирования, одновременно минимизирующая сходство между признаками, задаваемое квадратичной формой, и максимизирующая релевантность признаков, задаваемую линейным членом.

Для учета многоиндексной структуры данных предложена многоиндексная формулировка метода QPFS. Для адаптации данного метода к выбору призна-

ков задаче декодирования электрокортикограмм вводятся функции сходства между признаками отдельно для каждой из многоиндексных размерностей. Такой подход снижает размерность оптимизационной задачи по сравнению с использованием плоских матриц, что позволяет рассматривать исходное признаковое описание более высокой размерности.

Задача отбора объектов и признаков решается для фиксированного семейства моделей.

Для оценки объема выборки вводится критерий устойчивости оценок параметров модели. Устойчивость оценивается с помощью расстояния Кульбака-Лейблера [129, 58] между апостериорными распределениями параметров модели, оцененными по подвыборкам различного объема. Предлагаемый в данной диссертационной работе метод оценки объема выборки используется для выбора адекватной модели. Выбор модели осуществляется на основе сравнения оценки достаточного объема выборки с его фактическим значением. В непараметрическом случае для оценки близости эмпирических распределений двух выборок используется расстояние Кульбака-Лейблера между гистограммами этих выборок. При увеличении размерности выборки малого объема повышается разреженность гистограмм, поэтому предлагается использовать устойчивые модификации расстояния Кульбака-Лейблера [59]. Вводится статистический критерий принадлежности двух подвыборок одному распределению на основе расстояния Кульбака-Лейблера между распределениями параметров (или гистограммами распределений). Обоснование предложенного метода опирается на известные [60, 61] свойства распределения Кульбака-Лейблера, а также полученные [62] на их основе результаты.

Основные положения, выносимые на защиту.

1. Предложен метод выбора признаков, учитывающий многоиндексное мультикоррелирующее представление данных. Метод позволяет учитывать связи между признаками независимо от модели прогнозирования, не перебирая всех комбинаций признаков.
2. Исследованы свойства предложенного метода. Сформулированы условия локальной сходимости предложенного алгоритма. Доказана приближенная оптимальность низкорангового релаксированного решения.
3. Разработан метод выделения сегментов временных рядов для интерпретируемого сегментирования квазипериодических сигналов.
4. Разработан метод снижения признакового пространства, сохраняющий локальную структуру близости между объектами. Решена проблема учета разметки при выполнении вложения для классификации.
5. Приведены рекомендации по практическому использованию разработанных алгоритмов. Рассмотрены практические задачи декодирования временных рядов электрокортикограмм, выделения внутреннего плагиата, задачи агрегирования и прогнозирования временных рядов грузовых железнодорожных перевозок.

Научная новизна.

1. Предложен способ выбора признаков, основанный решении задачи квадратичного программирования. Метод учитывает многоиндексную структуру данных.
2. Предложен способ снижения размерности признакового пространства в задаче классификации, учитывающий локальную структуру близости между объектами и целевую переменную.
3. Предложен способ сегментирования квазипериодических временных рядов с помощью анализа фазовой траектории главных компонент временного ряда.
4. Предложен способ оценки объема достаточного выборки, основанный на сравнении оценок функции распределения данных, полученных на различных подвыборках исследуемой выборки. На основе получаемых оценок достаточного объема выборки предлагается выбирать оптимальный подход к построению модели классификации. Данный подход является оригинальным и предлагается впервые.

Теоретическая значимость. Работа в целом носит теоретический характер. Предложен метод выбора признаков с помощью квадратичного программирования, учитывающий многоиндексную структуру данных. Предложен приближенный алгоритм решения, основанный на низкоранговом разложении многоиндексной матрицы 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. Исследованы теоретические свойства предложенных методов. Предложен критерий информативности добавляемого в выборку объекта, исследованы его статистические свойства.

Практическая значимость. Предложенные в работе методы предназначены для построения точных и вычислительно простых моделей классификации, прогнозирования и декодирования временных рядов. Методы предназначены для создания нейрокомпьютерных интерфейсов и в частности, нейропротезирования. В работе приведены примеры использования предложенных методов для решения задач определения заимствований в авторских текстах, прогнозировании железнодорожных перевозок и автоматической сегментации биофизических временных рядов.

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 [63].
2. Оценка объема выборки в задачах классификации // Математические методы распознавания образов, г. Светлогорск, Калининградская область, 2015 [64].
3. Bayesian Sample Size Estimation for Patient Classification Survey // 20th Conference of the International Federation of Operational Research Societies, Barcelona, 2014 [65].
4. Small CVD sample set classification: generative versus discriminative // 26th European conference on operation research, Rome, 2013 [66].
5. Multiclass classification of cardio-vascular disease patients with sample size estimation // 25th European conference on operation research, Vilnuis, 2012 [67].

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

1. A. Motrenko, V. Strijov. Multi-way Feature Selection for ECoG-based Brain-Computer Interface // 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.

Личный вклад. Результаты получены самостоятельно под научным руководством д.ф.-м.н. В. В. Стрижова. Личный вклад автора в работах с соавторами заключается в следующем: в работе [68] проделана экспериментальная часть работы, в работах [69, 62, 70, 71] сформулированы и получены теоретические результаты.

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.

Глава 1
Снижение размерности в задачах прогнозирования и классификации мультикоррелирующих временных рядов

1.1. Задача декодирования сигналов высокой размерности

Работа посвящена разработке методов анализа наборов мультикоррелирующих временных рядов: классификации, прогнозирования и *декодирования* временного ряда.

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 восстановлении 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 признакового описания при решении задачи. Отличие от задачи регрессии в том, что признаковое описание не определено, и его построение является частью решения задачи декодирования.

В работе рассмотрены ключевые задачи создания и выбора моделей прогнозирования и классификации временных рядов.

- Построение признакового пространства включает обработку исходного сигнала, разбиение на сегменты и извлечение признаков. Извлечение признаков может носить экспертный характер, либо совершаться автоматически [72]. Методы извлечения признаков из временных рядов рассмотрены в главах 3 и 5.
- Выбор признаков. Так как в решаемых задачах признаковое пространство избыточно, требуется снизить его размерность для уменьшения сложности регрессионной модели и повышения ее устойчивости. Методы выбора признаков и снижения размерности рассмотрены в данной главе и в главе 2.
- Выбор объектов. Так как частота сэмплирования исходного сигнала в решаемых задачах превышает частоту сэмплирования прогнозируемого временного ряда, требуется выбрать способ, ставящий в соответствие прогнозируемым отсчетам времени отсчеты или интервалы времени исходного временного ряда. Методы построения выборки оптимального состава рассмотрены в главе 4.

1.2. Задача регрессии для многоиндексных данных

Пусть из временных рядов $\mathbf{s}(t)$, $\mathbf{y}(t)$ составлена выборка $(\underline{\mathbf{D}}, \mathbf{Y})$

$$\underline{\mathbf{D}} \in \mathbb{R}^{M \times n_1 \times \cdots \times n_D},\ D_{(m,:,:,:)} = \underline{\mathbf{X}}_m, \quad \mathbf{Y} = [\mathbf{y}_1^\intercal, \ldots, \mathbf{y}_M^\intercal]^\intercal, \tag{1.1}$$

где $\mathbf{y}_m = \mathbf{y}(t_m)$, а $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ – многоиндексная матрица. Требуется восстановить значения $\mathbf{y}_m$ при заданной матрице $\underline{\mathbf{X}}_m$, $m = 1, \ldots M$. Приблизим целевой временной ряд с помощью линейной модели

$$\hat{\mathbf{y}}_m = \mathrm{vec}\big(\underline{\mathbf{X}}_m\big)^\intercal \hat{\mathbf{W}}, \tag{1.2}$$

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

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \frac{1}{M \cdot N_y} ||\hat{\mathbf{Y}} - \mathbf{Y}||_F^2. \tag{1.3}$$

В задаче регрессии (1.2) предполагается, что признаки содержат в себе информацию, достаточную для восстановления целевых переменных, то есть существует неизвестная зависимость между целевой переменной $\mathbf{y}$ и признаками $\boldsymbol{\chi}_j$.

Определение 3. Признак $\boldsymbol{\chi}_i$ называется *релевантным* в задаче (1.2), если для некоторого достаточно малого $\delta > 0$ и вектора целевых переменных $\boldsymbol{y}_j$ выполнено

$$\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{y}_j) \geq 1 - \delta.$$

Признаки, нерелевантные поставленной задаче, являются шумовыми и должны быть отфильтрованы. Введем также понятие мультикоррелирующих признаков.

Определение 4. Признаки $\boldsymbol{\chi}_i$, $\boldsymbol{\chi}_j$ называются *коррелирующими*, если для некоторого достаточно малого $\delta > 0$ выполнено

$$\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \geq 1 - \delta.$$

Определение 5. Множество признаков $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_K$ называется *мультикоррелирующими*, если любая пара признаков $\boldsymbol{\chi}_i$, $\boldsymbol{\chi}_j$, $i, j \in \{1, \ldots, K\}$ из этого множества коррелируют.

Определение 6. Признаковое описание в задаче (1.2) называется *избыточным*, если существует такое преобразование признаков $\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_N] \mapsto \mathbf{Z} = [\boldsymbol{\chi}'_1, \ldots, \boldsymbol{\chi}'_K]$, такое что $N < K$, и

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

где f и g – оптимальные модели (1.2), построенные для признаковых описаний $\mathbf{X}$ и $\mathbf{Z}$.

Мультикоррелирующие признаки содержат избыточную информацию. Поэтому решение задачи (1.2), (1.3) неустойчиво. Чтобы уменьшить вычислительные затраты и повысить устойчивость решения воспользуемся методами отбора признаков (1.3).

1.3. Задача выбора признаков для многоиндескных данных

В данном разделе сформулирована задача выбора признаков для многоиндексного признакового описания $\underline{\mathbf{X}} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$.

Определение 7. Пусть функция

$$\varphi : \{1, \ldots, n_1\} \times \cdots \times \{1, \ldots, n_D\} \mapsto \{1, \ldots, (n1 \cdot \cdots \cdot n_D)\}$$

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

$$\mathbf{x} = \mathrm{vec}(\underline{\mathbf{X}}), \quad \text{где } x_{\varphi(i_1,\ldots,i_D)} = x_{i_1 \ldots i_D}, \ \mathbf{x} \in \mathbb{R}^{n_1 \cdot \cdots \cdot n_D}, \tag{1.4}$$

называется *векторизацией*.

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

$$\mathbf{X} = [\mathbf{x}_1^{\mathsf{T}}, \ldots, \mathbf{x}_M^{\mathsf{T}}]^T \in \mathbb{R}^{M \times (n_1 \cdot \cdots \cdot n_d)}, \text{ где } \mathbf{x}_m = \mathrm{vec}(\underline{\mathbf{X}}_m), \tag{1.5}$$

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

Обозначим через $\mathbf{X} \in \mathbb{R}^{M \times (n_1 \cdot \cdots \cdot n_D)}$ уплощение матрицы $\underline{\mathbf{D}} \in \mathbb{R}^{M \times (n_1 \cdot \cdots \cdot n_D)}$. Определим индикаторную переменную $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$, кодирующую вхождения признаков $\boldsymbol{\chi}_{i_1 \ldots i_D}$ в модель (1.2). Соответствующую двухиндексную матрицу обозначим $\mathbf{X}_{\underline{\mathbf{A}}}$

$$\mathbf{X}_{\underline{\mathbf{A}}} = [\ldots, \boldsymbol{\chi}_{i_1 \ldots i_D}, \ldots], \text{ где } a_{i_1 \ldots i_D} = 1.$$

Матрица $\mathbf{X}_{\underline{\mathbf{A}}}$ составлена из столбцов-признаков $\boldsymbol{\chi}_{i_1 \ldots i_D}$, для которых индикаторная переменная $a_{i_1 \ldots i_D}$ принимает ненулевое значение.

В принятых обозначениях задача выбора признаков формулируется следующим образом:

$$\underline{\mathbf{A}} = \underset{\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times \cdots \times n_D}}{\arg\min} \ \mathcal{L}\left(\mathbf{X}_{\underline{\mathbf{A}}} \mathbf{W}_{\underline{\mathbf{A}}}, \mathbf{Y}\right),$$

где $\mathcal{L}(\hat{\mathbf{Y}}, \mathbf{Y})$ — функция потерь, а матрица весов $\mathbf{W}_{\underline{\mathbf{A}}} \in \mathbb{R}^{N \times N_y}$, $N = \sum a_{i_1 \ldots i_D}$ минимизирует квадратичные потери (1.3) для $\mathbf{X}_{\underline{\mathbf{A}}}$.

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

$$\mathrm{sMSE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} ||\hat{\mathbf{y}}_m - \mathbf{y}_m||_2}{\sum_{m=1}^{M} ||\bar{\mathbf{y}} - \mathbf{y}_m||_2}, \quad \bar{\mathbf{y}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{y}_m, \tag{1.6}$$

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

$$\operatorname{corr}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{N_y} \sum_{n=1}^{N_y} \frac{\operatorname{cov}(\hat{\boldsymbol{y}}_n, \boldsymbol{y}_n)}{\sqrt{\operatorname{cov}(\hat{\boldsymbol{y}}_n, \hat{\boldsymbol{y}}_n)\operatorname{cov}(\boldsymbol{y}_n, \boldsymbol{y}_n)}}, \tag{1.7}$$

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

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

1.4. Выбор признаков в условиях мультикоррелирующего признакового описания

Для уменьшения избыточности признакового описания применяются методы выбора признаков. В данной работе развивается фильтрационный метод QPFS [13]. Фильтрационные методы ранжируют признаки вне зависимости от модели, поэтому они более эффективны чем методы, требующие настройки модели. С другой стороны, так как признаки оцениваются независимо, фильтрационные методы склонны к выбору коррелирующих признаков. Метод QPFS позволяет сбалансировать сходство между признаками и релевантность выбранных признаков задаче регрессии, не перебирая всех возможных комбинаций признаков. Зададим *сходство* между i-м и j-м признаками как

$$q_{ij} = |\operatorname{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)| \approx \frac{1}{M} \left[\mathbf{X}^{\intercal}\mathbf{X}\right]_{ij}, \tag{1.8}$$

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

$$b_i = \frac{1}{N_y} \sum_{n=1}^{N_y} |\operatorname{corr}(\boldsymbol{\chi}_i, \boldsymbol{y}_n)|. \tag{1.9}$$

Задача выбора признаков формулируется в виде задачи квадратичного программирования

$$\mathbf{a} = \underset{\mathbf{a}\in\{0,1\}^N}{\arg\min} \left(\mathbf{a}^{\intercal}\mathbf{Q}\mathbf{a} - \mathbf{b}^{\intercal}\mathbf{a}\right), \tag{1.10}$$

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 $\mathbf{b}$. В работе [13] кроме коэффициента корреляции также рассмотрены взаимная информация и нормализованная значимость признаков в качестве мер сходства и

релевантности. 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 (1.10) уравновешивает сходство между признаками из выбранного набора и их предсказательную способность, варьируя значения индикаторной переменной $\mathbf{a} \in [0,1]^N$, определяющей набор выбранных признаков:

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

1.4.1. Анализ мер сходства и релевантности при выборе модели прогнозирования временных рядов

Рассмотрим случаи организации признакового пространства:

- Все признаки независимы и каждый вносит одинаковый вклад в целевую переменную. Пусть, например:

  $$y = x_1 + x_2, \quad x_1, x_2 \ \sim \mathcal{N}(0, \sigma^2).$$

  Тогда, в случае (1.8) и (1.9) получаем $\mathbf{Q} = \mathbf{I}_2$, $\mathbf{b} = [1/\sqrt{2}, 1/\sqrt{2}]^{\mathsf{T}}$. Функция $F(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\mathbf{Q}\mathbf{x} - \mathbf{b}^{\mathsf{T}}\mathbf{x} = (x_1 + x_2)(1 - 1/\sqrt{2})$ имеет минимум в точке $(0,0)$, хотя все признаки должны быть выбраны.
- Целевая переменная y и признаки x_1, x_2 равны с точностью до случайного шума. Тогда $q_{ij} \equiv 1$, $b_i \equiv 1$ и $F(\mathbf{x}) = \sum_{i \neq j} x_i x_j$. Для функции двух переменных $F(x_1, x_2)$ минимум достигается сразу в трех точках $(0,0)$, $(0,1)$ и $(1,0)$. При этом решение $(0,0)$ не является оптимальным с точки зрения решаемой задачи прогнозирования.

Подобные решения возникают при использовании одной и той же функции для определения сходства $\mathbf{Q}$ и релевантности $\mathbf{b}$. Так как сходство признака с самим собой всегда принимает максимальное значение, а b_i, $q_{ij} \geq 0$, одним из решений задачи минимизации (1.10) будет решение $\mathbf{0}$. В связи с этим предлагается при определении релевантности добавлять к вектору $\mathbf{b}$ величину $1 - c \in [0,1]$, где c — некоторая константа, интерпретируемая как минимальное допустимое значение релевантности признака. На основании приведенного примера рекомендуется выбирать $c = 1/\sqrt{N}$, где N — максимальное допустимое количество активных признаков.

1.5. Многоиндексный метод выбора признаков на основе низкорангового тензорного разложения

Задача (1.10) сформулирована для двухиндексных матриц. При решении практических задач в данной работе рассматриваются многоиндексные признаковые описания $\underline{\mathbf{X}}_m \in \mathbb{R}^{n_1 \times \cdots \times n_D}$.

Базовым способом адаптации многоиндексных признаковых описаний к задаче (1.10) является уплощение (1.5) матрицы признаков $\underline{\mathbf{D}}$ векторизацией (1.4) каждой D-индексной матрицы $\underline{\mathbf{X}}_m$, и последующее применение исходного метода QPFS (1.10). Ограниченная применимость такого подхода связана с вычислением матрицы сходства $\mathbf{Q}$. Так как матрица $\mathbf{X}$ содержит значительное число коррелирующих признаков, с ростом количества признаков $\mathbf{Q}$ становится близка к вырожденной, даже будучи положительно определенной по построению (1.8). Формулировка оптимизационной задачи, т.е вычисление матриц $\mathbf{Q}$ и $\mathbf{b}$, становится самой нагруженной частью метода QPFS. Для решения этой проблемы предлагается использовать многоиндексную структуру признаков $\underline{\mathbf{X}}_m$ в тензорной модификации метода QPFS. В частности, для каждого направления $d = 1, \ldots, D$ вводится отдельная матрица сходства: $\mathbf{Q}_d \in \mathbb{R}^{n_d \times n_d}$. Многоиндексная матрица релевантности $\underline{\mathbf{B}}$ имеет ту же размерность, что и $\underline{\mathbf{X}}_m$.

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 многоиндекной матрицы $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times \cdots \times n_D}$ и матрицы $\mathbf{B} \in \mathbb{R}^{m \times n_d}$ вдоль направления d

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

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

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

Пусть заданы матрицы сходства $\mathbf{Q}_d \in \mathbb{R}^{n_d \times n_d}$ для каждого направления $d = 1, \ldots, D$ многоиндексной матрицы $\underline{\mathbf{X}}$ и задана матрица релевантности $\underline{\mathbf{B}}$. Переформулируем задачу (1.10) в виде

$$\underline{\mathbf{A}} = \underset{\underline{\mathbf{A}} \in \{0,1\}^{n_1 \times \cdots \times n_D}}{\arg\min} \left(\sum_{d=1}^{D} (\underline{\mathbf{A}} \times_d \mathbf{Q}_d) * \underline{\mathbf{A}} - \underline{\mathbf{B}} * \underline{\mathbf{A}} \right) \times_1 \mathbf{1}_{n_1} \ldots \times_D \mathbf{1}_{n_D}, \qquad (1.11)$$

Операция $\underline{\mathbf{A}} \times_1 \ldots \times_D \mathbf{1}_{n_D}$ эквивалентна суммированию всех элементов $\underline{\mathbf{A}}$:

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

Предложено приближенное решение (1.11), основанное на низкоранговом приближении $\underline{\mathbf{A}}$, которое позволяет решать задачу (1.11) итеративно, таким образом что на каждом шаге решается задача квадратичного программирования.

Чтобы сформулировать приближенное решение задачи, используется бинарность переменной $\underline{\mathbf{A}}$. Для бинарной многоиндексной матрицы существует точное низкоранговое разложение с некоторым R

$$\underline{\mathbf{A}} = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \circ \cdots \circ \mathbf{a}_D^{(r)}, \quad \mathbf{a}_d^{(r)} \in \mathbb{R}^{n_d}, \quad d = 1, \ldots, D. \tag{1.12}$$

Перепишем функцию потерь (1.11) с учетом (1.12) в виде

$$F(\underline{\mathbf{A}}) = \sum_{r=1}^{R} \sum_{d=1}^{D} \left(\mathbf{a}_d^{(r)\intercal} \mathbf{Q}_d \mathbf{a}_d^{(r)} \prod_{j \neq d} \|\mathbf{a}_j^{(r)}\|_2^2 \right) - \underline{\mathbf{B}} \times_1 \mathbf{a}_1^{(r)} \ldots \times_D \mathbf{a}_D^{(r)}. \tag{1.13}$$

Эта задача решается итеративно. На каждом шаге $k = 1, \ldots, K$ решается задача квадратичного программирования. Пусть вектор

$$\boldsymbol{\alpha}_d = [\mathbf{a}_d^{(1)\intercal}, \ldots, \mathbf{a}_d^{(R)\intercal}]^\intercal \in \mathbb{R}^{nR} \tag{1.14}$$

объединяет значения компонент разложения (1.13) индикаторной переменной $\underline{\mathbf{A}}$ для d-го направления, $d = 1, \ldots, D$. Пусть вектор $\boldsymbol{\alpha}^{(0)} = \mathbf{1}_{n_i R}$ выбран в качестве начальной аппроксимации вектора $\boldsymbol{\alpha}_i$. При фиксированных $\boldsymbol{\alpha}_1^{(k)}, \ldots, \boldsymbol{\alpha}_{d-1}^{(k)}$ и $\boldsymbol{\alpha}_{d+1}^{(k-1)}, \ldots, \boldsymbol{\alpha}_D^{(k-1)}$ относительно $\boldsymbol{\alpha}_d$ решается следующая задача

$$\boldsymbol{\alpha}_d^{(k)} = F(\boldsymbol{\alpha}_1^{(k)}, \ldots, \boldsymbol{\alpha}_{d-1}^{(k)}, \boldsymbol{\alpha}_d, \boldsymbol{\alpha}_{d+1}^{(k-1)}, \ldots, \boldsymbol{\alpha}_D^{(k-1)}), \quad \text{где} \tag{1.15}$$

$$F(\boldsymbol{\alpha}_1^{(k)}, \ldots, \boldsymbol{\alpha}_{d-1}^{(k)}, \boldsymbol{\alpha}_d, \boldsymbol{\alpha}_{d+1}^{(k-1)}, \ldots, \boldsymbol{\alpha}_D^{(k-1)}) = \boldsymbol{\alpha}_d^\intercal \left(\tilde{\mathbf{Q}}_d^{(k-1)} + \tilde{\mathbf{I}}_d^{(k-1)} \right) \boldsymbol{\alpha}_d - \tilde{\mathbf{B}}_d^{(k-1)} \boldsymbol{\alpha}_d.$$

Здесь $\tilde{\mathbf{Q}}_d^{(k)}$ 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

$$\tilde{\mathbf{Q}}_d^{(k,r)} = \mathbf{Q}_d \prod_{\substack{j=1 \\ j \neq d}}^{D} \|\mathbf{a}^{(k,r)}\|_j^2, \tag{1.16}$$

$$\tilde{\mathbf{I}}_d^{(k,r)} = \sum_{\substack{j=1 \\ j \neq d}}^{D} \left(\prod_{\substack{j'=1 \\ j' \neq d, j' \neq j}}^{D} \|\mathbf{a}^{(k,r)}\|_{j'}^2 \cdot \mathbf{a}_j^{(k,r)\intercal} \mathbf{Q}_j \mathbf{a}_j^{(k,r)} \right) \mathbf{I}_{n_d}, \tag{1.17}$$

и

$$\tilde{\mathbf{B}}_d^{(k)} = [\tilde{\mathbf{B}}_d^{(k,1)\intercal}, \ldots, \tilde{\mathbf{B}}_d^{(k,R)\intercal}]^\intercal, \quad \tilde{\mathbf{B}}_d^{(k,r)} = \underline{\mathbf{B}} \prod_{\substack{j=1 \\ j \neq d}}^{D} \times_j \mathbf{a}_j^{(k,r)}. \tag{1.18}$$

С учетом полученного значения $\boldsymbol{\alpha}_d^{(k)}$ пересчитываются матрицы $\tilde{\mathbf{Q}}_d^{(k)}$, $\tilde{\mathbf{I}}_d^{(k)}$ и $\tilde{\mathbf{B}}_d^{(k)}$. Задача (1.15) решается относительно $\boldsymbol{\alpha}_{d+1}$. Циклическая оптимизация повторяется до сходимости $\hat{\underline{\mathbf{A}}}$.

1.5.1. Линейная релаксация задачи выбора признаков

Задача целочисленной оптимизации (1.11) не является выпуклой. Хотя решение невыпуклых задач может привести к более качественным результатам, чем решение релаксированных задач [73], высокая вычислительная сложность затрудняет решение исходных. Чтобы задача решалась более эффективно, используется релаксация невыпуклых ограничений $\underline{\mathbf{A}} \in \{0,1\}^{n_1 \times \cdots \times n_D}$, т.е. их замена на $\underline{\hat{\mathbf{A}}} \in [0,1]^{n_1 \times \cdots \times n_D}$. Вместо задач (1.15) решаются задачи

$$\boldsymbol{\alpha}_d^{(k)} = \underset{\boldsymbol{\alpha} \in [0,1]^{nR}}{\arg\min} F(\boldsymbol{\alpha}_1^{(k)}, \ldots, \boldsymbol{\alpha}_{d-1}^{(k)}, \boldsymbol{\alpha}_d, \boldsymbol{\alpha}_{d+1}^{(k-1)}, \ldots, \boldsymbol{\alpha}_D^{(k-1)}), \tag{1.19}$$

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

$$\underline{\mathbf{A}}(\epsilon) = [a_{i_1 \ldots i_D}], \quad a_{i_1 \ldots i_D} = \begin{cases} 1 \text{ если } \hat{a}_{i_1 \ldots i_D} \geq \epsilon, \\ 0 \text{ иначе.} \end{cases} \tag{1.20}$$

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 $\underline{\hat{\mathbf{A}}}$ релаксированной задачи задает на множестве признаков линейный порядок

$$\boldsymbol{\chi}_{i_1 \ldots i_D} \preceq \boldsymbol{\chi}_{i'_1 \ldots i'_D} \Leftrightarrow \hat{a}_{i_1 \ldots i_D} \leq \hat{a}_{i'_1 \ldots i'_D}. \tag{1.21}$$

Algorithm 1 Многоиндексный алгоритм выбора признаков NQPFS.

Дано: начальное приближение $\boldsymbol{\xi}^{(0)} = [\boldsymbol{\alpha}_1^{(0)}, \ldots, \boldsymbol{\alpha}_D^{(0)}]$, матрицы $Q_1, \ldots, Q_D$, $\underline{\mathbf{B}}$, параметры R, ϵ, ε_F.

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

$k = 0$
while $|F(\boldsymbol{\xi}^{(k)}) - F(\boldsymbol{\xi}^{(k+1)})| > \varepsilon_F$ **do**
  for $1 \leq d \leq D$ **do**
    $\boldsymbol{\alpha}_d^{(k)} = \underset{\boldsymbol{\alpha}_d \in [0,1]^{n_d R}}{\arg\min} F(\boldsymbol{\alpha}_1^{(k)}, \ldots, \boldsymbol{\alpha}_{d-1}^{(k)}, \boldsymbol{\alpha}_d^{(k)}, \boldsymbol{\alpha}_{d+1}^{(k-1)}, \ldots, \boldsymbol{\alpha}_D^{(k-1)})$
  end for
  $k \leftarrow k + 1$
end while
$\underline{\hat{\mathbf{A}}} = \sum_{r=1}^{R} \mathbf{a}_1^{(r)} \times \cdots \times \mathbf{a}_D^{(r)}$
$\underline{\mathbf{A}} = \underline{\mathbf{A}}(\epsilon)$

1.5.2. Анализ сложности многоиндексного метода выбора признаков

Решение задачи квадратичного программирования (1.10) требует полиномиального времени по $n_1 \cdot \dots \cdot n_D$. Чтобы сформулировать задачу оптимизации (1.10) необходимо вычислить $(n_1 \cdot \dots \cdot n_D)^2$ элементов q_{ij} матрицы Q. Предложенный алгоритм совершает несколько итераций, на каждой из которых решается задача (1.15). Задача (1.15), решаемая относительно d-го направления, требует вычисления n_d^2 значений, что существенно меньше чем $(n_1 \cdot \dots \cdot n_D)^2$. Решение каждой из задач (1.15) полиномиально по $n_d \cdot R$. Здесь число R ограничено характеризует желаемую разреженность набора признаков $n_d \cdot R \leq n_d^2$. Даже с учетом числа итераций, асимптотическая сложность $\mathcal{O}\left(n_1^{p_1}\right) + \dots + \mathcal{O}\left(n_D^{p_D}\right)$ предложенного алгоритма, учитывающего многоиндексную структуру данных, меньше, чем сложность $\mathcal{O}\left((n_1 \cdot \dots \cdot n_D)^p\right)$ исходного алгоритма на векторизированных данных. Экспериментальное исследование предложенного алгоритма показало, что одной итерации, как правило, достаточно для сходимости алгоритма.

1.6. Анализ сходимости релаксированной задачи выбора признаков

Задача минимизации (1.13) решается итеративно. В данном разделе сформулированы условия, при которых последовательность $\{\boldsymbol{\alpha}^{(k)}\}$, сгенерированная алгоритмом 1, сходится.

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

$$\|\boldsymbol{\xi}_{k+1} - \boldsymbol{\xi}^*\| \leq \rho \|\boldsymbol{\xi}_k - \boldsymbol{\xi}^*\|.$$

Теорема 1. Для любого $k > 0$ выполнено

$$F\left(\underline{\mathbf{A}}^{(k+1)}\right) \leq F\left(\underline{\mathbf{A}}^{(k)}\right),$$

где $\{\boldsymbol{\alpha}^{(k)}\}$ — последовательность решений, сгенерированных алгоритмом, $\{\underline{\mathbf{A}}^{(k)}\}$ — соответствующая (1.12), (1.14) последовательность индикаторных переменных.

Доказательство. Так как каждое r-е слагаемое в (1.13) вносит независимый в значение $F(\underline{\mathbf{A}})$, будем рассматривать отдельные слагаемые. Для упрощения выкладок рассмотрим случай $D = 3$. Обозначим $\mathbf{a}_1^{(k,r)} = \mathbf{x}^{(k)}, \mathbf{a}_2^{(k,r)} = \mathbf{y}^{(k)}, \mathbf{a}_3^{(k,r)} = \mathbf{z}^{(k)}$.

$$F^{(r)}\left(\underline{\mathbf{A}}^{(k+1)}\right) \equiv F^{(r)}\left(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)}, \mathbf{z}^{(k+1)}\right) = \min_{\mathbf{z}} F^{(r)}\left(\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)}, \mathbf{z}\right) \leq$$

$$F^{(r)}\left(\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)},\mathbf{z}^{(k)}\right)=\min_{\mathbf{y}}F^{(r)}\left(\mathbf{x}^{(k+1)},\mathbf{y},\mathbf{z}^{(k)}\right)\leq F^{(r)}\left(\mathbf{x}^{(k+1)},\mathbf{y}^{(k)},\mathbf{z}^{(k)}\right)=$$

$$\min_{\mathbf{x}}F^{(r)}\left(\mathbf{x},\mathbf{y}^{(k)},\mathbf{z}^{(k)}\right)\leq F^{(r)}\left(\mathbf{x}^{(k)},\mathbf{y}^{(k+1)},\mathbf{z}^{(k)}\right)\equiv F\left(\underline{\mathbf{A}}^{(k)}\right).$$

Тогда

$$F\left(\underline{\mathbf{A}}^{(k+1)}\right)\leq\sum_{r=1}^{R}F^{(r)}\left(\underline{\mathbf{A}}^{(k+1)}\right)\leq\sum_{r=1}^{R}F\left(\underline{\mathbf{A}}^{(k)}\right)\leq F\left(\underline{\mathbf{A}}^{(k)}\right).$$

□

Чтобы показать сходимость $\{\boldsymbol{\alpha}^{(k)}\}$, воспользуемся теоремой из работы [74].

Теорема 2. (Bezdek, Hathaway, 2003) Пусть $\boldsymbol{\alpha}^*$ – локальный минимум функции $F:\mathbb{R}^n\mapsto\mathbb{R}$, такой что

1) $F\in C^2$ в окрестности точки $\boldsymbol{\alpha}^*$,

2) гессиан $\triangle F$ положительно определен в точке $\boldsymbol{\alpha}^*$,

3) F строго выпукла в некоторой окрестности $\boldsymbol{\alpha}^*$.

Тогда существует такая окрестность точки $\mathbf{x}^*$, из которой последовательность решений $\boldsymbol{\alpha}^{(k)}$, определенных последовательной оптимизацией $F(\boldsymbol{\alpha})$ по группам переменных $\boldsymbol{\alpha}_d$, сходится к $\boldsymbol{\alpha}^*$ линейно.

Заметим, что выполнено следующее утверждение. На каждом шаге решается задача выпуклой минимизации (1.19).

Утверждение 1. Пусть матрицы $Q_1,\ldots,Q_D$ положительно определены. Тогда для задачи (1.19) существует единственное решение.

Доказательство. Из положительной определенности Q_d следует, что каждый блок $\tilde{\mathbf{Q}}_d^{(k-1)}+\tilde{\mathbf{I}}_d^{(k-1)}$ матрицы $\tilde{\mathbf{Q}}_d^{(k)}$ положительно определен. Следовательно, $\tilde{\mathbf{Q}}_d^{(k)}$ — положительно определенная матрица, и $\det\tilde{\mathbf{Q}}_d^{(k)}>0$. Поэтому в задаче (1.19) существует глобальный минимум

$$\boldsymbol{\alpha}_d^{(k)}=\frac{1}{2}\tilde{\mathbf{B}}_d^{(k-1)}\left(\tilde{\mathbf{Q}}_d^{(k-1)}+\tilde{\mathbf{I}}_d^{(k-1)}\right)^{-1}. \tag{1.22}$$

□

Обозначим векторизацию индикаторной переменной $\underline{\mathbf{A}}$ через $\boldsymbol{\xi}=\mathrm{vec}\,(\underline{\mathbf{A}})=[\boldsymbol{\alpha}_1^{\mathsf{T}},\ldots,\boldsymbol{\alpha}_D^{\mathsf{T}}]^{\mathsf{T}}\in\mathbb{R}^{(n_1+\ldots n_D)R}$.

Теорема 3. Пусть $Q_1,\ldots,Q_D$ — положительно определенные матрицы с положительными элементами. Пусть точка $\boldsymbol{\xi}_*\in[0,1]^{(n_1+\cdots+n_D)R}$ такова что

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

$$\boldsymbol{\nabla}_{\xi}F|_{\boldsymbol{\xi}_*}=\mathbf{0}_{n_1+\cdots+n_D}, \tag{1.23}$$

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

$$\boldsymbol{\alpha}_d^* = \underset{\|\boldsymbol{\alpha}_d - \boldsymbol{\alpha}_d^*\| < \delta}{\arg\min} F(\boldsymbol{\xi}_{*\backslash d}), \quad \text{где} \quad \boldsymbol{\xi}_{*\backslash d} = \begin{bmatrix} \boldsymbol{\alpha}_1^* \\ \vdots \\ \boldsymbol{\alpha}_{d-1}^* \\ \boldsymbol{\alpha}_d \\ \boldsymbol{\alpha}_{d+1}^* \\ \vdots \\ \boldsymbol{\alpha}_D^* \end{bmatrix}, \tag{1.24}$$

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

$$\underline{\mathbf{B}} \times_1 \left(\mathbf{a}_1^{(r)}\right)^* \cdots \times_D \left(\mathbf{a}_D^{(r)}\right)^* > 0. \tag{1.25}$$

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

Доказательство. Покажем, что существует окрестность $B_\varepsilon(\boldsymbol{\xi}_*) = \{\boldsymbol{\xi} : \ \|\boldsymbol{\xi} - \boldsymbol{\xi}_*\| < \varepsilon\}$, в которой матрица $\mathbf{H}$ вторых производных функции F положительно определена.

Для упрощения выкладок рассмотрим случай $D = 3$. Все выкладки остаются справедливы для произвольного числа $d \geq 2$.

Рассмотрим вначале случай $R = 1$.

Переобозначим $\boldsymbol{\alpha}_1 = \mathbf{x}^{(1)}$, $\boldsymbol{\alpha}_2 = \mathbf{y}^{(1)}$, $\boldsymbol{\alpha}_3 = \mathbf{z}^{(1)}$. Представим F в виде:

$$F(\mathbf{x}, \mathbf{y}, \mathbf{z}) = F_x(\mathbf{x}, \mathbf{y}, \mathbf{z}) + F_y(\mathbf{x}, \mathbf{y}, \mathbf{z}) + F_z(\mathbf{x}, \mathbf{y}, \mathbf{z}) - F_B(\mathbf{x}, \mathbf{y}, \mathbf{z}),$$

где

$$F_x(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \|\mathbf{y}\|^2 \|\mathbf{z}\|^2 \mathbf{x}^\intercal Q_1 \mathbf{x}, \quad F_y(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \|\mathbf{x}\|^2 \|\mathbf{z}\|^2 \mathbf{y}^\intercal Q_2 \mathbf{y},$$
$$F_z(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \mathbf{z}^\intercal Q_3 \mathbf{z}, \quad F_B(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \underline{\mathbf{B}} \times_1 \mathbf{x} \times_2 \mathbf{y} \times_3 \mathbf{z}.$$

Для первых частных производных F_x имеем:

$$\frac{\partial F_x}{\partial x_i} = 2\|\mathbf{y}\|^2 \|\mathbf{z}\|^2 \mathbf{x}^\intercal \mathbf{q}_i, \quad \frac{\partial F_x}{\partial y_j} = 2y_j \|\mathbf{z}\|^2 \mathbf{x}^\intercal Q_1 \mathbf{x}, \quad \frac{\partial F_x}{\partial z_k} = 2z_k \|\mathbf{y}\|^2 \mathbf{x}^\intercal Q_1 \mathbf{x}.$$

Отсюда

$$\frac{\partial^2 F_x}{\partial x_i \partial x_j} = 2\|\mathbf{y}\|^2 \|\mathbf{z}\|^2 q_{ij}, \quad \frac{\partial^2 F_x}{\partial y_i \partial y_j} = 2\|\mathbf{z}\|^2 \mathbf{x}^\intercal Q_1 \mathbf{x} \delta_{ij}, \quad \frac{\partial^2 F_x}{\partial z_i \partial z_k} = 2\|\mathbf{y}\|^2 \mathbf{x}^\intercal Q_1 \mathbf{x} \delta_{ik},$$

$$\frac{\partial^2 F_x}{\partial x_i \partial y_j} = 4y_j \|\mathbf{z}\|^2 \mathbf{x}^\intercal \mathbf{q}_i, \quad \frac{\partial^2 F_x}{\partial x_i \partial z_k} = 4z_k \|\mathbf{y}\|^2 \mathbf{x}^\intercal \mathbf{q}_i, \quad \frac{\partial^2 F_x}{\partial y_j \partial y_k} = 4y_j z_k \mathbf{x}^\intercal Q_1 \mathbf{x}.$$

Аналогично для производных F_y, F_z. Для F_B

$$\frac{\partial^2 F_B}{\partial x_i \partial y_j} = \sum_k B_{ijk} z_k, \quad \frac{\partial^2 F_B}{\partial x_i \partial z_k} = \sum_j B_{ijk} y_j, \quad \frac{\partial^2 F_B}{\partial y_j \partial z_k} = \sum_i B_{ijk} x_i.$$

Таким образом, матрица H вторых производных функции F имеет вид

$$\mathbf{H}(\mathbf{x},\mathbf{y},\mathbf{z}) = 2\begin{bmatrix} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{B}^\intercal & \mathbf{D} & \mathbf{E} \\ \mathbf{C}^\intercal & \mathbf{E}^\intercal & \mathbf{F} \end{bmatrix} - \triangle F_B, \tag{1.26}$$

где

$$\mathbf{A} = \|\mathbf{y}\|^2\|\mathbf{z}\|^2 Q_1 + I_{n_1}\|\mathbf{z}\|^2\mathbf{y}^\intercal Q_2\mathbf{y} + I_{n_1}\|\mathbf{y}\|^2\mathbf{z}^\intercal Q_3\mathbf{z},$$
$$\mathbf{D} = I_{n_2}\|\mathbf{z}\|^2\mathbf{x}^\intercal Q_1\mathbf{x} + \|\mathbf{x}\|^2\|\mathbf{z}\|^2 Q_2 + I_{n_2}\|\mathbf{x}\|^2\mathbf{z}^\intercal Q_3\mathbf{z},$$
$$\mathbf{F} = I_{n_3}\|\mathbf{y}\|^2\mathbf{x}^\intercal Q_1\mathbf{x} + I_{n_3}\|\mathbf{z}\|^2\mathbf{y}^\intercal Q_2\mathbf{y} + \|\mathbf{x}\|^2\|\mathbf{y}\|^2 Q_3,$$
$$\mathbf{B} = 2\|\mathbf{z}\|^2 Q_1\mathbf{x}\mathbf{y}^\intercal + 2\|\mathbf{z}\|^2\mathbf{x}Q_2\mathbf{y}^\intercal + 2\mathbf{x}\mathbf{y}^\intercal\mathbf{z}^\intercal Q_3\mathbf{z}.$$
$$\mathbf{C} = 2\|\mathbf{y}\|^2 Q_1\mathbf{x}\mathbf{z}^\intercal + 2\|\mathbf{y}\|^2\mathbf{x}Q_3\mathbf{z}^\intercal + 2\mathbf{x}\mathbf{z}^\intercal\mathbf{y}^\intercal Q_2\mathbf{y},$$
$$\mathbf{E} = 2\|\mathbf{x}\|^2 Q_2\mathbf{y}\mathbf{z}^\intercal + 2\|\mathbf{x}\|^2\mathbf{y}Q_3\mathbf{z}^\intercal + 2\mathbf{y}\mathbf{z}^\intercal\mathbf{x}^\intercal Q_1\mathbf{x}.$$

Рассмотрим точку $\boldsymbol{\xi} = (\mathbf{x},\mathbf{y},\mathbf{z}) = (\mathbf{x}_* + \Delta\mathbf{x}, \mathbf{y}_* + \Delta\mathbf{y}, \mathbf{z}_* + \Delta\mathbf{z}) = \boldsymbol{\xi} + \Delta\boldsymbol{\xi}$.

$$\frac{1}{2}\boldsymbol{\xi}^\intercal\mathbf{H}\boldsymbol{\xi} = \frac{1}{2}[\mathbf{x}^\intercal,\mathbf{y}^\intercal,\mathbf{z}^\intercal]\mathbf{H}(\mathbf{x}_*,\mathbf{y}_*,\mathbf{z}_*)[\mathbf{x}^\intercal,\mathbf{y}^\intercal,\mathbf{z}^\intercal]^\intercal = \tag{1.27}$$

$$= \|\mathbf{y}_*\|^2\|\mathbf{z}_*\|^2\mathbf{x}^\intercal Q_1\mathbf{x} + \|\mathbf{x}\|^2\|\mathbf{z}_*\|^2\mathbf{y}_*^\intercal Q_2\mathbf{y}_* + \|\mathbf{x}\|^2\|\mathbf{y}_*\|^2\mathbf{z}_*^\intercal Q_3\mathbf{z}_* + \tag{1.28}$$
$$+ \|\mathbf{y}\|^2\|\mathbf{z}_*\|^2\mathbf{x}_*^\intercal Q_1\mathbf{x}_* + \|\mathbf{x}_*\|^2\|\mathbf{z}_*\|^2\mathbf{y}^\intercal Q_2\mathbf{y} + \|\mathbf{x}_*\|^2\|\mathbf{y}\|^2\mathbf{z}_*^\intercal Q_3\mathbf{z}_* + \tag{1.29}$$
$$+ \|\mathbf{y}_*\|^4\|\mathbf{z}\|^2\mathbf{x}_*^\intercal Q_1\mathbf{x}_* + \|\mathbf{x}_*\|^4\|\mathbf{z}\|^2\mathbf{y}_*^\intercal Q_2\mathbf{y}_* + \|\mathbf{x}_*\|^4\|\mathbf{y}_*\|^2\mathbf{z}^\intercal Q_3\mathbf{z} + \tag{1.30}$$
$$+ 4\|\mathbf{z}_*\|^2\mathbf{x}_*^\intercal Q_1\mathbf{x}\mathbf{y}_*^\intercal\mathbf{y} + 4\|\mathbf{z}_*\|^2\mathbf{x}_*^\intercal\mathbf{x}\mathbf{y}_*^\intercal Q_2\mathbf{y} + 2\mathbf{x}_*^\intercal\mathbf{x}\mathbf{y}_*^\intercal\mathbf{y}\mathbf{z}_*^\intercal Q_3\mathbf{z}_* + \tag{1.31}$$
$$+ 4\|\mathbf{y}_*\|^2\mathbf{x}_*^\intercal Q_1\mathbf{x}\mathbf{y}_*^\intercal\mathbf{z} + 4\|\mathbf{y}_*\|^2\mathbf{x}_*^\intercal\mathbf{x}\mathbf{z}_*^\intercal Q_3\mathbf{z} + 2\mathbf{x}_*^\intercal\mathbf{x}\mathbf{z}_*^\intercal\mathbf{z}\mathbf{y}_*^\intercal Q_2\mathbf{y}_* + \tag{1.32}$$
$$+ 4\|\mathbf{x}_*\|^2\mathbf{z}_*^\intercal Q_3\mathbf{z}\mathbf{y}_*^\intercal\mathbf{y} + 4\|\mathbf{x}_*\|^2\mathbf{z}_*^\intercal\mathbf{z}\mathbf{y}_*^\intercal Q_2\mathbf{y} + 2\mathbf{z}_*^\intercal\mathbf{z}\mathbf{y}_*^\intercal\mathbf{y}\mathbf{x}_*^\intercal Q_1\mathbf{x}_* - \tag{1.33}$$
$$- \underline{\mathbf{B}}\mathbf{x}\mathbf{y}\mathbf{z}_* - \underline{\mathbf{B}}\mathbf{x}\mathbf{y}_*\mathbf{z} - \underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}\mathbf{z}. \tag{1.34}$$

Так $\det Q_d > 0$, числа обусловленности $Q_d < +\infty$. Поэтому

$$\|\mathbf{z}_*\|^2\mathbf{x}_*^\intercal Q_1\mathbf{x}\mathbf{y}_*^\intercal\mathbf{y} = \|\mathbf{z}_*\|^2\mathbf{x}_*^\intercal Q_1(\mathbf{x}_* + \Delta\mathbf{x})\mathbf{y}_*^\intercal(\mathbf{y}_* + \Delta\mathbf{y}) =$$
$$\|\mathbf{y}_*\|^2\|\mathbf{z}_*\|^2\mathbf{x}_*^\intercal Q_1\mathbf{x}_* + \mathcal{O}\left(\|\Delta\mathbf{x}\|\right) + \mathcal{O}\left(\|\Delta\mathbf{y}\|\right) = \|\mathbf{y}_*\|^2\|\mathbf{z}_*\|^2\mathbf{x}_*^\intercal Q_1\mathbf{x}_* + \mathcal{O}\left(\|\Delta\xi\|\right),$$
$$\underline{\mathbf{B}}\mathbf{x}\mathbf{y}\mathbf{z}_* = \underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}_*\mathbf{z}_* + \mathcal{O}\left(\|\Delta\xi\|\right).$$

С учетом определения функции $F(\mathbf{x},\mathbf{y},\mathbf{z})$ суммарный вклад внедиагональных блоков (1.31)–(1.34) матрицы $\mathbf{H}$ в (1.27) равен

$$12F(\mathbf{x}_*,\mathbf{y}_*,\mathbf{z}_*) + 9\underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}_*\mathbf{z}_* + \mathcal{O}\left(\|\Delta\boldsymbol{\xi}\|\right),$$

или, в общем случае,

$$2D(D-1)F(\boldsymbol{\xi}_*) + D(2D-1)F_B(\boldsymbol{\xi}_*) + \mathcal{O}\left(\|\Delta\boldsymbol{\xi}\|\right).$$

Вклад (1.28)–(1.30) диагональных блоков равен

$$F(\mathbf{x},\mathbf{y}_*,\mathbf{z}_*)+\underline{\mathbf{B}}\mathbf{x}\mathbf{y}_*\mathbf{z}_*+F(\mathbf{x}_*,\mathbf{y},\mathbf{z}_*)+\underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}\mathbf{z}_*+F(\mathbf{x}_*,\mathbf{y}_*,\mathbf{z})+\underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}_*\mathbf{z}\geq$$

(с учетом условия (1.24))

$$3F(\mathbf{x}_*,\mathbf{y}_*,\mathbf{z}_*)+\underline{\mathbf{B}}\mathbf{x}\mathbf{y}_*\mathbf{z}_*+\underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}\mathbf{z}_*+\underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}_*\mathbf{z}=3F(\mathbf{x}_*,\mathbf{y}_*,\mathbf{z}_*)+3\underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}_*\mathbf{z}_*+\mathcal{O}\left(\|\Delta\boldsymbol{\xi}\|\right).$$

В общем случае, диагональные блоки матрицы вносят вклад

$$\sum_{d=1}^{D}F(\boldsymbol{\xi}_{*\backslash d})+F_B(\boldsymbol{\xi}_{*\backslash d})\geq D\cdot F(\boldsymbol{\xi}_*)+D\underline{\mathbf{B}}\mathbf{x}_*\mathbf{y}_*\mathbf{z}_*+\mathcal{O}\left(\|\Delta\boldsymbol{\xi}\|\right).$$

Из выполнения условий (1.23) получаем:

$$2F(\boldsymbol{\xi}_*)=\mathbf{x}_*^{\intercal}\boldsymbol{\nabla}_{\mathbf{x}}(F_x+F_y+F_z)|_{\xi_*}=F_B(\boldsymbol{\xi}_*).$$

Итого

$$\frac{1}{2}\boldsymbol{\xi}^{\intercal}\mathbf{H}\boldsymbol{\xi}\geq(2D^2-D)F(\boldsymbol{\xi}_*)+2D^2F_B(\boldsymbol{\xi}_*)+\mathcal{O}\left(\|\Delta\boldsymbol{\xi}\|\right)=\frac{1}{2}(6D^2-D)F_B(\xi_*)+\mathcal{O}\left(\varepsilon\right).$$

По условию (1.25) найдется такое ε, что $\frac{1}{2}(6D^2-D)F_B(\xi_*)+\mathcal{O}\left(\varepsilon\right)>0$, и для любого $\xi\in B_\varepsilon(\boldsymbol{\xi}_*)$ выполнено $\frac{1}{2}\boldsymbol{\xi}^{\intercal}\mathbf{H}\boldsymbol{\xi}>0$.

Таким образом, утверждение теоремы доказано для $R=1$. Заметим, что

$$F(\xi)=\sum_{r=1}^{R}F^{(r)}(\xi^r),$$

откуда следует, что матрица $\mathbf{H}$ имеет блочно-диагональный вид, причем для каждого блока выполнены условия теоремы для $R=1$. Следовательно, существует окрестность

$$B_\varepsilon(\boldsymbol{\xi}_*)=\cup_{r=1}^{R}B_\varepsilon^{(r)}\left((\boldsymbol{\xi}_*)^{(r)}\right),$$

в которой матрица $\mathbf{H}$ положительно определена, а функция F — строго выпукла.

□

1.6.1. Вычисление матриц сходства и релевантности в многоиндексном случае

В данной работе матрицы сходства $\mathbf{Q}_d$, для d-го направления $d=1,\ldots,D$ определяются с помощью сингулярного разложения высшего порядка HOSVD (higher order SVD):

$$\underline{\mathbf{X}}=\sum_{l=1}^{L}\lambda_l\cdot\mathbf{u}_0^{(l)}\circ\mathbf{u}_1^{(l)}\circ\cdots\circ\mathbf{u}_D^{(l)}.$$

Матрицы сходства $\mathbf{Q}_d$ вычисляются как

$$\mathbf{Q}_d = \frac{1}{D(L-1)}\mathbf{U}_d\Sigma\mathbf{U}_d^{\mathsf{T}}, \quad \text{где } \Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_L), \tag{1.35}$$

$$\mathbf{U}_d = [\mathbf{u}_d^{(1)}, \ldots, \mathbf{u}_d^{(L)}] \in \mathbb{R}^{n_d \times L}, \; d = 1, \ldots, D.$$

Определение релевантности (1.9) обобщается на многоиндексный случай следующим образом:

$$\underline{\tilde{\mathbf{B}}} = [\tilde{b}_{i_1 \ldots i_D}], \quad \tilde{b}_{i_1 \ldots i_D} = \frac{1}{D}\sum_{n=1}^{D} |\operatorname{corr}(\boldsymbol{\chi}_{i_1 \ldots i_D}, \mathbf{y}_n)|.$$

В соответствии с рекомендациями, приведенными для двумерного случая, в данной работе используется поправка к релевантности:

$$\underline{\mathbf{B}} = [b_{i_1 \ldots i_D}] = \left[\tilde{b}_{i_1 \ldots i_D} + (1-c)\left(q_{i_1 i_1}^{(1)} + \cdots + q_{i_D i_D}^{(D)}\right)\right].$$

Глава 2
Снижение размерности с сохранением локальной структуры близости между объектами выборки

В данном разделе описан метод снижения размерности признакового пространства, сохраняющий локальную структуру близости между объектами в исходном пространстве. Метод основан на предположении, что данные расположены вблизи некоторого многообразия, так что метрически близкие объекты могут быть геодезически далекими. Применение линейных методов снижения размерности в этом случае приведет к нарушению исходной структуры близости между объектами. В работе предлагается модификация метода t-SNE, решающего задачу снижения размерности с сохранением вероятностного распределения отношения ближайшего соседства между объектами. Предлагаемая модификация учитывает также метки классов объектов, что позволяет повысить качество решения задачи классификации при использовании метода t-SNE.

2.1. Стохастическое вложение объектов в пространство меньшей размерности

Обозначим $\mathbb{X} \subset \mathbb{R}^n$ множество всех возможных векторов $\mathbf{x}$ признаков изучаемых объектов. Предполагается, что объекты $\mathbb{X}$ подчиняются *гипотезе многообразия*: найдется гладкое отображение $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$ такое, что

$$\text{для } \mathbf{x} \in \mathbf{X} \text{ существует } \mathbf{z}^* \in \mathbb{R}^d : \mathbf{x} = \mathbf{f}(\mathbf{z}^*) + \boldsymbol{\varepsilon},$$

где $\boldsymbol{\varepsilon}$ — случайный вектор с нулевым матожиданием и конечной матрицей корреляций.

Предполагается, что существует гладкое отображение $f : \mathbb{R}^d \to \mathbb{R}^n$, $d < n$, 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

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

Будем называть d эффективной размерностью исходного пространства $\mathbb{X}$. Она определяется природой признакового пространства. Поскольку d заранее неизвестно, введем понятие результирующего пространства $\mathbb{R}^k$, в котором выполняется поиск решения. В общем случае $k \neq d$. Процесс поиска образов объектов выборки в результирующем пространстве назовем вложением в него.

Рассмотрим выборку из m объектов, заданную матрицей

$$\mathbf{X} = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_m]_{\mathsf{T}}, \quad \mathbf{x}_i \in \mathbb{X}, \quad i = 1 \ldots m. \tag{2.1}$$

Пусть $p_{ij} = P(\mathbf{x}_i, \mathbf{x}_j)$ и $q_{ij} = Q(\mathbf{z}_i, \mathbf{z}_j)$ — расстояния между объектами в $\mathbb{R}^n$ и $\mathbb{R}^k$ соответственно:

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

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

Параметр σ_i в условном распределении p_{ij} задан для каждого i и зависит от расположения $\mathbf{x}_i$ относительно других объектов в исходном пространстве. Если он расположен в области высокой концентрации исходных данных, то коэффициент σ_i имеет меньшие значения, чем если бы концентрация была низкой.

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

$$\mathbf{Z} = [\mathbf{z}_1 \ \ldots \ \mathbf{z}_m]_\intercal \subset \mathbb{R}^k \tag{2.2}$$

минимизируют дивергенцию Кульбака-Лейбера

$$\mathbf{Z}_{\min} = \underset{\mathbf{Z} \in \mathbb{R}^{m \times k}}{\operatorname{argmin}} C(\mathbf{X}, \mathbf{Z}), \tag{2.3}$$

$$C(\mathbf{X}, \mathbf{Z}) = \mathrm{KL}(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{2.4}$$

Заметим, что минимизация происходит только по координатам объектов $\mathbf{z}_1 \ \ldots \ \mathbf{z}_m$ как по переменным, а координаты $\mathbf{x}_1 \ \ldots \ \mathbf{x}_m$ считаются известными константами.

Задача решается градиентными методами [23]. Для инициализации начальных точек $\mathbf{Z}^{(0)} = [\mathbf{z}_1^{(0)} \ \ldots \ \mathbf{z}_m^{(0)}]_\intercal$ градиентного спуска в стандартной реализации было предложено [23] два метода: инициализировать случайными точками, либо использовать для задания начальной инициализации метод главных компонент. От качества начальной инициализации, в случае с невыпуклой задачей оптимизации, зависит не только скорость сходимости к оптимуму, но и локальный минимум, к которому будет сходиться градиентный метод.

2.2. Модификация метода стохастических вложений для обучения с учителем

Рассмотрим задачу классификации с обучающей выборкой $\mathbf{X}$ (2.1) и тестовой выборкой из m' объектов $\mathbf{X}' = [\mathbf{x}_{m+1} \ \ldots \ \mathbf{x}_{m+m'}]_\intercal \subset \mathbb{X}$. Соответственно, метки классов $y_i \in \{0, 1\}, \quad i = 1, \ldots, m$ известны, а $\hat{y}_i, \quad i = m+1, \ldots, m+m'$ необходимо оценить. Так как на этапе обучения данные $\mathbf{X}'$ могут быть недоступны, метод непараметрического t-SNE не применим для снижения размерности

в задачах классификации. Назовем это проблемой непросмотренных объектов ("out-of-sample problem"). Для ее решения предлагается минимизировать (2.4) независимо по различным подмножествам объектов.

Для повышения качества классификатора в результирующем пространстве, предлагается перед вложением обучающей выборки добавить в ней метки классов в качестве признаков и улучшить таким образом начальное приближение градиентного метода. Идея такого подхода заключается в том, что, поскольку t-SNE сохраняет только локальную структуру схожести между объектами, после проведения процедуры понижения размерности, классифицируемые объекты отображаются в кластеры, предварительно разнесенные учетом меток. При этом используется предположение, что объекты из $\mathbf{X}'$ больше схожи с объектами $\mathbf{X}$ того же класса, чем с объектами противоположного. Таким образом удается увеличить расстояние между образами классифицируемых объектов из различных классов, что упрощает их классификацию. На диаграммах (2.5) и (2.6) показаны основные отображения оригинального непараметрического t-SNE

$$\mathbf{X} \in \mathbb{R}^{m\times n} \qquad \mathbf{Z} \in \mathbb{R}^{m\times k} \tag{2.5}$$

$$\mathbf{X}' \in \mathbb{R}^{m'\times n} \qquad \mathbf{Z}' \in \mathbb{R}^{m'\times k}$$

и предложенной модификации

$$\mathbf{X}|\mu\mathbf{y} \in \mathbb{R}^{m\times(n+1)} \quad \begin{matrix}\text{Начальная партия}\\ \text{Дополнительные партии}\end{matrix} \quad \mathbf{Z} \in \mathbb{R}^{m\times k} \tag{2.6}$$

$$\mathbf{X}' \in \mathbb{R}^{m'\times n} \qquad \mathbf{Z}'^{(0)} \in \mathbb{R}^{m'\times k} \qquad \mathbf{Z}' \in \mathbb{R}^{m'\times k}$$

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

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

где μ — вес меток как признаков. В модифицированном алгоритме на основе расширенной матрицы $\tilde{X}$ выполняется поиск образов $\mathbf{Z}$ (2.4), на которых обучается классификатор. Таким образом, при построении вложения обучающей выборки решается задача

$$\mathbf{Z}_{\min} = \operatorname*{argmin}_{\mathbf{Z}\in\mathbb{R}^{m\times k}} C\left(\left(\mathbf{X} \mid \mu\mathbf{y} \right), \mathbf{Z}\right).$$

2.3. Вложение новых объектов в пространство со сниженной размерностью

Обозначим через $\mathbf{Z}' = [\mathbf{z}_{m+1} \ \dots \ \mathbf{z}_{m+m'}]^{\mathsf{T}}$ образы $\mathbf{X}'$ в результирующем пространстве. Аналогично (2.3), сформулируем задачу поиска $\mathbf{Z}'$ в виде m' задач k-

мерной минимизации, которые могут быть решены независимо:

$$\mathbf{z}_i^{\min} = \underset{\mathbf{z}_i \in \mathbb{R}^{m'}}{\operatorname{argmin}} C\left(\left[\frac{\mathbf{X}}{\mathbf{x}_i{}^{\mathsf{T}}}\right], \left[\frac{\mathbf{Z}}{\mathbf{z}_i{}^{\mathsf{T}}}\right]\right), \qquad i = m+1, \ldots, m+m',$$

где матрицы $\left[\frac{\mathbf{X}}{\mathbf{x}_i{}^{\mathsf{T}}}\right]$ и $\left[\frac{\mathbf{Z}}{\mathbf{z}_i{}^{\mathsf{T}}}\right]$ получены из $\mathbf{X}$ и $\mathbf{Z}$ добавлением строк $\mathbf{x}_i{}^{\mathsf{T}}$ и $\mathbf{z}_i{}^{\mathsf{T}}$ соответственно. При использовании такого подхода предполагается, что в обучающая выборка $\mathbf{X}$ (2.1) достаточно репрезентативна.

Для инициализации образов $\mathbf{z}_{i'}$ классифицируемых объектов предлагается использовать метод взвешенного среднего:

$$\mathbf{z}_{i'}^{(0)} = \sum_{i=1}^{m} \mathbf{z}_i w_{ii'}, \quad \sum_{i=1}^{m} w_{ii'} = 1, \quad i' = m+1, \ldots, m+m'$$

где веса $w_{ii'}$ зависят от расстояния между $\mathbf{x}_i$ и $\mathbf{x}_{i'}$. В работе рассмотрены два способа задания весов:

$$w_{ii'}^{\text{softmax}} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_{i'}\|)}{\sum\limits_{k=1}^{m} \exp(-\|\mathbf{x}_k - \mathbf{x}_{i'}\|)}, \text{ или} \tag{2.7}$$

$$w_{ii'}^{\text{stud}} = \frac{(1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}{\sum\limits_{k=1}^{m} (1 + \|\mathbf{x}_k - \mathbf{x}_{i'}\|^2)^{-1}}. \tag{2.8}$$

Для ускорения процедуры вложения при работе с большими данными предлагается процедура поэтапного вложения объектов блоками, размер которых, S_s для первого по очереди и S_b для всех остальных, много меньше размера m всей выборки.

2.4. Экспериментальные исследования предложенного алгоритма снижения размерности

В данном разделе приведены результаты экспериментального исследования свойств предложенного алгоритма. Рассмотрены способы инициализации вложения тестовых данных. Использовались четыре различных подхода:

- случайный — инициализация случайным образом,
- "PCA" — инициализация образами при снижении размерности методом главных компонент,
- "Softmax" и "Student" — задаваемые по формулам (2.7) и (2.8).

На рисунках 2.1, 2.2, 2.3, 2.4, 5.11, 5.12 изображены результаты экспериментов, проведенных при использовании перечисленных способов инициализации.

Все инициализированные таким образом объекты далее преобразуются, минимизируя (2.4) при фиксированных образах объектов обучающей выборки. После этого происходит классификация полученных образов.

Методы инициализации (2.7) и (2.8) были предложены так, чтоб инициализированные данные обладали свойством сохранения локальной структуры исходной выборки. Предполагается, что это улучшит сходимость градиентного метода, используемого для минимизации (2.4), по сравнению с инициализациями PCA и random.

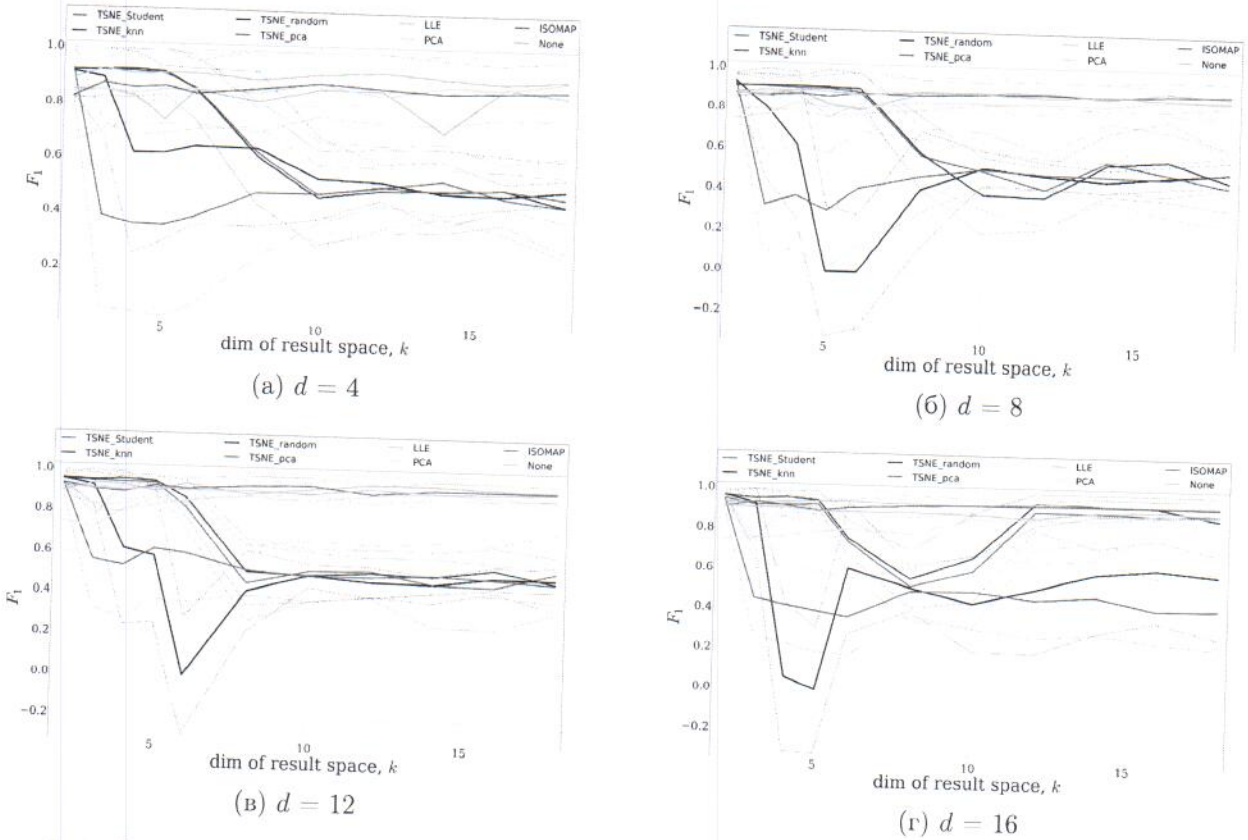
В данном разделе для эмпирического исследования свойств предлагаемого алгоритма использовались синтетические выборки $\mathbf{X} = [\mathbf{x}_1 \ \ldots \ \mathbf{x}_m]_{\mathrm{T}}$. Для любого вектора $\mathbf{x}_i$ компоненты сгенерированы как стандартные нормальные распределения на гранях гиперкуба. При этом эффективная размерность выборки составляет d, а оставшиеся признаки являются шумовыми. Далее выборка сворачивалась в спираль по одной из размерностей. Это делалось для того, чтоб реализовать предположение о существовании многообразия меньшей размерности, в котором содержится выборка. Генерировалось одинаковое количество объектов разных классов, а на обучение и контроль выборка разбивалась в соотношении $1:4$.

В этом разделе описывается исследование качества классификации с применением предлагаемого алгоритма в зависимости от основных его параметров и специфики выборки. Для сравнения предлагаемого алгоритма и его исследования рассматривается классификация в комбинации с другими методами снижения размерности: анализ главных компонент [19], локальное линейное вложение LLE [30], изометрическое отображение ISOMAP [28], а также без применения снижения размерности. Для построения классификатора использовался метод логистической регрессии на основе стохастического градиентного спуска [75].

На рис. 2.1 изображена зависимость меры качества F_1 от размерности вложения k при различных значениях эффективной размерности d. Пунктиром отмечено стандартное отклонение, срезанное по уровню единицы. На графиках видно, что качество значительно ухудшается при увеличении k, независимо от соотношения k и d. Эксперимент проведен при постоянных $m = 500, n = 20, S_b = 100, S_s = 400, \mu = 150$.

На рис. 2.2 изображена зависимость F_1 от веса μ меток класса y в стартовой выборке при различных значениях размерности выборки n. Из них можно сделать вывод, что качество классификации повышается с ростом μ, при этом скорость роста падает с ростом n. Также видно, что разработанный метод при достаточно больших значениях μ показывает в среднем лучшие результаты среди всех рассмотренных методов снижения размерности, а также превосходит по качеству классификацию в исходном пространстве. Эксперимент проведен при постоянных $m = 500, k = 3, S_b = 100, S_s = 400$. В этом эксперименте все исходные признаки были информативными.

Для исследования зависимости качества классификации от величины отношения размера стартовой части к размеру выборки S_s/m был поставлен экспе-



(а) $d = 4$

(б) $d = 8$

(в) $d = 12$

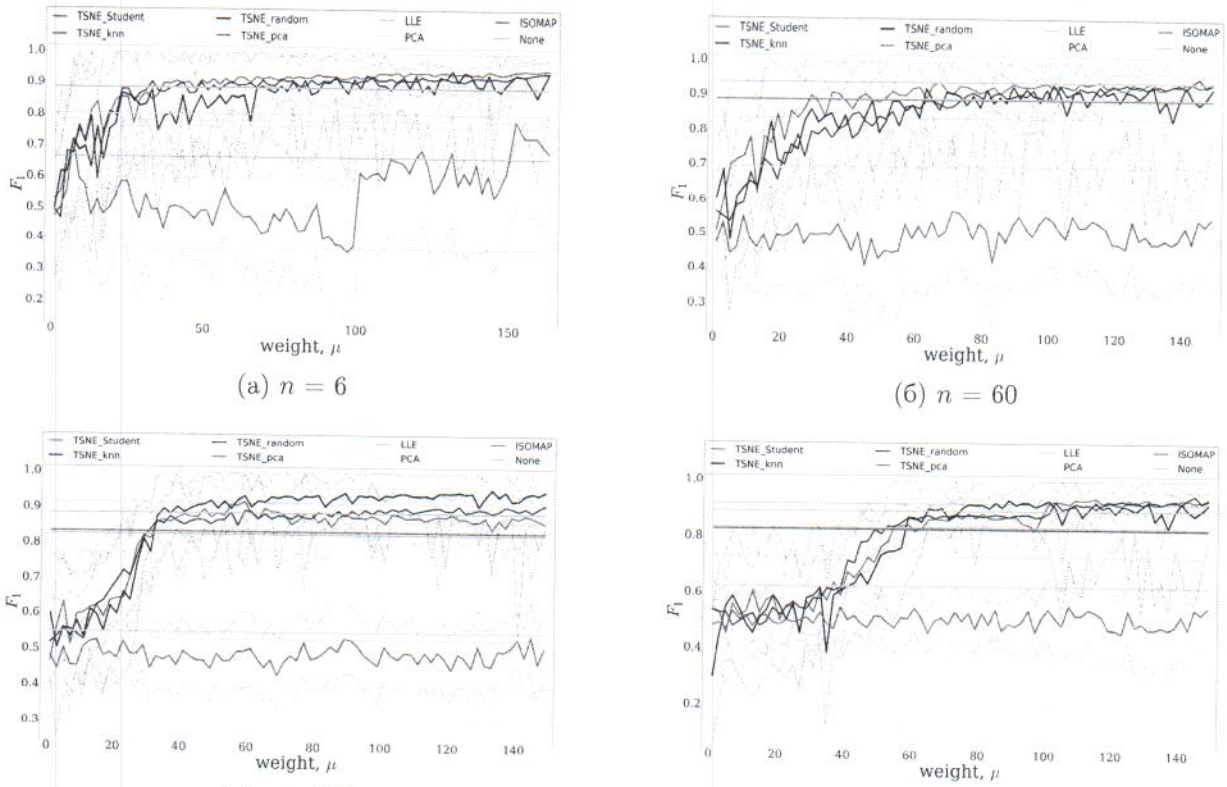
(г) $d = 16$

Рис. 2.1. Зависимость F_1 от k при различных эффективных размерностях выборки d.

римент, где при постоянных $n = 6, k = d = 3, \mu = 150$ исследовалась зависимость меры качества F_1 от размера выборки m и размера начального вложения S_s. При этом размер дополнительно вкладываемых блоков S_b принимался заведомо больше размера выборки m, так что дополняющая часть не разбивалась на блоки. На рис. 2.3 выведены результаты. Можно видеть, что зависимость от этих параметров незначительна. При этом скорость работы алгоритма увеличивается при наличии разбиений на стартовую и дополняющую часть. Таким образом показано, что предложенная модификация алгоритма позволяет значительно ускорить его работу без существенного снижения качества.

Графики на рисунке 2.3 демонстрируют, что самое высокое качество классификации достигается при использовании методов инициализации с помощью (2.8) и случайной инициализации. Результаты метода инициализации PCA не вынесены на рисунок, так как метод показал результаты порядка 0.5.

Целью эксперимента, результаты которого приведены на рис. 2.4, было исследование зависимости значения функции качества классификации от S_b. Он был проведен при постоянных $n = 6, k = d = 3, \mu = 150, S_s = 200$. В результате было обнаружено, что предлагаемый метод устойчив относительно параметра S_b.



(а) $n = 6$

(б) $n = 60$

(в) $n = 300$

(г) $n = 600$

Рис. 2.2. Зависимость F_1 от μ при различных размерностях выборки n.

Глава 3
Сегментация квазипериодических временных рядов

В данном разделе предложен метод разбиения квазипериодического временного ряда на интерпретируемые сегменты.

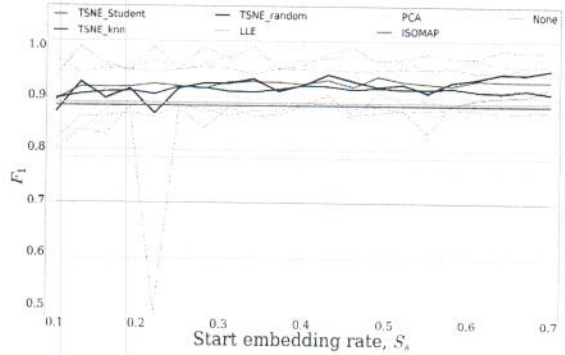
3.1. Задача выделения границ периодов как задача классификации

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.

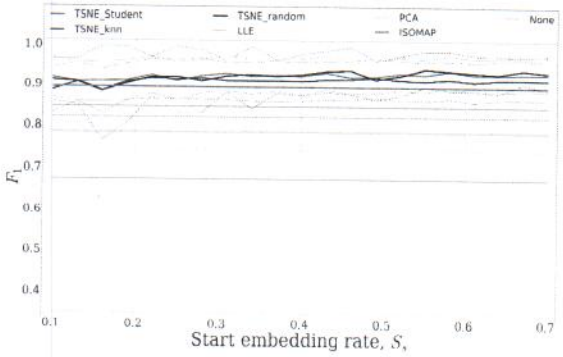
Определение 12. *Квазипериодический временной ряд* $\mathbf{s}$ определяется набором $\langle \mathbf{s}^*, a(i,s), f(i) \rangle$, так что

$$s(i) = a(i, s_{[f(t)]}), \tag{3.1}$$

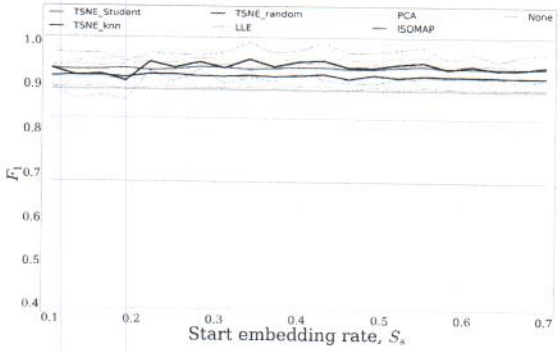
где вектор $\mathbf{s}^* = [s_1, \ldots, s_T]^{\mathsf{T}}$ описывает базовый сегмент, функция $a(i,s), i \in \{1, \ldots, m\}$ модифицирует форму базового сегмента, а кусочно-заданная функция $f(i) \mapsto \{1, \ldots, T\}$ задает масштабирование по времени. Функция f имеет конечное множество точек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



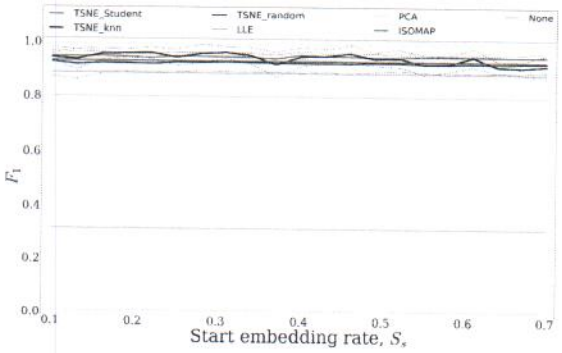
(а) $m = 1000$



(б) $m = 2000$

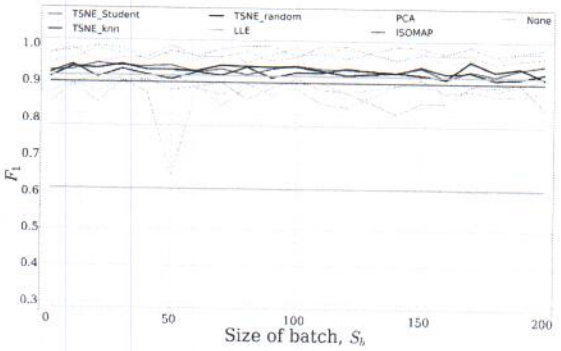


(в) $m = 3000$

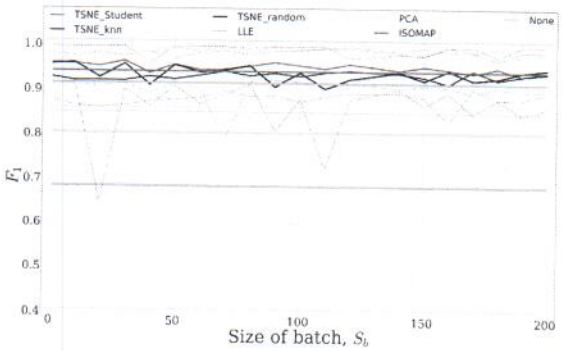


(г) $m = 4000$

Рис. 2.3. Зависимость F_1 от $\frac{S_s}{m}$ при различных размерах выборки m.



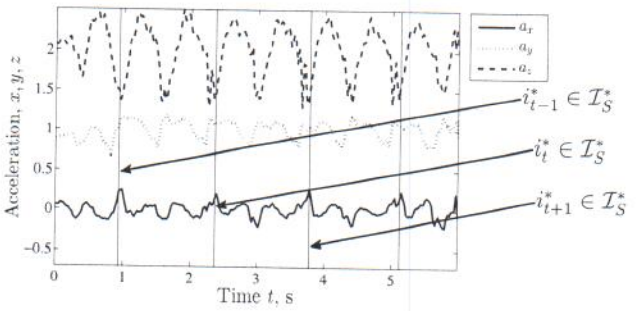
(а) $m = 500$



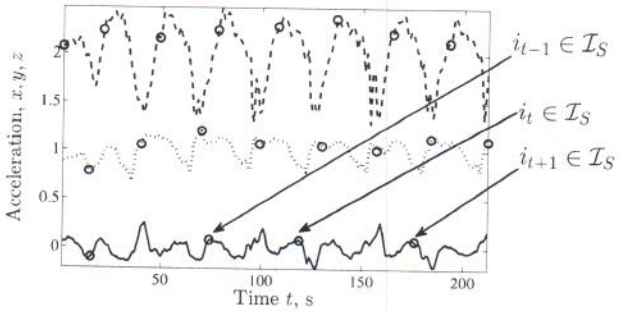
(б) $m = 1000$

Рис. 2.4. Зависимость F_1 от S_b при различных размерах выборки m.

Точки $\mathcal{I}_{\mathcal{S}}^* = \{i_1^*, \ldots, i_S^*\} = \{i_t^* \mid f(i_t + 1) < f(i_t)\}$ считаются концами периодов, которые алгоритму сегментирования необходимо найти. Результатом решения задачи сегментирования квазипериодического ряда является разбиение $\mathcal{I}_{\mathcal{S}} = \{i_1, \ldots, i_S\}$ временного ряда $\mathbf{s}$ на последовательность $\mathcal{S} = \{S_1, \ldots, S_T\}$ сегментов $S_t = [s(i_t), \ldots, s(i_{t+1} - 1)]^\intercal$, соответствующих преобразованному базовому периоду (3.1). На рисунке 3.1 изображен пример временного ряда, разбитого на периоды. На рисунке изображены временные ряды проекции a_x, a_y и a_z ускорения, измеренного при ходьбе, на оси ординат. Временные ряды были измерены с помощью акселерометра мобильного телефона и сегментирова-



(а) Истинные границы временных рядов $\mathcal{I}_S{}^*$



(б) Автоматически выделенные временные ряды $\mathcal{I}_S$

Рис. 3.1. Временные ряды ускорения a_x, a_y и a_z. (a) Вручную сегментированные временные ряды; концы сегментов $i^*_t \in \mathcal{I}^*_S$ обозначены вертикальными линиями. (b) Результаты автоматического сегментирования. Концы i_t выделенных сегментов обозначены точками.

ны вручную. Точки сегментации — концы периодов — $i^*_t \in \mathcal{I}^*_S$, выбранные вручную, называются *истинными*. Истинные точки сегментации показаны на рисунке 3.1(a) вертикальными линиями. На рисунке 3.1(b) изображены кружками отмечены выбранные алгоритмом точки сегментации $i_t \in \mathcal{I}_S$ для тех же временных рядов.

В данной работе термин *сегментирование* используется для обозначения любой процедуры разбиения временного ряда на сегменты; сегментирование временного ряда на сегменты, соответствующие периодом называется *выделение периодов*. Также, результат сегментирования называется сегментацией временного ряда; в случае выделения периодов полученные сегменты называются (выделенными) периодами.

Для оценки результатов процедуры выделения периодов, используется методика, описанная в работе [76]. Авторы работы [76] рассматривают сегментирование временного ряда как задачу двухклассовой классификации набора точек $\{s(1), \ldots, s(m)\}$. Для фиксированного набора истинных точек сегментации $\mathcal{I}^*_S = \{i^*_1, \ldots, i^*_S\}$ класс каждого объекта $x(i)$ определяется как

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

Тогда результаты сегментирования можно оценивать, используя стандартные

меры качества классификации, например точность и полноту, с некоторыми поправками на специфику задачи сегментирования: временные ряды измерены с конечной частотой сэмплирования. Предлагается ввести параметр ϵ, в рамках которого отклонение точек $i_t \in \mathcal{I}_\mathcal{S}$ от $\mathcal{I}^*_\mathcal{S}$ не считается ошибкой классификации. Таким образом, если некоторая точка множества $\mathcal{I}_\mathcal{S}$ достаточно близка к некоторой точке i^*_t, $|i_t - i^*_t| < \epsilon$, и никакие другие точки из $\mathcal{I}_\mathcal{S}$ в ϵ-окрестности i^*_t не содержатся, то ϵ-окрестность i^*_t содержит только правильно классифицированные, ровно одна из которых принадлежит положительному классу. Если $q \geq 1$ точек из $\mathcal{I}_\mathcal{S}$ принадлежат окрестности $(i^*_t - \epsilon, \ldots, i^*_t, \ldots, i^*_t + \epsilon)$, то $q - 1$ из них считаются классифицированными неправильно — это ложные положительные срабатывания алгоритма. Если $q = 0$, считается что в точке $s(i^*_t)$ сделана ошибка — ложно отрицательное срабатывание. Так как примеров отрицательного класса в задаче сегментирования временного ряда гораздо больше, чем положительных примеров, количество правильно распознанных негативных примеров неинформативно для оценки качества классификации. Поэтому в данной работе для оценки используется коэффициент F_1

$$F_1 = \frac{2TP}{2TP + FN + FP}. \tag{3.2}$$

Здесь TP, TN, FP и FN обозначают количество правильно классифицированных положительных, правильно классифицированных отрицательных неправильно классифицированных отрицательных и неправильно классифицированных положительных объектов. Коэффициент F_1 принимает значения от 0 до 1, причем более высокие значения F_1 соответствуют более высокому качеству классификации.

3.2. Декомпозиция временного ряда на периодическую и апериодическую компоненты

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

$$\mathbf{s} = \hat{\mathbf{s}} + \tilde{\mathbf{s}} + \boldsymbol{\varepsilon}, \tag{3.3}$$

где $\hat{\mathbf{s}}$ — нестационарный тренд, $\tilde{\mathbf{s}}$ — периодическая компонента временного ряда, а $\boldsymbol{\varepsilon}$ — шум. Каждая составляющая декомпозиции (3.3) может быть приближена [77] комбинацией главных компонент траекторной матрицы

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

центрированного временного ряда $\mathbf{s}$

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

Параметр N, определяющий длину наибольшего периода, детектируемого методом главных компонент будем называть длиной окна. Метод максимально точно воспроизводит периодическую компоненту с периодом T, если длина окна N кратна T.

Для аппроксимации временного ряда $\tilde{\mathbf{s}}$ с помощью сингулярного разложения $\mathbf{H}$

$$\frac{1}{N}\mathbf{H}^{\mathsf{T}}\mathbf{H} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N) \tag{3.4}$$

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

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

Для аппроксимации временного ряда $\tilde{\mathbf{s}}$ выбираются несколько компонент $\mathbf{y}_j$, из которых составляется матрица $\tilde{\mathbf{H}}$. Искомая составляющая $\tilde{\mathbf{s}}$ исходного временного ряда $\mathbf{s}$ восстанавливается путем анти-диагонального усреднения матрицы $\tilde{\mathbf{H}}$. Для извлечения периодической составляющей необходимо выделить соответствующие искомой периодике главные компоненты.

Теорема 4. Для временного ряда

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

вида

$$s(i) = A_i \cos(2\pi w i + \phi) \tag{3.5}$$

с $w \in (0, 1/2)$, $\phi \in [0, 2\pi)$, $m \cdot w \in \mathbb{N}$ и $A_i : \exists C \in \mathbb{R} |A_i| < C \, \forall i$ главные компоненты $\mathbf{y}_1$ и $\mathbf{y}_2$ могут быть представлены в виде

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

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

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

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

Доказательство. Воспользуемся результатом из работы [78] для временного ряда $\mathbf{s}$ вида

$$s(i) = Ae^{\alpha i}\cos(2\pi w i + \phi)$$

с $w \in (0, 1/2)$, $A, \ \alpha \in \mathbb{R}$, $\phi \in [0, 2\pi)$ и $m \cdot w \in \mathbb{N}$. Было показано, что при $\alpha \to 0$ и $m \to \infty$, так что $m \cdot \alpha \to \gamma \in \mathbb{R}$ главные компоненты $\mathbf{y}_1$ и $\mathbf{y}_2$ могут быть представлены в виде

$$y_1(l) = B_1 e^{\alpha l}\cos(2\pi w l + \phi_1),$$

$$y_2(l) = B_2 e^{\alpha l}\cos(2\pi w l + \phi_2),$$

$$\text{где } \phi_1, \ \phi_2 \in [0, 2\pi), \ l = 1, \ldots, m - N + 1,$$

и $|\phi_1 - \phi_2| \to \pi/2$.

Рассмотрим временной ряд более общего вида (3.5). Собственные векторы $\mathbf{v}_j$ траекторной матрицы $\mathbf{H}$ образуют [79] ортонормированный базис в линейном векторном пространстве, натянутом на столбцы матрицы $\mathbf{H}$. Поэтому для временного ряда вида (3.5) найдутся [77] два собственных вектора $\mathbf{v}_1, \mathbf{v}_2$ с ненулевыми собственными значениями $\lambda_1, \lambda_2 \neq 0$. Соответствующие главные компоненты $\mathbf{y}_1, \mathbf{y}_2$ имеют вид

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

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

Сравнивая выражение

$$\sum_{l=1}^{m-N+1} A_{i+l}A_{l+k}\cos(2\pi w(l+i)+\phi)\cos(2\pi w(l+k)+\phi)$$

для (i,k)-го элемента матрицы $\mathbf{H}^{\mathsf{T}}\mathbf{H}$ с $N\mathbf{v}_i^{\mathsf{T}}\mathbf{\Lambda}\mathbf{v}_k$, получаем, используя уравнение (3.4), что

$$B_j(l) \sim A_l/\sqrt{N\lambda_l}.$$

Тогда $|A_i| \leq C \Rightarrow C_2 \leq B_1(l), B_2(l) \leq C_1$. Условие ортогональности $\mathbf{v}_1^{\mathsf{T}}\mathbf{v}_2 = 0$ дает

$$0 = \frac{1}{N}\sum_{l=1}^{m-N+1} B_1(l)^2\cos(2\pi w l+\phi_1)\cos(2\pi w l+\phi_2) =$$

$$\sum_{l=1}^{m-N+1} B_1(l)^2\cos(4\pi w l+\phi_1+\phi_2) \tag{3.6}$$

$$+\cos(\phi_1-\phi_2)\sum_{l=1}^{m-N+1} B^2(l). \tag{3.7}$$

Рассмотрим (3.6). Так как $B_j(l) \leq C_1$,

$$\sum_{l=1}^{m-N+1} B_1(l)B_2(l)\cos(4\pi w l+\phi_1+\phi_2) \leq$$

$$C_1^2\sum_{l=1}^{m-N+1}\cos(4\pi w l+\phi_1+\phi_2) =$$

$$C_1^2\sum_{l=1}^{m-N+1}\cos(4\pi w\cdot l)\cos(\phi_1+\phi_2) -$$

$$C_1^2 \sum_{l=1}^{m-N+1} \sin(4\pi w \cdot l)\sin(\phi_1+\phi_2).$$

Для $\sum_l \cos(lx)$ и $\sum_l \sin(lx)$ имеем:

$$\sum_{l=1}^{L} \cos(lx) = \frac{\cos x - 1 - \cos\big((L+1)x\big) + \cos Lx}{2 - 2\cos x},$$

$$\sum_{l=1}^{L} \sin(lx) = \frac{3\sin x - \sin\big((L+1)x\big) - \sin Lx}{2 - 2\cos x}.$$

Тогда для любого m получаем ограничение сверху на сумму в уравнении (3.6):

$$\sum_{l=1}^{m-N+1} B_1(l)B_2(l)\cos(4\pi wl + \phi_1 + \phi_2) \leq$$

$$C_1^2\left(\frac{2\cos(\phi_1+\phi_2) + 5\sin(\phi_1+\phi_2)}{4}\right).$$

Для суммы в уравнении (3.7), так как $A_i \geq C_2$, получаем

$$\sum_{l=1}^{m-N+1} B^2(l) \geq (m-N+1)C_2^2.$$

Если $\cos(\phi_1 - \phi_2) \not\to 0$, то выражение в (3.7) неограниченно. Для выполнения условия ортогональности должно выполняться $|\phi_1 - \phi_2| \to \pi/2$.

Получим оценки λ_1, λ_2 для постоянных A_i: $A_i \equiv A$. Согласно (3.4),

$$\lambda_j \approx \frac{A^2}{N}\sum_{l=1}^{N}\cos^2(2\pi wl) = \frac{A^2}{2N}\sum_{l=1} N(1 + cos(4\pi wl))$$

или

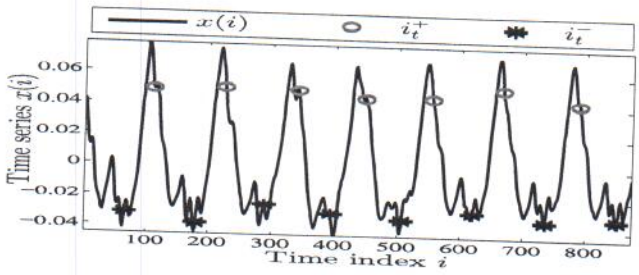
$$\lambda_j \approx \frac{A^2}{N}\sum_{l=1}^{N}\sin^2(2\pi wl) = \frac{A^2}{2N}\sum_{l=1} N(1 - cos(4\pi wl)).$$

Тогда значение λ_j можно оценить как

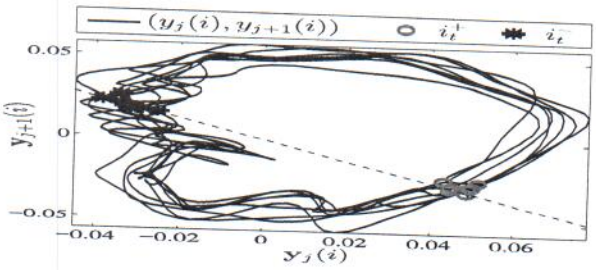
$$\lambda_{\cos} \approx \frac{A^2}{2N}\int_0^N \big(1 + \cos(4\pi wt)\big)dt = \frac{A^2}{2}\left(1 - \frac{\sin(4\pi wN)}{4\pi wN}\right)$$

для $y_j \sim \cos wj$,

$$\lambda_{\sin} \approx \frac{A^2}{2N}\int_0^N \big(1 - \cos(4\pi wt)\big)dt = \frac{A^2}{2}\left(1 + \frac{\sin(4\pi wN)}{4\pi wN}\right)$$



(а)



(б)

Рис. 3.2. Рассечение траектории главных компонент временного ряда $\mathbf{s}$ для выделения полупериодов. (a) Временной ряд $\mathbf{s}$ и выделенные полупериоды $S_t^- = [s(i_t^-), \ldots, s(i_{t+1}^+ - 1)]^\intercal$ и $S_t^+ = [s(i_t^+), \ldots, s(i_{t+1}^- - 1)]^\intercal$. Множества $\mathcal{I}_\mathcal{S}^-$ и $\mathcal{I}_\mathcal{S}^+$ по отдельности определяют способы разбиения временного ряда на периоды. (b) Траектория главных компонент рассеченная произвольной осью. Кружками и звездочками отмечены точки с одинаковыми фазовыми углами. Эти точки образуют множества $\mathcal{I}_\mathcal{S}^+$ и $\mathcal{I}_\mathcal{S}^-$.

$$\text{для } y_j \sim \sin wj.$$

Собственные значения асимптотически (при $N \to \infty$) равны. Более того, при $w = 1/2$ и $T = 2$, только одна главная компонента описывает периодику для X.

□

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. В приложении также показано, что парные периодические компоненты имеют (асимптотически) равные главные значения. Это позволяет рассматривать только последовательные пары главных компонент $(\mathbf{y}_j,\ \mathbf{y}_{j+1})$ в процедуре автоматического отбора периодических компонент, описанной в разделе "Автоматический выбор пары главных компонент".

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 $(\mathbf{y}_j,\ \mathbf{y}_{j+1})$.

Заметим, что условия теоремы выполняются для любых квазипериодических временных рядов без масштабирования времени. После масштабирования периодов по времени соотношение (3.5) лишь приближено описывает временной ряд, и утверждение теоремы может не выполняться.

3.2.1. Рассечение фазовой траектории главных компонент осью симметрии

Рассмотрим строго периодический временной ряд $\mathbf{s} = \{s(i), i = 1, \ldots, m\}$ с периодом T:

$$s(i+T) = s(i).$$

Согласно рассуждениям, приведенным выше, периодическая составляющая $\tilde{\mathbf{s}} = \mathbf{s}$ временного ряда $\mathbf{s}$ содержит две периодические компоненты $\mathbf{y}_1$, $\mathbf{y}_2$ с неотрицательными главными значениями $\lambda_1 \neq 0$, $\lambda_2 \neq 0$: синус и косинус с периодом T. Фазовая траектория нормализованных главных компонент $\mathbf{y}_1^*$ и $\mathbf{y}_2^*$

$$\mathbf{y}_j^* = \frac{\mathbf{y}_j}{\sqrt{\sum_{l=1}^{m-N+1} y_j^2(l)}}, \; j = 1, 2$$

имеет форму спирали с радиусом 1 и центром в точке (0,0):

$$y_1^*(l) \approx \cos(2\pi w l + \phi),$$

$$y_2^*(l) \approx \sin(2\pi w l + \phi).$$

Далее будем считать, что главные компоненты нормализованы.

Один цикл соответсвует полному периоду временного ряда. Если радиус меняется не сильно, то точки временного ряда $math\tilde{h}bfs$ с приблизительно равными координатами (y_1, y_2) соответствуют одной и той же фазе и могут быть выбраны в качестве точек сегментации.

Разрезая фазовую траекторию вдоль оси, проходящей через начало координат ($y_2 = ky_1$ or $y_1 = 0$),получаем разбиение $\mathcal{I}_{\mathcal{S}} = \{i_1^-, i_1^+, \ldots, i_T^-, i_T^+\}$ исходного временного ряда на отрицательные $S_t^- = [s(i_t^-), \ldots, s(i_{t+1}^- - 1)]^\intercal$:

$$y_2(i) - ky_1(i) < 0 \;\; \text{для всех } i \in \{i_t^-, \ldots, i_{t+1}^- - 1\}$$

$$(y_1(i) < 0 \;\; \text{для всех } i \in \{i_t^-, \ldots, i_{t+1}^- - 1\}), t \leq T/2$$

и положительные полупериоды $S_t^+ = [x(i_t^+), \ldots, x(i_{t+1}^+ - 1)]^\intercal$:

$$y_2(i) - ky_1(i) > 0 \;\; \text{для всех } i \in \{i_t^+, \ldots, i_{t+1}^+ - 1\}$$

$$(y_1(i) > 0 \;\; \text{для всех } i \in \{i_t^+, \ldots, i_{t+1}^+ - 1\}), t \leq T/2.$$

Пусть $i_t^- < i_t^+$, тогда $i_t^- = i_t^+ - 1$. Сшивая полупериоды S_t^- и S_t^+, получим сегментацию $\mathcal{I}_{\mathcal{S}} = \{i_1^-, \ldots, i_T^-\}$ временного ряда на искомые периоды.

Рисунок 3.2 иллюстрирует процедуру разбиения временного ряда на полупериоды. На рисунке 3.2(a) изображен квазипериодический временной ряд $\mathbf{s}$, определенный набором $\langle \mathbf{s}, a, f \rangle$. Кружками и звездами отмечены концы выделенных полупериодов: i_t^+ и i_y^-, соответственно. Фиксируем пару компонент

$(\mathbf{y}_j, \mathbf{y}_{j+1})$ временного ряда X (вопрос выбора пары компонент будет рассмотрен позже). Фазовая траектория выбранных компонент изображена на рисунке 3.2(b). Если длина периодов меняется со временем, нормализованные компоненты образуют двухмерную спираль с медленно меняющимся радиусом, причем циклы содержат различное число точек. Аналогично, точки с одинаковой фазой имеют в пространстве $(y_j,\ y_{j+1})$ лишь приблизительно одинаковые координаты. При разбиении траектории некоторой осью (на рисунке 3.2(b) ось отложена пунктиром) формируются полуциклы, соответствуют полупериоды. Каждое из множеств $\mathcal{I}_{\mathcal{S}}^{+}$ и $\mathcal{I}_{\mathcal{S}}^{-}$ конечных точек этих полупериодов определяет сегментацию временного ряда. На рисунке 3.2(a) изображены оба варианта сегментации. Разница заключается в разбросе фазовых координат: так как временной ряд не является строго периодическим, точки $s(i_t),\ i_t \in \mathcal{I}_{\mathcal{S}}$ имеют различные фазовые координаты $(y_j(i_t), y_{j+1}(i_t))$. В качестве $\mathcal{I}_{\mathcal{S}}$ выбирается множество $\mathcal{I}_{\mathcal{S}}^{\bullet}$ с меньшим расстоянием между $(y_j(i_t^{\bullet}), y_{j+1}^{\bullet})$:

$$\mathcal{I}_{\mathcal{S}} = \underset{\bullet\in\{+,-\}}{\arg\min} \sum_{i_t^{\bullet}, i_q^{\bullet}\in\mathcal{I}_{\mathcal{S}}^{\bullet}} ||\big(y_j(i_t^{\bullet}), y_{j+1}(i_t^{\bullet})\big) - \big(y_j(i_q^{\bullet}), y_{j+1}(i_q^{\bullet})\big)||.$$

Рисунок 3.2 демонстрирует, что результаты выделения периодов зависит от наклона оси. Для выбора наклона предлагается следующая эвристика: предлагается рассекать фазовую траекторию ее осью симметрии [80, 81].

3.2.2. Выбор пары главных компонент, отвечающих периодической составляющей временного ряда

Главные компоненты должны выбираться по некоторому формальному критерию. Используемая в работе процедура автоматического отбора основана на методе, предложенном в статье [78]. Идея этого метода состоит в том, чтобы обнаружить главные компоненты, соответствующие периодической составляющей с одинаковой частотой. Это делается путем сравнения спектральных плотностей главных компонент.

Спектральные плотности вычисляется с помощью быстрого преобразования Фурье временного ряда $\mathbf{s}$:

$$s(i) = \sum_{k=1}^{m} f(k)\mathrm{w}_m^{-(k-1)(i-1)}, \quad \mathrm{w}_m = \exp(-2\pi\mathrm{i}/m), \tag{3.8}$$

$$f(k) = \frac{1}{m}\sum_{i=1}^{m} s(i)\mathrm{w}_m^{-(k-1)(i-1)}.$$

Спектральная плотность содержит информацию о вкладах различных частот в сигнал. В качестве основного приближения фундаментальной частоты выберем минимальную ненулевую частоту $w_{\ min}$.

Опишем процедуру выбора пары основных компонент.

1. Формируется набор $\mathcal{Y}$ пар-кандидатов. Так как собственные значения главных компонент периодической пары почти равны, то главные компоненты пар-кандидатов имеют следующие индексы:

$$\mathcal{Y} = \{(\mathbf{y}_j, \mathbf{y}_{j+1}) | \lambda_j + 1 > 0,\ 1 \leq j \leq N.\}$$

2. Как было показано в разделе "Выбор основных компонент", периодическая пара главных компонент состоит из двух периодических функций, различающихся только по своей фазе. Спектральные плотности таких главных компонент имеют максимум амплитуды $|f(k)|$ в одних и тех же частотах $(k-1)/m$. Для каждой пары $(\mathbf{y}_j, \mathbf{y}_{j+1}) \in \mathcal{Y}$ проверяется, совпадают ли максимумы их спектральных плотностей.
Пусть k_j — номер частоты, соответствующий максимальной амплитуде $|f^j(k)|$ спектральной плотности компоненты $\mathbf{y}_j$

$$k_j = \underset{k \in \{0,\ldots,m-1\}}{\arg\max} |f^j(k)|.$$

Из-за зашумленности или других модуляций периодического сигнала спектральная плотность может стать размытой: спектральная мощность принимает ненулевые значения в частотах, близких к периодическим частотам, содержащимся в незашумленном сигнале. Поэтому точки максимума k_j и k_{j+1} могут не совпадать даже для пары компонент $(\mathbf{y}_j, \mathbf{y}_{j+1})$, отвечающих одной и той же частоте. Для определения пары компонент построим набор пар-кандидатов $\mathcal{Y}^*$, удовлетворяющих условие

$$|k_j - k_{j+1}| \leq \varepsilon M,$$

где M — количество измерений в секунду. Параметр ε контролирует приемлемый уровень размытия спектральной плотности. Следуя результатам вычислительных экспериментов из работы [78], в данном исследовании параметр ε был фиксирован равным $1/M$. Таким образом, множество $\mathcal{Y}^*$ пар-кандидатов:

$$\mathcal{Y}^* = \{(\mathbf{y}_j, \mathbf{y}_{j+1}) \in \mathcal{Y} |\ |k_j - k_{j+1}| \leq 1\}.$$

3. Из множества $\mathcal{Y}^*$ выбирается пара $(\mathbf{y}_j, \mathbf{y}_{j+1})$, для которой $(k-1)/m$ максимально близко к минимальной ненулевой частоте $w_{\min}$ периодической компоненты временного ряда X.

Глава 4
Оценка объема выборки

4.1. Оценка объема выборки в задачах классификации

Рассмотрим выборку вида $D_m = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, где $\mathbf{x}_i \in \mathbb{R}^n$ — описание i-го элемента выборки, а $y_i \in \mathcal{Y}$ — его метка класса. Введем обозначение $D_m = (\mathbf{X}, \mathbf{y})$, где $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$, $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_m^\intercal]^\intercal$ — матрица плана.

В задаче классификации необходимо построить отображение наблюдаемых данных $\mathbf{x} \in \mathbb{R}^m$ на множество меток класса $\mathcal{Y}$. Будем называть функцию $a : \mathbb{R}^m \to \mathcal{Y}$ *классификатором*. В данной работе рассматривается подход к построению классификатора, основанными на максимизации правдоподобия данных.

В данном разделе предполагается, если не указано иное, что выборка $(\mathbf{y}, \mathbf{X})$ простая, то есть объекты $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, m$ — случайные, независимые величины из одного распределения. Будем называть *гипотезой порождения данных* совокупность предположений о виде неизвестного распределения $\mathbb{P}(\mathbf{y}, \mathbf{X})$ элементов выборки D_m. В данной работе основное внимание уделяется случаю, когда предположения об истинном распределении данных формулируется с точностью до некоторого неизвестного параметра θ. Предполагается, что функция плотности распределения $p(\mathbf{y}|\mathbf{X})$ принадлежит параметрическому множеству функций

$$\mathcal{F} = \{f(\mathbf{y}, \mathbf{X}, \theta), \theta \in \Theta\},$$

то есть найдется такое значение параметра $\theta^* \in \Theta$, что

$$\mathbb{P}(\mathbf{y}, \mathbf{X}) = f(\mathbf{y}, \mathbf{X}, \theta^*).$$

В данных обозначениях предположения о независимости и однородности выборки означают, что каждая функция $f \in \mathcal{F}$ представима в виде произведения

$$f(\mathbf{y}, \mathbf{X}, \theta) = \prod_{i=1}^m g(y_i, \mathbf{x}_i, \theta)$$

по элементам выборки.

Пусть фиксировано некоторое значение параметра $\theta \in \Theta$. Тогда классификатор $a(\mathbf{x})$ имеет вид

$$a(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} (y, \mathbf{x}, \theta).$$

Таким образом, для решения задачи классификации в данной постановке необходимо оценить значение параметра θ, наиболее согласованного с наблюдаемыми данными.

Правдоподобием данных называется функция $p(\mathbf{y}, \mathbf{X}|\theta)$ параметра θ. Оценка максимального правдоподобия параметра θ

$$\hat{\theta} = \arg\max_{\theta} p(\mathbf{y}, \mathbf{X}|\theta)$$

состоятельна и асимптотически нормальна, и в пределе по $m \to \infty$ не превосходит по среднеквадратичному отклонению любую другую состоятельную оценку $\tilde{\theta}$ параметра θ:

$$\sqrt{\sum_{j=1}^{n}\left(\hat{\theta}_j - \theta_j\right)^2} \leq \sqrt{\sum_{j=1}^{n}\left(\tilde{\theta}_j - \theta_j\right)^2}.$$

Оптимальность в данном смысле оценок максимального правдоподобия обеспечивает их широкую применимость. Однако, так как в практических задачах объем выборки конечен, оценка максимального правдоподобия не всегда является оптимальной. Так как непосредственная максимизация качества классификации на обучающей выборке по параметру θ ведет к к переобучению, возникает необходимость ограничить множество параметров Θ. Оценка максимальной апостериорной вероятности

$$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{y}, X) = \arg\max_{\theta} p(\mathbf{y}, X|\theta) + p(\theta) \tag{4.1}$$

обеспечивает регуляризацию параметров за счет введения дополнительных предположений об их априорном распределении $p(\theta)$.

Определение 13. Совместное распределение $p(\mathbf{y}, \mathbf{X}, \theta) = p(\mathbf{y}, \mathbf{X}|\theta)p(\theta)$ будем называть *вероятностной моделью.*

Задача оценки объема выборки может возникать в трех различных постановках, в зависимости от типа решаемой прикладной задачи.

1. Задача оценки объема выборки при отсутствии наблюдений решается при планировании эксперимента до сбора каких-либо данных. В этом случае применяются классические статистические методы, в которых формула для оценки объема выборки задается в явном виде при фиксированных параметрах проверяемых гипотез. При наличии сложных гипотез порождения данных, например, в байесовской постановке, для оценки объем выборки применяется сэмплирование. Для сэмплированных выборок различного размера проверяется выполнение критерия, определяющего метод. Выбирается минимальный объем выборки, удовлетворяющий заданному критерию. Цель оценки объемы выборки в данном случае — ограничение объема выборки снизу.
2. При наличии потенциально неограниченного объема данных необходимо получить оценку сверху на объем выборки. В данном случае более эффективно применение методов, учитывающих наблюдаемые данные. Решается задача определения объема выборки, достаточного для выполнения интересующего исследователя критерия.
3. Задача оценки объема выборки с учетом выборки данных недостаточного объема. Данный случай является обобщением предыдущих двух. Наличие измерений позволяет скорректировать предположения о распределении данных, задать параметры методов, проверяемых гипотез. Так как число

измеренных объектов может быть недостаточным с точки зрения выбранного критерия (здесь используются методы второго типа), требуется оценить число дополнительных измерений, которые необходимо произвести (здесь используются методы первого типа).

В данной работе основное внимание уделяется байесовским методам оценки объема выборки, так как в байесовской постановке нет необходимости различать перечисленные случаи. 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

$$\int L(\mathbf{y}, \mathbf{X}) \prod_{i=1}^{m} \mathbb{P}(y_i, \mathbf{x}_i) \prod_{i=1}^{m} dy_i \dots d\mathbf{x}_i \leq \xi \tag{4.2}$$

Так как истинное порождающее распределение $\mathbb{P}(y_i, \mathbf{x}_i)$ неизвестно, используется выборочная оценка математического ожидания 4.2:

Определение 14. Назовем объем m^* выборки из неизвестного распределения $\mathbb{P}(y, \mathbf{x})$ *достаточным* согласно критерию T, если T выполняется для всех $m \geq m^*$.

4.2. Непараметрический метод оценки объема выборки

Задан набор временных рядов $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_S\}$, где каждый ряд $\mathbf{x}_j = \{x_j(i) \in \mathbb{R}\}_{i=1}^{m_j}$ — это последовательность реализаций некоторого стационарного случайного процесса. Требуется кластеризовать набор $\mathbf{X}$

$$\mathbf{X} = \bigsqcup_{k=1}^{K} \mathbf{X}_k, \quad \mathbf{X}_k = \{\mathbf{x}_j,\, j \in \mathcal{A}_k\}, \quad \{1, \dots S\} = \bigsqcup_{k=1}^{K} \mathcal{A}_k, \tag{4.3}$$

разбив набор $\mathbf{X}$ на K наборов $\mathbf{X}_k$ временных рядов, таких что все ряды в $\mathbf{X}_k$ принадлежат одному и тому же распределению. Здесь $\mathcal{A}_k$ — множество индексов временных рядов k-го кластера.

В силу стационарности случайного процесса $\mathbf{x}$, пренебрежем последовательностью значений ряда $\mathbf{x}$. Представим временной ряд $\mathbf{x}$ как выборку X реализаций некоторой случайной величины с распределением P

$$X = \{x \in \mathbb{R} | \text{ для некоторого } i \in \{1, \dots, m\} : x(i) = x\}. \tag{4.4}$$

Для решения задачи об агрегировании временных рядов $\mathbf{x}$ и $\mathbf{x}'$ будем сравнивать гистограммы, построенные по выборкам X и X', сопоставленным каждому из рядов в соответствии с (4.4). Опишем подробнее процедуру построения гистограммы.

Пусть объем выборок X и X' равен m и m' соответственно. Разобьем область значений случайной величины из P на N промежутков $[a_i, a_{i+1}]$, и обозначим $p_i = P(a_i < x \leq a_{i+1})$ вероятность случайной величине с распределением P принять значение из i-го промежутка; n_i и n'_i — количество объектов выборок X и X', попавших в i-тый промежуток. Обозначим $\hat{P}_m$ гистограмму, построенную по выборке X объема m из распределения P. Гистограмма $\hat{P}_m$ задается набором оценок:

$$\hat{P}_m(a_i < x \leq a_{i+1}) = \frac{n_i}{m} = \hat{p}_i, \quad i = 1, \ldots, N-1. \tag{4.5}$$

вероятности p_i.

Для решения задачи кластеризации (4.3) воспользуемся алгоритмом нахождения кратчайшего незамкнутого пути между временными рядами. Результатом применения алгоритма является минимальное остовное дерево: граф с $n-1$ ребрами, покрывающий все n вершин, ребра которого обладают минимальной суммарной длиной. Удалив из минимального остовного дерева $K-1$ самых длинных ребер, получим кластеризацию вершин графа на K кластеров $\mathbf{X}_k$. Вершинами графа являются исследуемые временные ряды $\mathbf{x}_j$; длина ребра, соединяющего две вершины, равна расстоянию между соответствующими временными рядами. Найдя расстояние между всеми парами рядов, получим матрицу парных расстояний D. В качестве расстояний между рядами $\mathbf{x}_r$ и $\mathbf{x}_s$ будем использовать симметризованное расстояние Кульбака-Лейблера между гистограммами $\hat{P}_{m_r}$ и $\hat{P}_{m_s}$, построенными по временным рядам:

$$D(r,s) = \frac{2m_r m_s}{m_r + m_s}\left(D_{\mathrm{KL}}(\hat{P}_{m_r}||\hat{P}_{m_s}) + D_{\mathrm{KL}}(\hat{P}_{m_s}||\hat{P}_{m_r})\right). \tag{4.6}$$

Первый сомножитель в правой части снимает зависимость от объема выборки.

Конечной целью кластеризации временных рядов с учетом расстояний между ними является повышение согласованности агрегированных прогнозов. Для оценки качества кластеризации будем рассматривать несогласованность

$$\delta(i) = \left|\sum_k^K \hat{\mathbf{X}}_k(i) - \sum_j^n \hat{\mathbf{x}}_j(i)\right| \tag{4.7}$$

при прогнозировании по наборам $\mathbf{X}_k$ временных рядов и отдельным временным рядам $\mathbf{x}_j$. Здесь $\hat{\mathbf{X}}_k(i)$ — прогноз агрегированного ряда в момент i, $\hat{x}_j(i)$ — прогноз j-го ряда в момент времени i. Чем меньше несогласованность, тем качественнее выполнена кластеризация. Очевидно, что наименьшее значение $\delta_{\min} = 0$ выражения (4.7) достигается при $K = S$, поэтому предлагается огра-

ничить число K или ввести в (4.7) штраф $h(K)$ за его повышение:

$$K = \arg\min_K \left(\sum_{i=1}^{m} \delta(i) + h(K) \right).$$

В данной работе ограничимся рассмотрением предложенного критерия принадлежности временных рядов к одному распределению и кластеризации временных рядов на основе расстояния между ними.

4.2.1. Статистическая значимость расстояния Кульбака-Лейблера

Чтобы показать, что результаты кластеризации временных рядов на основе расстояния Кульбака-Лейблера между ними статистически значимы, необходимо исследовать распределение расстояния Кульбака-Лейблера между гистограммами, построенными по выборкам X и X' из одного распределения P. В данном разделе будет показано, что хотя расстояние Кульбака-Лейблера не имеет предельного распределения, для него можно получить предельные оценки сверху.

Пусть пока выборки X и X' имеют одинаковый объем m. Рассмотрим расстояние $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m)$ между гистограммами $\hat{P}_m$ и $\hat{P}'_m$. По определению, расстояние Кульбака-Лейблера $D_{\mathrm{KL}}(Q||P)$ между распределениями Q и P равно

$$D_{\mathrm{KL}}(Q||P) = \int P \cdot f\left(\frac{Q}{P}\right), \tag{4.8}$$

где $f(t) = t \ln t$. Функция f строго выпукла и дважды дифференцируема в единице, и, повторяя рассуждения из [82], разложим подынтегральное выражение из правой части (4.8) по f в окрестности единицы:

$$P(x) f\left(\frac{Q(x)}{P(x)}\right) = f(1) + f'(1)(Q(x) - P(x)) + \frac{f''(1)}{2} \frac{(Q(x) - P(x))^2}{P(x)}$$

$$+ P(x) o\left(\left(\frac{Q(x)}{P(x)} - 1\right)^3\right),$$

где $f(1) = 0$, $f''(1) = 1$. Подставив вместо Q распределение $\hat{P}_m$, определяемое (4.5), и просуммировав по i, получим соотношение

$$D_{\mathrm{KL}}(\hat{P}_m||P) = \sum_{i=1}^{N} p_i f\left(\frac{\hat{p}_i}{p_i}\right) =$$

$$= \frac{1}{2} \sum_{i}^{N} \frac{(\hat{p}_i - p_i)^2}{p_i} + \sum_{i=1}^{N} p_i \cdot \varepsilon\left(\left(\frac{\hat{p}_i}{p_i} - 1\right)^3\right) \sim \frac{1}{2m} \sum_{i}^{N} \frac{(n_i - mp_i)^2}{mp_i}.$$

и следующий предельный переход:

$$2m \cdot D_{\mathrm{KL}}(\hat{P}_m||Q) \sim m \sum_{i=1}^{N} \frac{(\hat{p}_i - p_i)^2}{p_i)} = \sum_{i=1}^{N} \frac{(n_i - mp_i)^2}{mp_i} \to \chi^2_N \text{ при } m \to \infty. \tag{4.9}$$

Докажем следующую теорему:

Теорема 5. Случайная величина $2m \cdot D_{\mathrm{KL}}(Q||\hat{P}_m) \to \chi^2_N$ по распределению при $m \to \infty$.

Доказательство. Аналогично доказательству предельного перехода (4.9), разложим $D_{\mathrm{KL}}(Q||\hat{P}_m)$ по степеням $f(t)$ вблизи единицы и получим

$$D_{\mathrm{KL}}(Q||\hat{P}_m) \sim \frac{1}{2} \sum_{i=1}^{N} \frac{(\hat{P}_m(\xi_i) - Q(\xi_i))^2}{\hat{P}_m(\xi)} = \frac{1}{2m} \sum_{i=1}^{N} \frac{(n_i - mp_i)^2}{n_i}.$$

Пусть $G_m(x)$ — функция распределения случайной величины $\sum_{i=1}^{N} \frac{(n_i - mp_i)^2}{mp_i}$, $F_m(x)$ — случайной величины $\sum_{i=1}^{N} \frac{(n_i - mp_i)^2}{n_i}$. Так как $G_m(x)$ сходится поточечно к $F_{\chi^2_{N-1}}$ при $m \to \infty$, имеем

$$|G_m(x) - F_{\chi^2_{N-1}}| < \frac{\epsilon}{2} \quad \forall m > m'.$$

Покажем, $|G_m(x) - F_m(x)| \to 0$ при $m \to \infty$. Для этого покажем, что $\forall \varepsilon > 0$ найдется объем выборки m_0 такой, что для всех $m > m_0$ выполняется

$$\mathrm{P}\left(\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \le \frac{\varepsilon}{N}\right) > 1 - \varepsilon. \tag{4.10}$$

Согласно центральной предельной теореме,

$$\frac{n_i - mp_i}{p_i(1 - p_i)\sqrt{m}} \to \mathcal{N}(0, 1) \text{ по распределению при } m \to \infty,$$

причем для скорости сходимости имеет место неравенство Берри-Эссена:

$$|Q_m(x) - \Phi(x)| \le \frac{A}{\sqrt{m}},$$

где $Q_m(x)$ — функция распределения величины $\frac{n_i - mp_i}{p_i(1-p_i)\sqrt{m}}$, $\Phi(x)$ — функция стандартного нормального распределения, A — некоторая константа. Тогда вероятность

$$\mathrm{P}\left(\left|\frac{n_i - mp_i}{p_i(1 - p_i)\sqrt{m}}\right| < C\right) = Q_m(C) - Q_m(-C) \ge 2\Phi(C) - 1 - \frac{2A}{\sqrt{m}}. \tag{4.11}$$

Пусть, кроме того, выполняется $0 < 1 - p \le p_i \le p < 1$. Тогда, с вероятностью $\mathrm{P}_C \ge 2\Phi(C) - 1$ выполняется

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| = \frac{|n_i - mp_i|^3}{mn_ip_i} \le \frac{C^3(1-p_i)^3p_i^2}{n_i}\sqrt{m} \le \frac{C^3(1-p)^3p}{\sqrt{m} - C(1-p)}.$$

Обозначим $m_1 = [4C^2(1-p)^2]$, тогда при $m > m_1$ имеет место $\sqrt{m} - C(1-p_i) > \frac{1}{2}\sqrt{m}$ и

$$\left|\frac{(n_i - mp_i)^2}{n_i} - \frac{(n_i - mp_i)^2}{mp_i}\right| \le \frac{2C^3(1-p_i)^3p_i}{\sqrt{m}}.$$

Тогда, для фиксированного ε определим

$$C_\varepsilon = \frac{\epsilon^{1/3}m^{1/6}}{(1-p_i)(2p_iN)^{1/3}},\ \mathrm{P}_m(\varepsilon) = 2\Phi(C_\varepsilon) - 1 - \frac{2A}{\sqrt{m}}.$$

При заданном ε вероятность $\mathrm{P}_m(\varepsilon) \to 1$ при $m \to \infty$, поэтому найдется m_2 такое что для любого $m > m_2$ выполнено $\mathrm{P}_m(\varepsilon) > 1 - \varepsilon$. Выбрав $m_0 = \max(m_1, m_2)$, получим утверждение (4.10). Тогда

$$\left|\sum_{i=1}^{N}\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| \le \sum_{i=1}^{N}\left|\frac{(n_i - mp)^2}{n_i} - \frac{(n_i - mp)^2}{mp}\right| < \varepsilon \quad \text{при } m > m_0.$$

Из только что доказанного следует, что $|F_m(x) - G_m(x)| \to 0$ при $m \to 0$. Тогда, $\forall\epsilon > 0\ \exists m'' :$ при $m > m''$ выполняется

$$|F_m(x) - F_{\chi^2_{N-1}}| < |F_m(x) - G_m(x)| + |G_m(x) - F_{\chi^2_{N-1}}| < \frac{\epsilon}{2} + \frac{\epsilon}{2}.$$

□

Доказанная теорема и утверждение (4.9) позволяют получить оценки распределения случайных величин $2m \cdot D_{\mathrm{KL}}(\hat{P}_m||Q)$ и $2m \cdot D_{\mathrm{KL}}(Q||\hat{P}_m)$ при больших m. Для решения задачи кластеризации (4.3) нам потребуется также исследовать поведение расстояния Кульбака-Лейблера $D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l)$ между гистограммами, построенными по выборкам X, X' различных длин m и l. Воспользовавшись неравенством треугольника

$$D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \le D_{\mathrm{KL}}(\hat{P}_m||Q) + D_{\mathrm{KL}}(Q||\hat{P}_l),$$

получим следствия из теоремы 5:

Следствие 1: $2m \cdot D_{\mathrm{KL}}(\hat{P}_m||\hat{P}'_m) \le \chi^2_{2N}$ в пределе при $m \to \infty$.

Следствие 2: Пусть выборки X, X' растут таким образом, что $m/l \to \rho$, $0 < \rho < \infty$. Тогда

$$2\frac{ml}{m+l} \cdot D_{\mathrm{KL}}(\hat{P}_m||\hat{P}_l) \le \chi^2_{2N}$$

в пределе при $m, l \to \infty$.

Доказательство. Действительно, при выполнении условия $m/l \to \rho$, $0 < \rho < \infty$, имеем

$$\frac{l}{m+l} \to \frac{1}{1+\rho}, \quad \frac{m}{m+l} \to \frac{\rho}{1+\rho}$$

и

$$2\frac{ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m || \hat{P}_l) \leq \frac{l}{m+l} 2m D_{\mathrm{KL}}(\hat{P}_m || Q) + \frac{m}{m+l} 2l D_{\mathrm{KL}}(Q || \hat{P}_l) \to \chi^2_{2N}.$$

□

Обозначим величину $\frac{2ml}{m+l} D_{\mathrm{KL}}(\hat{P}_m || \hat{P}'_m)$ через $\xi_{m,l}$. Следствие 2 дает верхнюю оценку поведения случайно величины $\xi_{m,l}$ при больших m и l, а именно: пусть $\eta \sim \chi^2_{2N}$, тогда при достаточно больших m и l выполнено для любого элементарного исхода w из вероятностного пространства Ω выполнено $\xi_{m,l}(w) < \eta(w)$. Следовательно, для любого $x \in \mathbb{R}$ верно

$$\mathrm{P}(\xi_{m,l} < x) \geq \mathrm{P}(\eta < x). \tag{4.12}$$

В следующем разделе покажем, как этот факт будет использоваться для проверки принадлежности временных рядов одному распределению.

4.2.2. Проверка принадлежности временных рядов к одному распределению

Для решения задачи агрегирования временных рядов $\mathbf{x}$ и $\mathbf{x}'$ необходимо уметь принимать решение о принадлежности временных рядов одному распределению. Опишем процедуру проверки гипотезы о принадлежности выборок X и X', составленных (4.4) из временных рядов $\mathbf{x}$ и $\mathbf{x}'$. Пусть нулевая гипотеза H_0 состоит в принадлежности выборок X и X' к одному распределению:

$$H_0 : P(x) = P'(x).$$

Сформулируем критерий проверки гипотезы H_0 при альтернативе $H_1 : P(x) \neq P'(x)$. Для этого определим критическую область $U(\alpha)$ для статистики $t_{m,l}$ с уровнем значимости α:

$$U(\alpha) = \{t : \bar{t}_{1-\alpha} > t \text{ или } t > \bar{t}_\alpha\},$$

где критическое значение $\bar{t}_\alpha$ определяется соотношением

$$\mathrm{P}(t > \bar{t}_\alpha | H_0) = \alpha. \tag{4.13}$$

Так как предельное распределение величины $\xi_{m,l}$ неизвестно, будем использовать критическую область, задаваемую распределением χ^2_{2N}. Будем говорить,

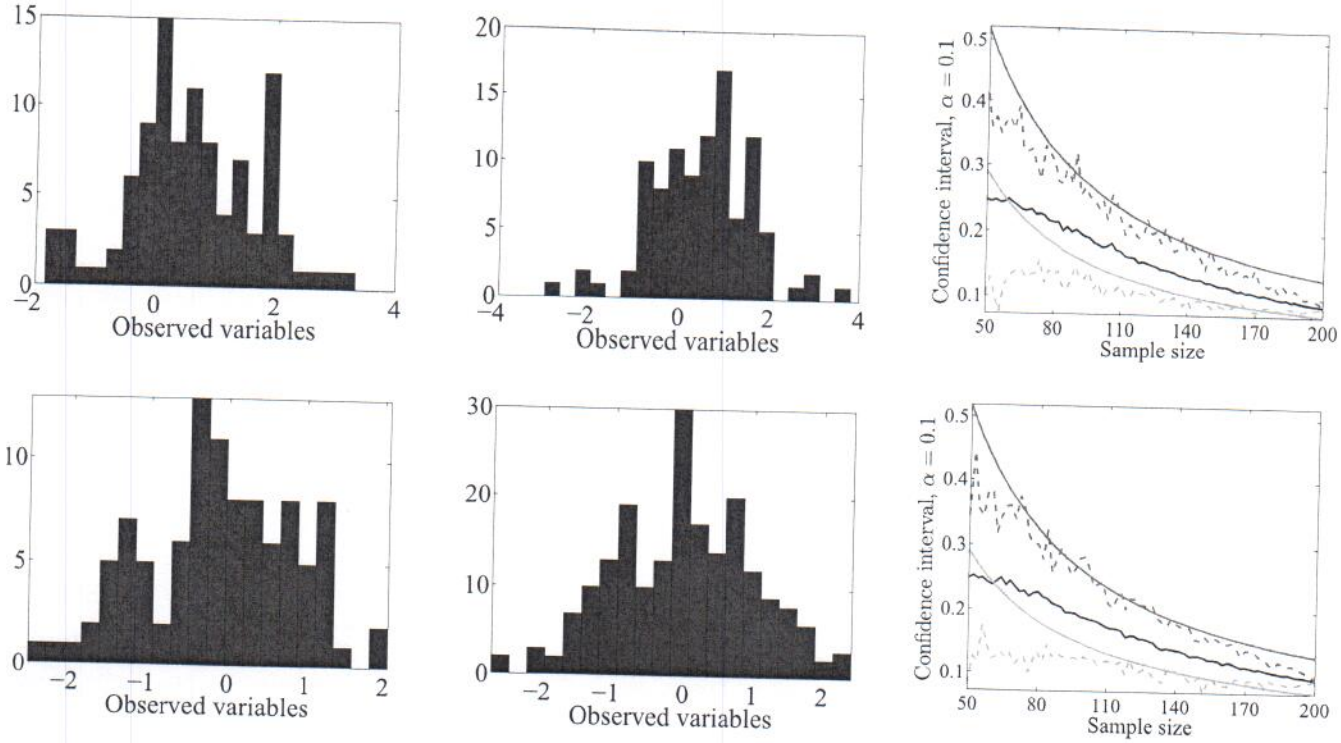


Рис. 4.1. Гистограммы, построенные по двум выборкам из нормального распределения и зависимость статистики t_m от объема выборки. Красным и зеленым отмечены границы доверительного интервала для t_m при $\alpha = 0.1$. На рисунках снизу выборки были зашумлены.

что данные отвергают гипотезу H_0 в случае, если статистика $t_{m,l}$ принадлежит критической области U^{χ^2}

$$U^{\chi^2}(\alpha) = \{t : \bar{t}^{\chi^2}_{1-\alpha} > t \text{ или } t > \bar{t}^{\chi^2}_{\alpha}\}, \tag{4.14}$$

где $\bar{t}^{\chi^2}$ — критическое значение величины $\chi^2_{2_N}$:

$$\mathrm{P}(t > \bar{t}^{\chi^2}_{\alpha} | t \sim \chi^2_{2_N}) = \alpha.$$

Из неравенства (4.12) и определения (4.13) критических значений следует, что критические области U и U^{χ^2} несравнимы, то есть

$$\bar{t}_{1-\alpha} < \bar{t}^{\chi^2}_{1-\alpha}, \quad \bar{t}_{\alpha} < \bar{t}^{\chi^2}_{\alpha}.$$

Это означает, что возможны следующие ситуации:

1. Случай $\bar{t}^{\chi^2}_{1-\alpha} < t_{m,l} < \bar{t}_{\alpha}$, когда статистика $t_{m,l}$ одновременно принадлежит истиной, но неизвестной критической области U, и вычислимой критической области U^{χ^2}.
2. Случай $\bar{t}_{1-\alpha} < t_{m,l} < \bar{t}^{\chi^2}_{1-\alpha}$, когда статистика $t_{m,l}$ принадлежит истиной, но неизвестной критической области U, и не принадлежит U^{χ^2}. Так как $t_{m,l} \in$

U, то с высокой вероятностью гипотеза H_0 неверна, и мы рискуем принять неверное решение об истинности гипотезы H_0. То есть зазор между $\bar{t}_{1-\alpha}$ и $\bar{t}^{\chi^2}_{1-\alpha}$ повышает вероятность ошибки второго рода.

3. Случай $\bar{t}_\alpha < t_{m,l} < \bar{t}^{\chi^2}_\alpha$, когда статистика $t_{m,l}$ попадает в U^{χ^2}, хотя на самом деле $t_{m,l}$ не принадлежит U. В этом случае велика вероятность, что H_0 верна, но решение будет принято в пользу H_1. Таким образом, зазор между $\bar{t}_\alpha$ и $\bar{t}^{\chi^2}_\alpha$ повышает вероятность ошибки первого рода.

Второй случай разрешается следующим образом: использование симметризованного расстояния позволяет перейти от двусторонних критериев U и U^{χ^2} вида (4.14) к односторонним критериям

$$U_1(\alpha) = \{t : t > \bar{t}_\alpha\}, \quad U_1^{\chi^2}(\alpha) = \{t : t > \bar{t}^{\chi^2}_\alpha\}.$$

В этом случае $U_1^{\chi^2} \subseteq U_1$ и справедливо следствие $t_{m,l} \in U_1^{\chi^2} \Rightarrow t_{m,l} \in U_1$. Кроме того, далее будет показано (теорема 4.2.2.), что при увеличении объема выборки m вероятность отклонить гипотезу H_0 с помощью критерия (4.14) в случае, если гипотеза H_0 неверна, стремится к единице. Влияние третьего случая на возможность применения критерия (4.14) для принятия нулевой гипотезы исследуется экспериментально. Эксперименты, приведенные ниже и в разделе «Вычислительный эксперимент» показывают, что при истинности нулевой гипотезы области U и U^{χ^2} достаточно близки для принятия верного решения.

Пример применения критерия (4.14) **при истинности** H_0**.** На рисунке 4.2.1. изображены гистограммы из для двух выборок из

- стандартного нормального распределения 4.2.1. (сверху)
- и стандартного нормального распределения с шумом $\varepsilon \sim 0.1 \cdot R[0,1]$ 4.2.1. (снизу),

а также зависимость расстояния Кульбака-Лейблера между гистограммами между выборками одинакового объема от объема выборки m и область допустимых значений с точки зрения критерия (4.14). Здесь вместо критических значений $\bar{t}^{\chi^2}_{1-\alpha}$, $\bar{t}^{\chi^2}_\alpha$ и $t_m = 2m \cdot D_{\mathrm{KL}}$ отложены величины $\bar{t}^{\chi^2}_{1-\alpha}/2m$, $\bar{t}^{\chi^2}_\alpha/2m$ и $t_m/2m$, чтобы продемонстрировать масштаб расстояния Кульбака-Лейблера и наличие сходимости. Рисунки показывают, что в данном случае использование распределения χ^2_{2N} в качестве оценки предельного распределения статистики t_m позволяет принять верное решение о принадлежности рядов к одному распределению. Пунктирная линия показывает границу области, в которую вошли $1 - \alpha = 90\%$ выборки и задает оценку критической области для $D_{\mathrm{KL}}(\hat{P}_m || \hat{P}'_m)$. Более подробно результаты описаны в разделе «Вычислительный эксперимент».

Покажем теперь, что критерий (4.14) так же можно использовать для отвержения гипотезы H_0.

Теорема 6. Критерий (4.14) состоятелен:

$$\lim_{m\to\infty} \mathrm{P}(t_m \in U | H_1) = 1,$$

Пусть $p_i > p'_i$. Согласно (4.11), при больших m с вероятностью $\mathrm{P} > (2\Phi(C_1) - 1)(2\Phi(C_2) - 1)$ выполнено

$$|n_i - mp_i| < C_1\sqrt{m}, \quad |n'_i - mp'_i| < C_2\sqrt{m}.$$

Для любого $\varepsilon > 0$ найдется константа $C_\varepsilon : \mathrm{P} > (2\Phi(C_\varepsilon) - 1)^2 > 1 - \varepsilon$. Выберем $C_1 = C_2 = C_\varepsilon$. Тогда $(n_i - n'_i) > m(p_i - p'_i) + O(\sqrt{m})$ и

$$\frac{(n_i - n'_i)^2}{n_i} > m\frac{(p_i - p'_i)^2}{p_i} + O(\sqrt{m}) > Cm. \tag{4.15}$$

Следовательно, для любого $\alpha \in (0, 1)$ при достаточно больших m

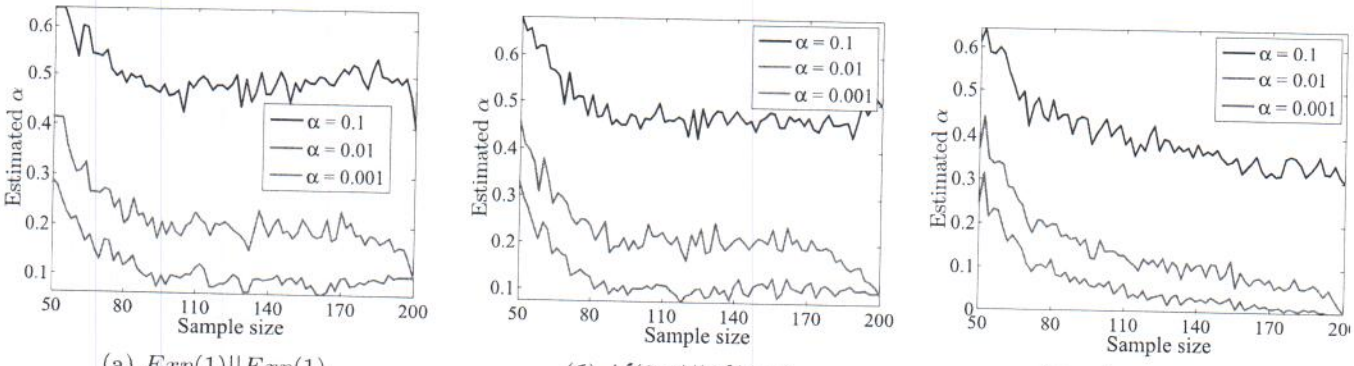
$$t_m = 2mD_{\mathrm{KL}}(\hat{P}^1_m || \hat{P}^2_m) \sim \sum_{i=1} \frac{(n_i - n'_i)^2}{n_i} > Cm > \bar{t}_\alpha$$

с вероятностью $\mathrm{P} > 1 - \epsilon$, то есть вероятность $\mathrm{P}(t_m > \bar{t}_\alpha) \to 1$ при $m \to \infty$. □

4.2.3. Анализ свойств предложенного критерия достаточного объема выборки

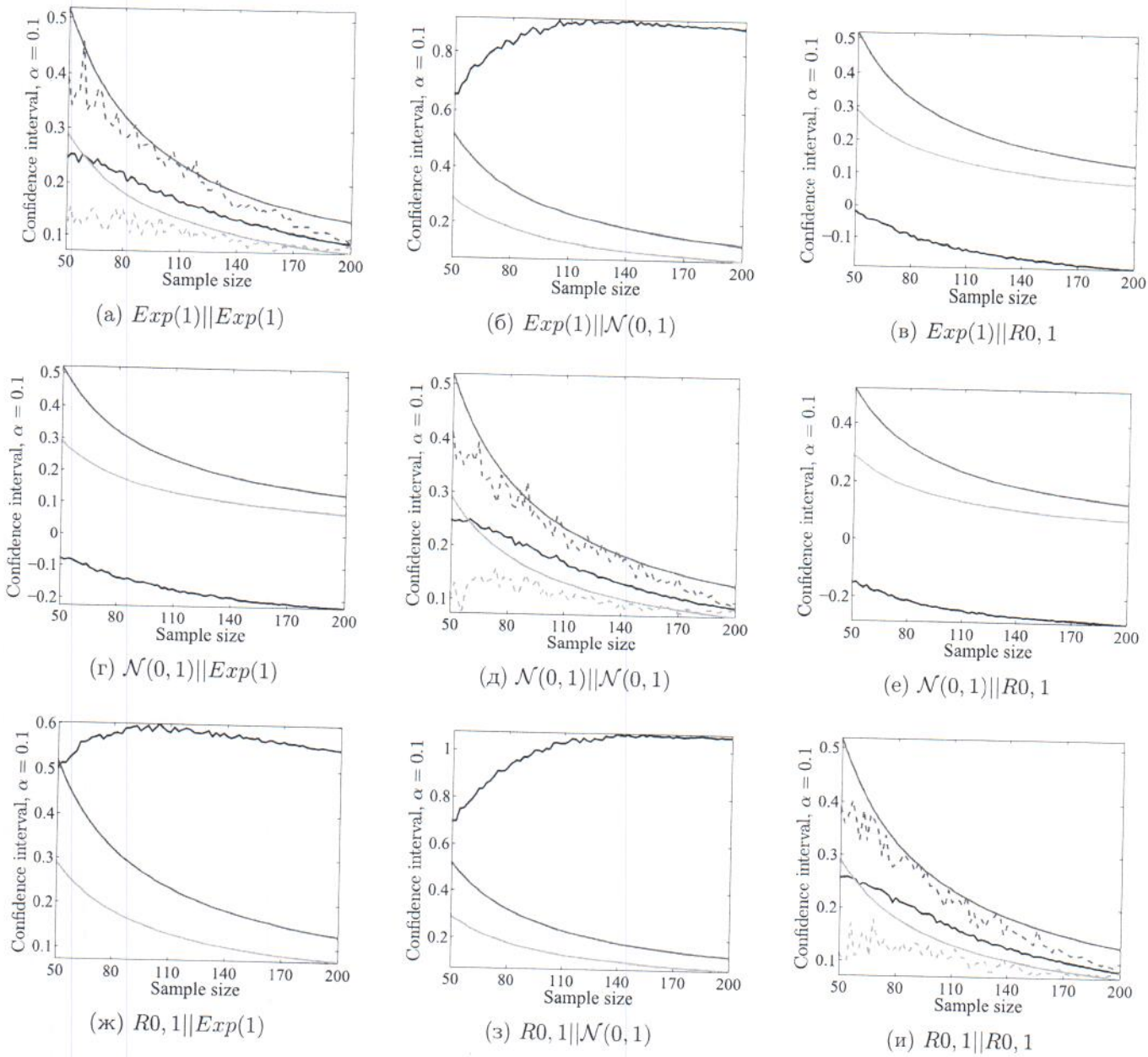
Работа критерия была рассмотрена на различных парах распределений. Для выбранной пары распределений повторяется следующая процедура:

1) генерируются выборки X и X' одинакового объема m;
2) по выборкам строятся гистограммы $\hat{P}_m$ и $\hat{P'}_m$ с фиксированным числом разбиений $N = 20$ и вычисляется расстояния Кульбака-Лейблера $D_{\mathrm{KL}}(\hat{P}_m || \hat{P'}_m)$;
3) расстояния усредняются по 1000 генерациям выборок
4) объем m выборки увеличивается.



(а) $Exp(1)||Exp(1)$ (б) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$ (в) $R(0,1)||R(0,1)$

Рис. 4.3. Зависимость фактического уровня значимости от объема выборки различных уровнях значимости критерия χ^2_{2N}.



(а) $Exp(1)||Exp(1)$ (б) $Exp(1)||\mathcal{N}(0,1)$ (в) $Exp(1)||R0,1$

(г) $\mathcal{N}(0,1)||Exp(1)$ (д) $\mathcal{N}(0,1)||\mathcal{N}(0,1)$ (е) $\mathcal{N}(0,1)||R0,1$

(ж) $R0,1||Exp(1)$ (з) $R0,1||\mathcal{N}(0,1)$ (и) $R0,1||R0,1$

Рис. 4.2. Зависимость статистики t_m от объема выборки для различных пар распределений. Красным и зеленым отмечены границы доверительного интервала для t_m при $\alpha = 0.1$.

то есть вероятность отвергнуть гипотезу H_0, если распределения временных рядов X и X' не совпадают, с увеличением выборки стремится к единице.

Доказательство. Пусть функции распределения P и P' временных рядов не совпадают. Тогда найдется $x^* \in \mathbb{R}$, в котором значения этих функций различны $P(x^*) \neq P'(x^*)$. Тогда найдется такой способ разбиения пространства $\mathbb{R}$, что для некоторого i вероятность попадания в i-тый промежуток не одинакова для рассматриваемых случайных величин:

$$P(a_i < x \leq a_{i+1}) = p_i \neq p'_i = P'(a_i < x \leq a_{i+1}).$$

На рисунке 4.2.2. отложены расстояния $D_{\mathrm{KL}}(\hat{P}_m|\hat{P'}_m)$ в зависимости от объема выборки и критические значения $\bar{t}^{\chi^2}/2m$ при заданном уровне значимости α. Заметим, что в случае различных распределений расстояние $D_{\mathrm{KL}}(\hat{P}_m|\hat{P'}_m)$ быстро попадает в критическую область, и характер его зависимости от m согласуется с оценкой (4.15). Отрицательные значения, нехарактерные для расстояния Кульбака-Лейблера, возникают при численном приближении интеграла (4.8), когда распределение в знаменателе под знаком логарифма имеет большую область определения. Именно из-за отрицательных значений был использован двусторонний критерий. Для симметризованного расстояния Кульбака-Лейблера используется односторонний критерий. В случае с выборками из одного распределения, на рисунке также отложены пунктиром оценки критических значений $\bar{t}$, полученные экспериментально. Видно, что, несмотря на то критические области U^{χ^2} и U не совпадают, при истинности гипотезы H_0 зависимость не попадает ни в U^{χ^2}, ни в U. **Оценка фактического значения** α. Чтобы

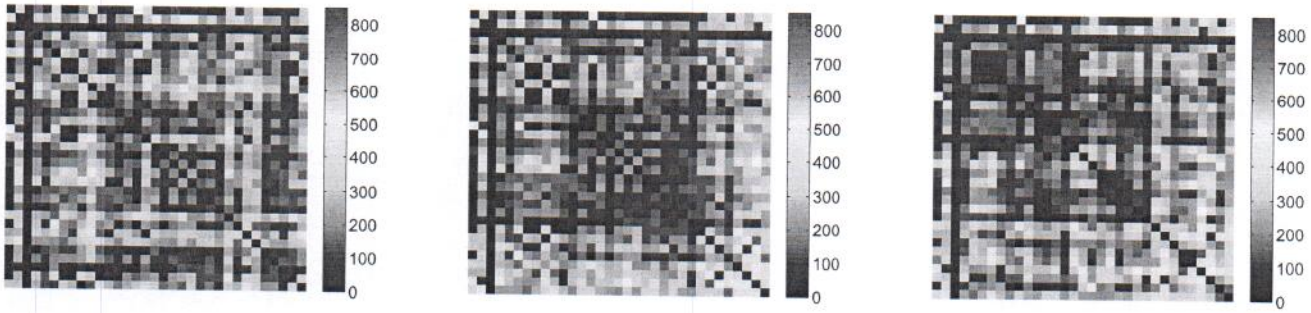


Рис. 4.4. Симметризованные матрицы попарных расстояний Кульбака-Лейблера между временными рядами для различных групп грузов: до кластеризации и после для $K = 5$ и $K = 10$ кластеров.

оценить реальный уровень значимости решения о принятии или отвержении гипотезы H_0, подсчитаем среднюю долю объектов выборки, попавших в U^{χ^2} при заданном α. Результаты отражены на рисунке 4.3. Из рисунков следует, что для достижения уровня значимости $\alpha = 0.1$ нужно использовать в качестве оценки U критическую область U^{χ^2} с уровнем значимости 0.001.

4.3. Активный выбор объектов

Пусть даны размеченная выборка $L_m = \{\mathbf{z}_i\}_{i=1}^m = \{(y_i, \mathbf{x}_i)\}_{i=1}^m$ и неразмеченная выборка $U_m = \{(\mathbf{x}_i)\}_{i=m+1}^M$, причем $\mathbf{x} = (\mathbf{x}^{(1)}|\ldots|\mathbf{x}^{(L)})$, $L \geq 1$.

В задаче активного обучения предполагается, что неразмеченные объекты $\mathbf{x}_i \in U$ доступны в большом количестве, в то время как разметка объектов является затратной операцией. При этом объекты неравнозначны с точки зрения полезности для повышения качества классификации: предполагается, что раз-

метка одних объектов дает больше информации, чем разметка других объектов. В связи с этим в литературе предлагаются различные

Определим выборку $D_m = L_m \cup U_m$. Пусть $\ell(\mathbf{x}|D_m)$ — ожидаемые потери при измерении объекта $\mathbf{x}$

$$\ell(\mathbf{x}|D_m) = \int p(y|\mathbf{x}, D_m)c(y, \mathbf{x})dy, \quad c(y, \mathbf{x}) = \sum_{l=1}^{L} c\left(y, \mathbf{x}^{(l)}\right).$$

Здесь $c(y, \mathbf{x})$ — стоимость измерений, учитывающая как затраты на изменение объекта, так и его предполагаемую полезность. В данной работе при оценке объема выборки по набору измеренных данных требуется ответить на вопрос, проводить ли дополнительные измерения? Измерения проводятся, пока ожидаемые потери

$$\int p(\mathbf{x}|D_m)\ell(\mathbf{x})\mathbf{dx}$$

существенно меньше текущих

$$\ell(D_m) = \sum_{i=1}^{m} c(y_i, \mathbf{x}_i) + \sum_{i=m+1}^{M} \int p(y|\mathbf{x}_i, D_m)c(y, \mathbf{x}_i)dy.$$

Пусть на основе размеченных L и неразмеченных U данных построена модель f_m. Новый объект для разметки выбирается согласно правилу

$$\mathbf{x} = \arg \max_{\mathbf{x}_i,\, i \in \{1,\ldots,M\} \setminus \mathcal{L}} Q(\mathbf{x}_i, f_m),$$

после чего модель уточняется $f_m \to f_{m+1}$ с использованием обновленной выборки $L_{m+1} = L_m \cup \{\mathbf{x}, f^*(\mathbf{x})\}$ и выбор объектов продолжается пока не будет выбрано достаточно объектов или достигнут некий критерий останова $\tilde{Q}(U, L, f_m) < \epsilon$. Рассмотрены следующие стратегии выбора объектов:

- Least confidence sampling. Согласно данной стратегии, для разметки выбирается тот объект, относительно разметки которого текущая модель наименее уверена. Для вероятностных моделей новый объект выбирается, например, согласно

$$\mathbf{x} = \arg \max_{i \in \{1,\ldots,M\} \setminus \mathcal{L}} 1 - P(f_m(\mathbf{x})|\mathbf{x}_i). \qquad \text{(least confidence sampling)}$$

- Max entropy. Идея такая же, но неуверенность измеряется как энтропия предсказаний метки класса:

$$\mathbf{x} = \arg \max_{i \in \{1,\ldots,M\} \setminus \mathcal{L}} -\sum_{k=1}^{K} P(y_k|\mathbf{x}_i) \log P(y_k|\mathbf{x}_i). \qquad \text{(max entropy sampling)}$$

- Query-by-Committee. Данная стратегия подразумевает, что заданы несколько моделей (или гипотез) $f^{(1)}, \ldots, f^{(C)}$. Выбирается объект, для которого результаты моделирования $f^{(c)}(\mathbf{x})$ наименее согласованы с друг с другом, например

$$\mathbf{x} = \arg \max_{i \in \{1,\ldots,M\} \setminus \mathcal{L}} - \sum_{k=1}^{K} V_k(\mathbf{x}_i) \log V_k(\mathbf{x}_i), \quad V_k(\mathbf{x}) = \frac{1}{C} \sum_{c=1}^{C} \mathbb{I}[f^{(c)_m}(\mathbf{x}) = y_k], \tag{vote entropy}$$

или

$$\mathbf{x} = \arg \max_{i \in \{1,\ldots,M\} \setminus \mathcal{L}} \frac{1}{C} \sum_{c=1}^{C} D_{KL} \left(P^{(c)}(\cdot|\mathbf{x}_i), \bar{P}(\cdot|\mathbf{x}_i) \right), \tag{KL from concensus}$$

где $\bar{P}(\cdot|\mathbf{x}_i)$ — усредненное распределение

$$\bar{P}(\cdot|\mathbf{x}_i) = \frac{1}{C} \sum_{c=1}^{C} P^{(c)}(\cdot|\mathbf{x}_i).$$

- Ожидаемое изменение качества.

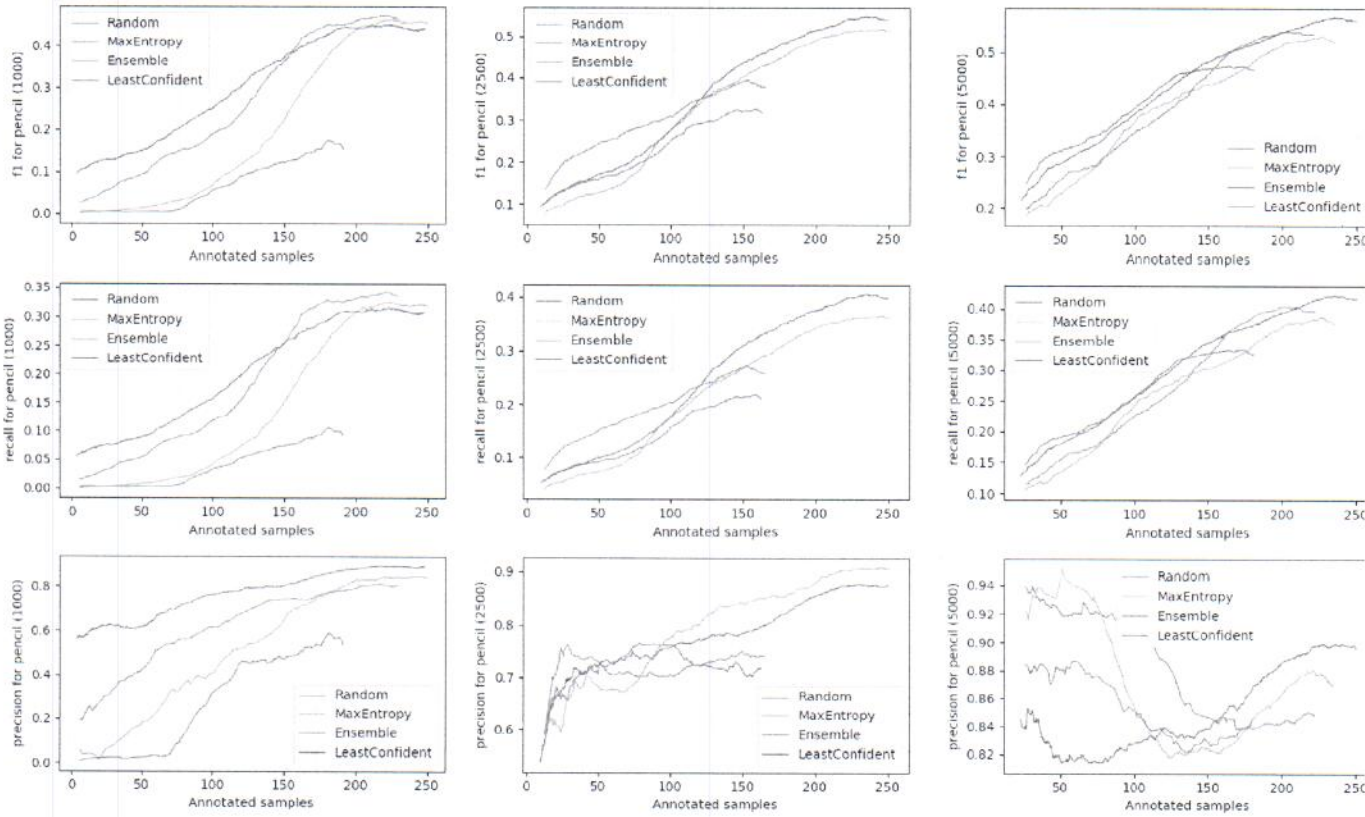
$$\mathbf{x} = \arg \max_{i \in \{1,\ldots,M\} \setminus \mathcal{L}} \sum_{k} F_1 \left(L \cup \{\mathbf{x}_i, k\} \right) p(y_i = k|\mathbf{x}_i), \tag{Expected F1 change}$$

или оптимистичный вариант:

$$\mathbf{x} = \arg \max_{i \in \{1,\ldots,M\} \setminus \mathcal{L}} F_1 \left(L \cup \{\mathbf{x}_i, \hat{y}_i\} \right), . \tag{Optimistic F1 change}$$

Для большинства функций ошибки ожидание аналитически вычислить не удается, поэтому стратегия уменьшения ожидаемой ошибки применяется вместе с методами сэмплирования, и требует большого объема вычислений. Этот критерий вычислительно наиболее сложный, так как для каждого нового объекта нужно заново обучать модель. Для уменьшения сложности выбираются кандидаты из случайного подмножества неразмеченных объектов или из множества, выбранного более простым критерием.

На рисунках 4.6 и 4.6 показано, как меняется значение метрик качества в зависимости от числа объектов класса, добавленных в выборку.



(а) $m = 1000$ (б) $m = 2500$ (в) $m = 5000$

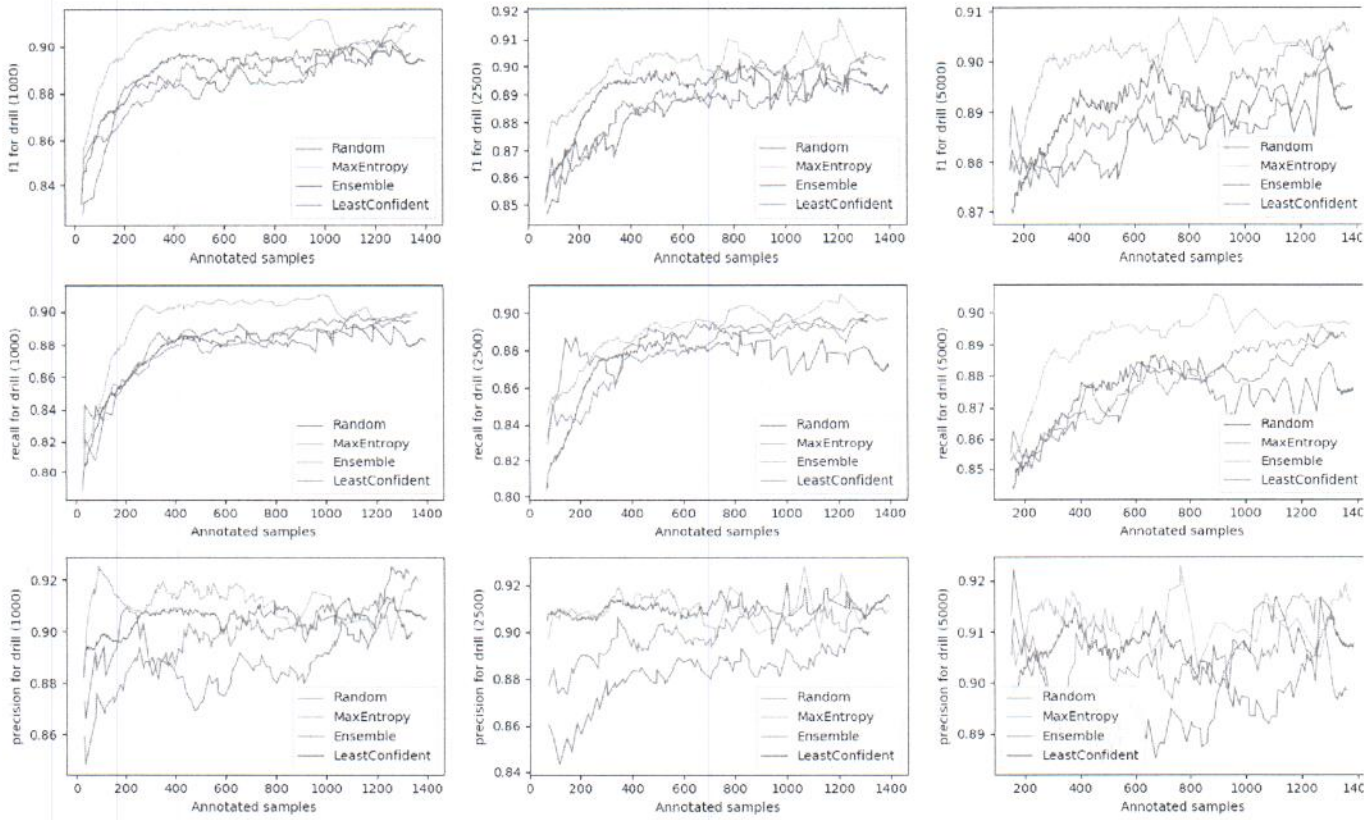
Рис. 4.5. Зависимость метрик качества от числа объектов, добавленных в выборку каждым из критериев, класс 'pencil'.

Глава 5
Прикладные задачи

В данном разделе описаны результаты применения разработанного метода оценки объема выборки и критерия идентичности двух выборок к прикладным задачам. Рассмотрены задача классификации пациентов с нарушениями работы сердечно-сосудистой системы [68] и задача прогнозирования объемов грузовых железнодорожных перевозок [62, 63].

5.1. Построение оптимальной выборки для задачи восстановления движений по сигналам электрокортикограмм

Данный раздел посвящен задаче выбора признаков при решении задачи регрессии для декодирования движений на основе сигналов ECoG. Задача состоит в том, чтобы предсказать траектории руки по временным рядам кортикальной активности. Задача прогнозирования или восстановления трехмерных траекторий движения решалась в работах [9, 37, 38, 39, 10]. Используются линейные модели и выделение признаков, основанных на экспертной информации. При-



(а) $m = 1000$ (б) $m = 2500$ (в) $m = 5000$

Рис. 4.6. Зависимость метрик качества от числа объектов, добавленных в выборку каждым из критериев 'drill'.

знаки описывают спектральную мощность сигналов в определенных полосах частот и для заданных электродов. Спектральные признаки извлекаются из сигналов электрокортикограмм с помощью спектральных преобразований или авторегрессии. В работе [37] в качестве признаков наряду с исходными сигналами использовались сигналы электрокортикограмм, преобразованные в частотную область с помощью авторегрессионной модели. Для повышения качества распознавания используются иерархические модели [39, 10], в которых модели нижнего уровня отвечают за различные виды активности — например, движение или покой. В работе [39] в качестве моделей нижнего уровня иерархии также использовалась линейная регрессия, а в качестве признаков — предобработанные сигналы электрокортикограмм. Признаковое описание, полученное с помощью спектральных преобразований имеет высокую размерность. Соседние по времени измерения одного того же сигнала или сигналов, снятых с близко расположенных электродов не являются независимыми. Описанные выше способы [9, 37, 39, 10] построить матрицу признаков в задаче декодирования нейросигналов порождают избыточное признаковое описание. Признаковое описание каждой точки содержит пространственную, временную и частотную компоненты, описывающие изменение спектральных характеристик сигналов, снимаемых с различных электродов. Такое признаковое описание обладает высокой

размерностью, сами признаки сильно скоррелированы. Кроме того, сложность модели является существенным ограничением в задаче декодирования движений с целью разработки нейрокомпьтерного интерфейса.

5.1.1. Построение многоиндексного признакового описания в задачах анализа сигналов высокой размерности

Исходные временные ряды ECoG содержат многомерные временные ряды $\mathbf{s}(t) \in \mathbb{R}^{N_{\text{ch}}}$ с напряженностью электрического поля, измеряемого каждым электродом $1, \ldots, N_{\text{ch}}$. Целевые временные ряды $\mathbf{y}(t) \in \mathbb{R}^3$ содержат трехмерные координаты[1] позиций запястья, локтя и плеча обезьяны. В экспериментах рассматривались траектории запястья. Электроды установлены в контралатеральном полушарии. В данной работе использовались способы извлечения признаков из сигналов электрокортикограмм для классификации и прогнозирования движений, а также реконструкции мысленных движений, наиболее часто описываемые в литературе [37, 9, 10]. Признаковое описание включает признаки в частотной и временной областях. Признаки из частотной области характеризуют зависящий от времени вклад каждой частоты из некоторого рассматриваемого диапазона в анализируемый сигнал. Такие признаки получают с помощью вейвлет-преобразований, быстрого преобразования Фурье или авторегрессионного анализа. Признаки из временной области, или LMP (local motor potentials) [37] представляют собой сглаженные временные ряды электрокортикограмм $\mathbf{s}(t)$. Как LMP, так и частотные признаки вычисляются с некоторой задержкой по времени относительно прогнозируемой точки.

Вычисление оптимальной задержки. Оптимальная задержка по времени выбирается исходя из максимизации линейной кросс-корреляции между временными рядами электрокортикограмм $\mathbf{s}(t)$ и целевым временными рядами $\mathbf{y}(t)$:

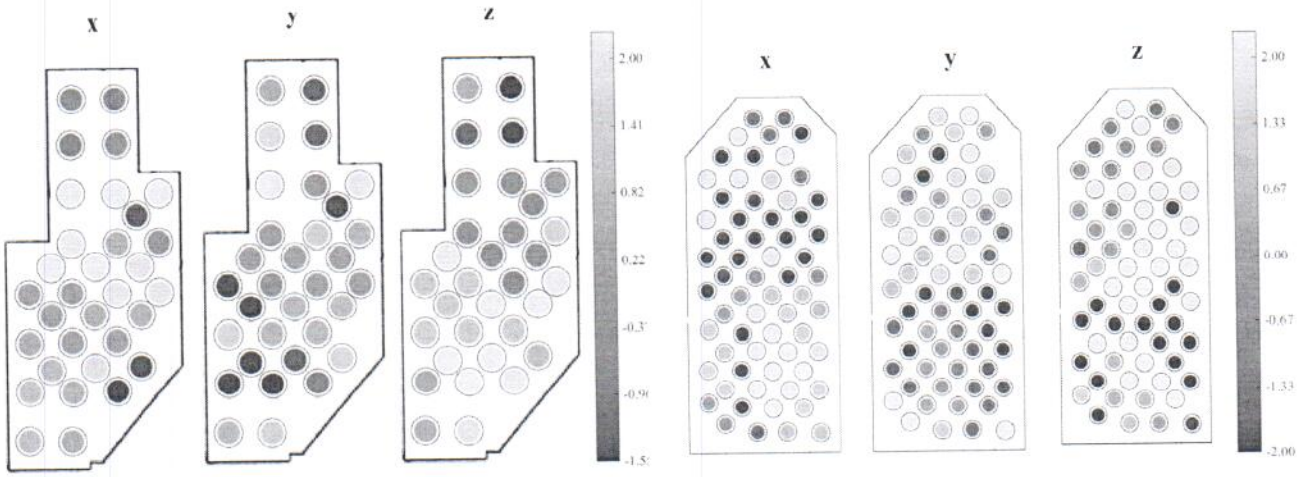
$$\tau_n^* = \underset{\tau \in [\tau_{\min}, \tau_{\max}]}{\arg\max} \frac{\left|\sum_{i=1}^m s_n(t_i + \tau) y(t_i)\right|}{\sqrt{\sum_{i=1}^m s_n(t_i + \tau) s_n(t_i + \tau)} \sqrt{\sum_{i=1}^m y(t_i) y(t_i)}}.$$

Здесь $y(t)$ – целевой временной ряд для фиксированной декартовой координаты, а $s_n(t)$ – временной ряд электрокортикограмм для n-го электрода.

Рисунок 5.1 демонстрирует, что оптимальная задержка τ^* может принимать как положительные, так и отрицательные значение. Положительное значение τ^* означает, что наибольшая корреляция регистрируемой активности s_n с прогнозируемым положением запястья $y(t)$ is наблюдается уже после момента t построенная.

Оптимальное значение задержки τ^* зависит от положения электрода, причем пространственное распределение τ^* различно для координат x, y, z целевых временных рядов и различных испытуемых. Поэтому во всех экспериментах используется нулевая задержка.

[1] В задаче декодирования речь идет о данных, описанных в работе [12]



(а) A (б) K1

Рис. 5.1. Абсолютные значения кросс-корреляции между временными рядами электрокортикограмм и целевыми временными рядами (координаты запястья) для обезьяны A (32 электрода) и K1 (64 электрода).

Вычисление частотных признаков. Для формирования из временных рядов $\mathbf{s}(t)$, $\mathbf{y}(t)$ выборки $(\underline{\mathbf{D}}, \mathbf{Y})$ выбирается M отсчетов времени $t_1, \ldots, t_M$ с шагом δt.

Признаки $\underline{\mathbf{X}}_m$ для $m-$го отсчета времени содержат пространственную, темпоральную и частотную информацию о временных рядах $\mathbf{s}(t)$ в интервале времени $[t_m - \Delta, t_m]$. Рисунок 5.2 иллюстрирует процесс извлечения признаков. Пространственная компонента представлена N_{ch} электродами. Каждый временной ряд электрокортикограмм $\mathbf{s}_n(t)$, $n = 1, \ldots, N_{\text{ch}}$ отображается в частотную область с помощью вейвлет-преобразования. В данной работе используется непрерывной вейвлет-преобразование с вейвлетом Морле. Для получения $T \times F$ признаков используется следующая процедура. Выбираются F базовых частот (параметров масштабирования) f_j, $j = 1, \ldots, F$ и к каждому временному ряду $s_n(t)$, $n = 1, \ldots, N_{\text{ch}}$ применяется вейвлет-преобразование с центрами $t_1 \leq t_i \leq t_M$ и параметрами масштабирования f_j, $j = 1, \ldots, F$:

$$W_{ijn} = \frac{1}{\sqrt{|f_j|}} \sum_{t \leq t_M} \psi\left(\frac{t - t_i}{f_j}\right) s_n(t). \tag{5.1}$$

Временные ряды напряженности $\mathbf{s}(t)$ сэмплируются с частотой 1000Hz для $N_{\text{ch}} = 32$ или $N_{\text{ch}} = 64$ электродов. Координаты запястья сэмплируются с частотой 120Hz. В вычислительных экспериментах были рассмотрены два способа извлечения признаков, обозначаемые 2D и 3D.

1. Выборка 2D содержит задержанные ($\tau = 0.65s$) временные ряды ECoG и

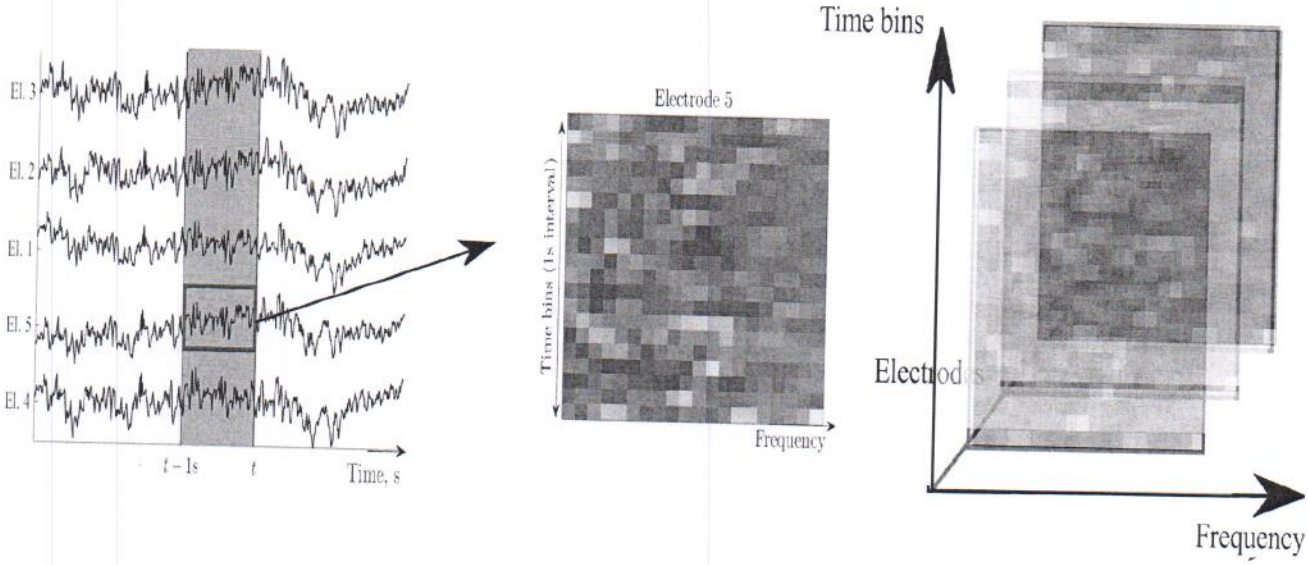


Рис. 5.2. Пример процедуры извлечения признаков. Для каждого электрода применение к односекундному интервалу $[t_m - \Delta t, t_m]$ дискретного вейвлет-преобразования (5.1) и порождает признаковое описание для точки t_m в частотно-временной области. Объединение частотно-временных признаков по всем электродам задает трехиндексную признаковую матрицу для точки t_m.

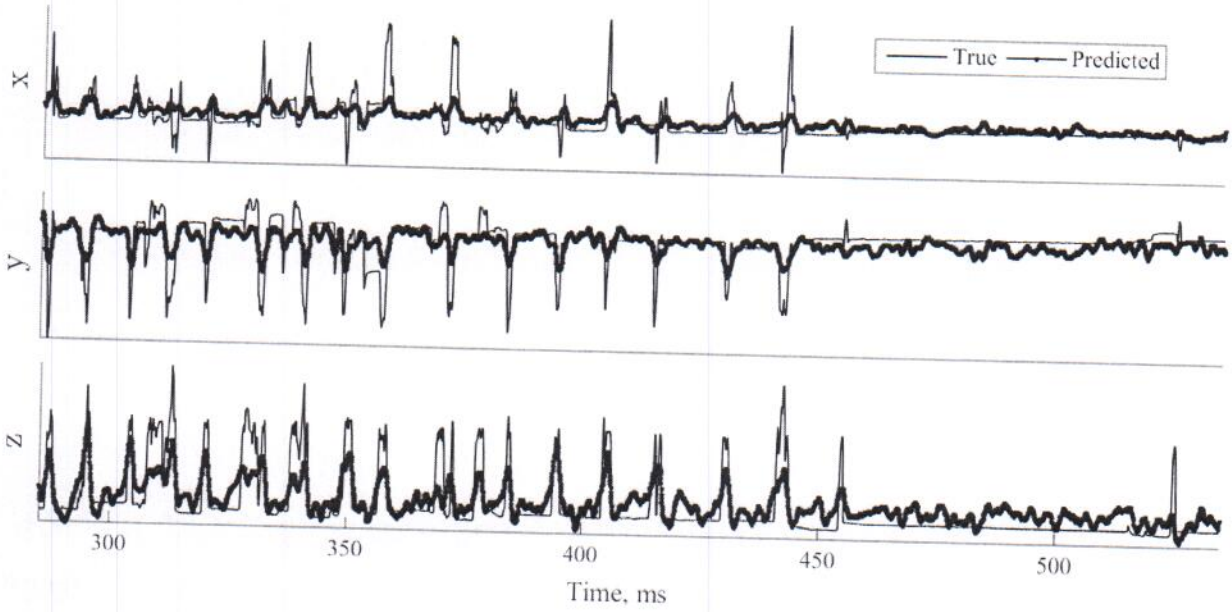


Рис. 5.3. Сегмент спрогнозированного временного ряда. Линейная регрессия, 50 лучших признаков согласно многоиндексному QPFS.

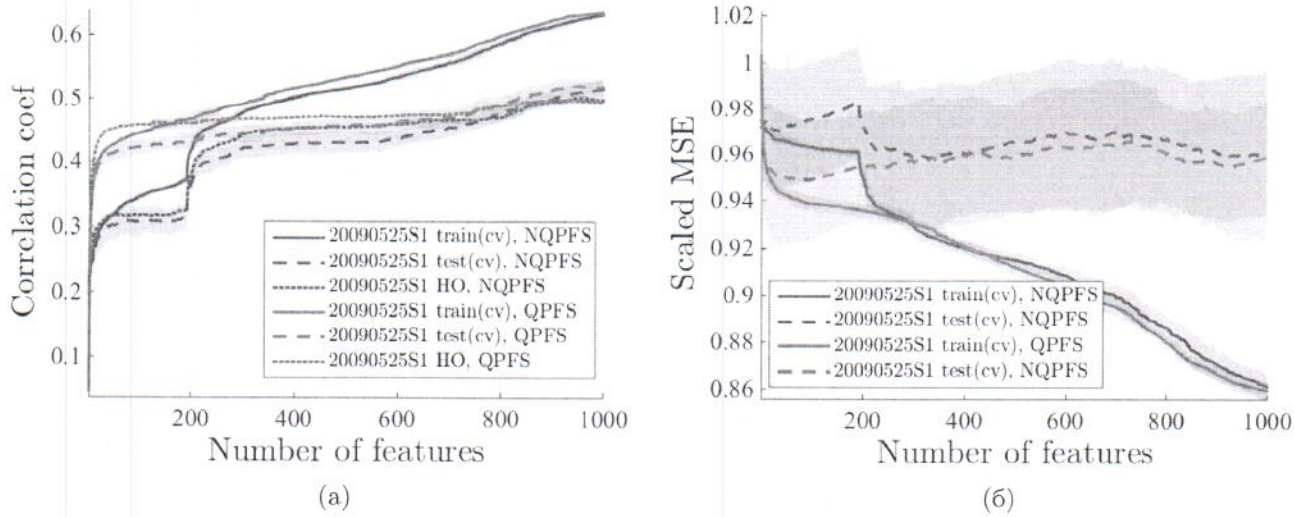


Рис. 5.4. Зависимость качества прогнозирования от сложности модели. Признаки добавляются по одному в порядке (1.21), заданном методом QPFS. Качество измеряется как (a) коэффициент корреляции между прогнозом и прогнозируемым временным рядом (b) ошибка sMSE.

коэффициенты вейвлет-преобразования:

$$\underline{\mathbf{X}}_m \in \mathbb{R}^{F\times N_{\text{ch}}}, \quad \underline{\mathbf{X}}_{mjn} = \begin{cases} s_n(t_m+\tau), j=1, \\ W_{mjn} \text{ for } j=2,\dots,F+1, \end{cases} \quad n=1,\dots,N_{\text{ch}}. \tag{5.2}$$

Для анализа временных рядов частота сэмплирования была уменьшена в 10 раз. Отсчеты времени для формирования выборки были выбраны с шагом $\delta t = 0.05s$. Были рассмотрены полосы частот 0.5–8Hz с шагом 0.5Hz, 9–18Hz с шагом 3Hz и 20-45 с шагом 5Hz.

2. Выборка 3D содержит трехиндексные признаки без временной задержки. Признаки 3D в явном виде включают предысторию $\Delta_m = [t_m - \Delta t, t_m]$ коэффициентов вейвлет-преобразования. Для формирования выборки 3D в точках $t_1,\dots,t_M$, выбирается более подробная сетка t_i, такая что $|t_i \in \Delta_m| \geq T$, где T — параметр процедуры, контролирующий загрубление информации о коэффициентах вейвлет-преобразования в интервале Δ_m. Интервал Δ_m разбивается на T последовательных интервалов δt_i, $i = 1,\dots,T$. Для $n-$го электрода (i,j,n)-й элемент трехиндексной матрицы $\underline{\mathbf{X}}_m \in \mathbb{R}^{T\times F\times N_{\text{ch}}}$ получается усреднением $W_{i'jn}$ по интервалу δ_i:

$$X_{mijn} = \frac{1}{|\delta t_i|} \sum_{i':\, t_{i'}\in\delta t_i} W_{i'jn}. \tag{5.3}$$

Коэффициенты вейвлет-преобразования вычислялись без уменьшения частоты сэмплирования со следующими параметрами: длительность исторического интервала для каждой точки выбиралась равной $\Delta t = 1$с с шагом

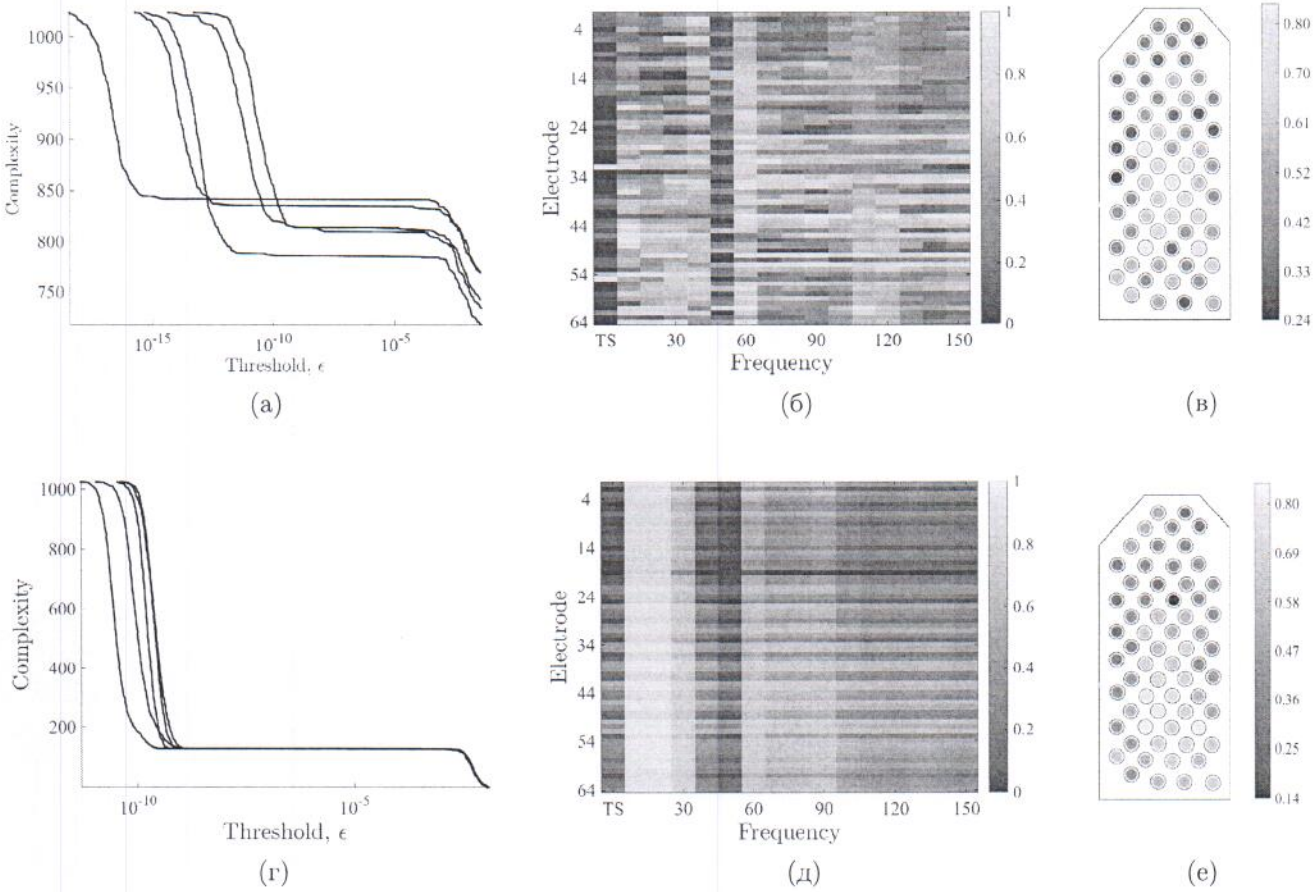


Рис. 5.5. (a) Сложность модели в зависимости от величины порога ϵ. (b) Оценки важности пар электрод-частота. Важность измеряется как усредненный по разбиениям кроссвалидации ранг (1.21). (c) Ранги отдельных электродов усредненные по частотам.

$\delta t = 0.05s$, $T = 20$, $F = 20$. Частоты были распределены логарифмически в интервале 10 – 500 Hz.

Результаты отбора признаков с помощью метода QPFS В данном разделе проводится сравнение исходного метода QPFS и предлагаемой тензорной модификации NQPFS. Код для проведения вычислительных экспериментов находится по адресу [83]. Рисунок 5.5 иллюстрирует результат выбора признаков с помощью методов QPFS и NQPFS для выборки 2D (5.2). Для оценки качества результатов выбора признаков методом (N)QPFS часть выборки, соответствующая отсчетам с 5 до 645 секунд была разделена на $K = 5$ частей, каждая из которых 1 раз входила в контрольную выборку и 4 раза — в обучающую. Оставшиеся данные (до 950 секунд) использовались для валидации результатов.

Для каждой обучающей выборки решалась релаксированная задача (1.11). Полученные значения структурной переменной определяли некоторое ранжирование признаков (1.21). Будем говорить, что $\chi_{(i,j)}$ имеет ранг n, если он не хуже чем $n-1$ признаков. Признаки с более высоким рангом с большей вероятностью добавляются в модель, поэтому важность признаков измерялась как средний по разбиениям кроссвалидации ранг.

Выбор активного набора признаков $\mathbf{X}_{\underline{\mathbf{A}}(\epsilon)}$ определяется бинаризацией переменной $\underline{\hat{\mathbf{A}}}$ по некоторому порогу $\epsilon \in [0, 1)$. Качество выбранного набора признаков $\underline{\mathbf{A}}(\epsilon)$ оценивалось как качество $\mathcal{Q}(\mathbf{X}^{\text{test}}_{\underline{\mathbf{A}}(\epsilon)}\mathbf{w}_{\underline{\mathbf{A}}}, \mathbf{Y}^{\text{test}})$ линейной регрессии (1.2) с параметрами $\mathbf{w}_{\underline{\mathbf{A}}(\epsilon)}$, оцененными на обучающей выборке.

Рисунки 5.3, 5.4 и 5.5 иллюстрируют результаты методов (N)QPFS для выборки 2D. Рисунок 5.3 демонстрирует пример восстановленных проекций траектории запястья обезьяны. Прогноз получен по 25 лучшим признакам, выбранным методом NQPFS для выборки 2D. Заметим, что хотя восстановленная траектория менее амплитудна, чем истинная при реконструкции активной части движений, при прогнозировании временного ряда, соответствующего состоянию покоя наблюдаются значительные шумовые колебания. Для устранения этого эффекта используются смеси моделей, таким образом чтобы отдельные модели отвечали за однородные участки временных рядов. Вопрос о построении многоуровневых моделей более подходящей моделей остается за рамками данной работы.

Рисунок 5.4 демонстрирует кривые качества для методов QPFS и NQPFS. На рисунок 5.4(a) изображена зависимость коэффициента корреляции между восстановленной $\hat{\mathbf{Y}}$ и $\mathbf{Y}$ истинной траекториями запястья от сложности $N(\epsilon) = \sum_{i,j} a_{ij}(\epsilon)$ модели. Качество на тестовой выборке перестает увеличиваться после добавления примерное 300 признаков; качество на отложенной валидационной выборке не меняется после добавления примерно 100 признаков. Рисунок 5.4(a) демонстрирует аналогичную зависимость для ошибки sMSE.

Рисунки 5.5(a) и (d) демонстрируют зависимость сложности N от порога

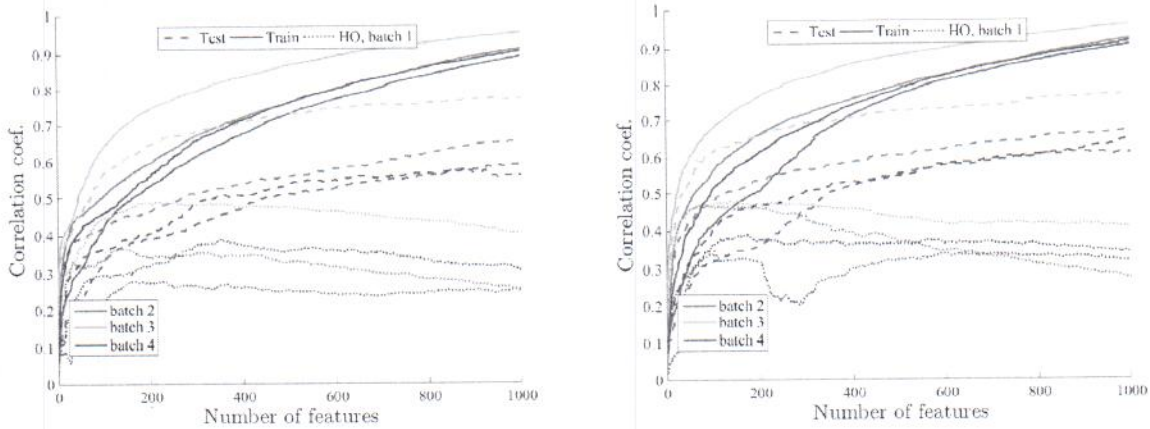


Рис. 5.6. Зависимость коэффициента корреляции между прогнозом и прогнозируемым временным рядом от сложности модели. Признаки добавляются по одному в порядке (1.21), заданном методом (a) QPFS и (b) NQPFS.

ϵ, используемого для бинаризации релаксированной структурной переменной для методов QPFS и NQPFS. Каждая кривая соответствует одному разбиению на кроссвалидации. На рисунках 5.5(b) и (e) цветом закодированы значения важности признаков (пар электрод-частота). Первый столбец матрицы соответствует задержанным временным рядам ECoG. На рисунки 5.5(c) и (f) цветом показаны усредненные по частотам значения важности для каждого электрода. Электроды изображены в соответствии с их расположением в эксперименте (обезьяны A). Заметим, что хотя кривые качества 5.4 для двух методов очень похожи, рисунок 5.5 демонстрирует, что методы QPFS b NQPFS выбирают различные наборы признаков.

В случае выборки 3D данные обрабатывались блоками для в связи в большой размерностью выборки. Рисунок 5.6 демонстрирует кривые качества для каждого из блоков. На рисунке 5.7 показано, как меняется распределение важности пар электрод-частота от блока к блоку. Справа на каждом подрисунке 5.7 цветом закодированы усредненные по частотам важности электродов. Согласно рисункам, QPFS придает большее значение низким частотам и передним электродам, чем NQPFS. Для демонстрации на рисунках была выбран 10−й интервал, но для остальных интервалов наблюдались похожие распределения.

Сравнение качества прогнозирования при отборе признаков с помощью QPFS и регрессии методом частичных наименьших квадратов

Для оценки качества предложенного метода было проведено сравнение качества линейной регрессии на признаках, выбранных методами QPFS и NQPFS с качеством PLS и NPLS регрессии. В рамках PLS регрессии снижение размерности достигается не за счет выбора исходных признаков, а путем выбора факторов — линейных комбинаций исходных признаков. Преобразования признаков выполняются таким образом, чтобы уменьшить избыточность признаков, од-

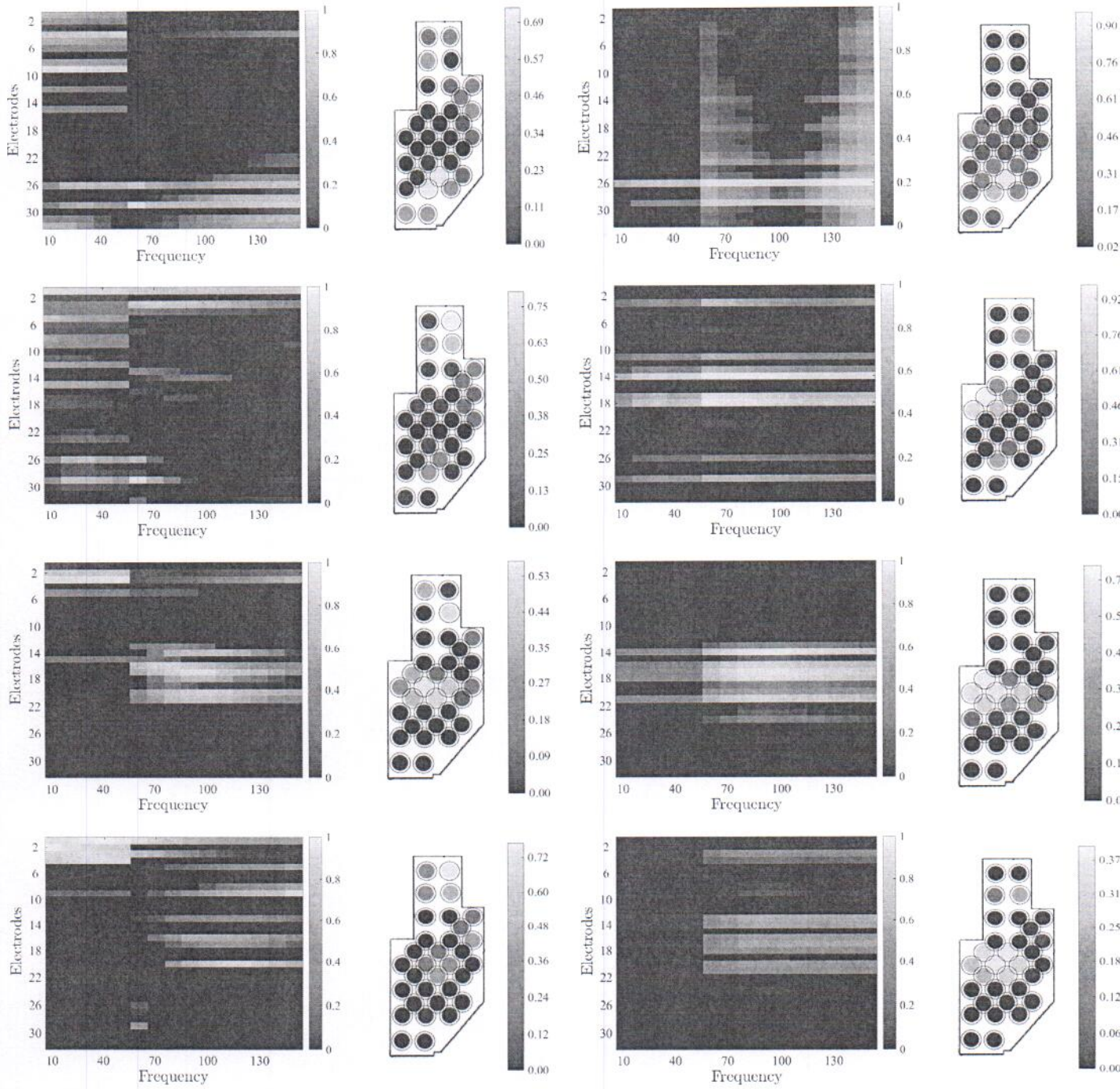


Рис. 5.7. Важность пары электрод-частота, интервал номер 10. Слева: QPFS, справа: NQPFS.

новременно с настройкой параметров модели. Для этого PLS одновременно декомпозирует $\mathbf{X}$ (1.5) и $\mathbf{Y}$ (1.1) в N факторов $\mathbf{T} \in \mathbb{R}^{M \times N}$ и $\mathbf{U} \in \mathbb{R}^{M \times N}$,

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\intercal} + \mathbf{E}, \quad \mathbf{P}^{\intercal}\mathbf{P} = \mathbf{I}_N,$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\intercal} + \mathbf{F}, \quad \mathbf{Q}^{\intercal}\mathbf{Q} = \mathbf{I}_N,$$

so that $\mathbf{t}_i^{\intercal}\mathbf{u}_i = \beta_i \to \max$, $i = 1, \ldots, N$. Решение задачи регрессии задается как

$$\hat{\mathbf{Y}} = \hat{\mathbf{T}}\mathrm{diag}(\boldsymbol{\beta})\mathbf{Q}^{\intercal} = \mathbf{X}\mathbf{W},$$

где $\hat{\mathbf{T}} = \mathbf{X}\mathbf{P}(\mathbf{P}^{\intercal}\mathbf{P})^{-1}$.

В данной работе используется NPLS, модификации PLS для многоиндексных данных, предложенная в [84]. Подробное описание алгоритма содержится в [20].

В таблице 5.1 приведены значения коэффициента корреляции (1.7) между восстановленной и истинной траекториями запястья для методов QPFS, NQPFS, PLS и NPLS. Каждый метод выбирал по $N = 10, 25, 200$ или 500 признаков (факторов, в случае PLS и NPLS). Для QPFS и NQPFS затем оценивались параметры линейной регрессии (1.2) без дополнительной регуляризации. В случае с PLS и NPLS, параметры регрессии являются часть решения задачи. Для каждого эксперимента (5 экспериментов для обезьяны A и 3 эксперимента для обезьяны K) и каждого значения N лучший результат выделен жирным шрифтом.

В дополнение к рассмотренным ранее критериям (1.7) и (1.6) были рассмотрены расстояние DTW и ошибка MADE (mean average difference error) между $\hat{\mathbf{Y}}$ и $\mathbf{Y}$. Расстояние DTW используется как мера отклонения между последовательностями, в том случае, когда поточечное сравнение не имеет смысла или невозможно. В этом случае выполняется процедура выравнивания последовательностей, стоимость которой принимается за расстояние. Ошибка MADE оценивает гладкость восстановленной траектории как

$$\mathrm{MADE}(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{\sum_{m=1}^{M} |\hat{\mathbf{y}}'_m - \mathbf{y}'_m|}{\sum_{m=1}^{M} |\bar{\mathbf{y}}' - \mathbf{y}'_m|}.$$

Эта ошибка важна в контексте разработки методов для создания НКИ. Результаты сравнения по всем четырем критериям приведены на рисунке 5.8. На рисунке 5.8(a) представлены значения критериев для $N = 10, 25, 50, 100, 200, 500$. Для каждого метода подсчитан средний ранг: если давал лучший результат, назначался ранг 1 и так далее до четырех. На рисунке 5.8(b) приведены усредненные по восьми экспериментам результаты.

Кроме того, была проведена оценка быстродействия сравниваемых методов. Так как методы PLS и NPLS включают оценку параметров модели, время на оценку параметров регрессии (1.3) также учитывалось при анализе методов QPFS и NQPFS. На рисунке 5.9 изображена зависимость времени исполнения от

Таблица 5.1. Коэффициент корреляции между восстановленной и истинной траекториями.

Эксперимент	Метод	$N = 10$	$N = 25$	$N = 200$	$N = 500$
A, 20090116	QPFS	$\mathbf{0.333 \pm 0.007}$	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	$\mathbf{0.511 \pm 0.020}$
	PLS	0.306 ± 0.007	$\mathbf{0.376 \pm 0.012}$	$\mathbf{0.476 \pm 0.017}$	0.51 ± 0.010
	NPLS	0.0367 ± 0.006	0.125 ± 0.016	0.329 ± 0.021	0.358 ± 0.054
K, 20090525	QPFS	$\mathbf{0.252 \pm 0.079}$	0.288 ± 0.096	0.392 ± 0.144	$\mathbf{0.467 \pm 0.132}$
	NQPFS	0.185 ± 0.040	0.225 ± 0.063	0.374 ± 0.119	0.416 ± 0.112
	PLS	0.24 ± 0.042	$\mathbf{0.288 \pm 0.095}$	$\mathbf{0.429 \pm 0.128}$	0.409 ± 0.122
	NPLS	0.0359 ± 0.026	0.115 ± 0.007	0.278 ± 0.092	0.336 ± 0.072
K, 20090527	QPFS	$\mathbf{0.346 \pm 0.020}$	$\mathbf{0.391 \pm 0.007}$	0.403 ± 0.007	$\mathbf{0.441 \pm 0.006}$
	NQPFS	0.185 ± 0.012	0.191 ± 0.007	0.384 ± 0.006	0.413 ± 0.009
	PLS	0.297 ± 0.010	0.347 ± 0.013	$\mathbf{0.428 \pm 0.018}$	0.417 ± 0.028
	NPLS	0.118 ± 0.019	0.23 ± 0.016	0.354 ± 0.025	0.371 ± 0.018
K, 20090602	QPFS	$\mathbf{0.357 \pm 0.017}$	$\mathbf{0.382 \pm 0.014}$	0.424 ± 0.014	0.468 ± 0.010
	NQPFS	0.29 ± 0.007	0.306 ± 0.003	0.381 ± 0.006	$\mathbf{0.474 \pm 0.017}$
	PLS	0.341 ± 0.012	0.369 ± 0.010	$\mathbf{0.442 \pm 0.008}$	0.416 ± 0.015
	NPLS	0.111 ± 0.009	0.167 ± 0.007	0.306 ± 0.012	0.344 ± 0.026
A, 20081127	QPFS	0.281 ± 0.014	0.323 ± 0.017	0.397 ± 0.028	0.438 ± 0.066
	NQPFS	0.205 ± 0.008	0.227 ± 0.012	$\mathbf{0.436 \pm 0.012}$	$\mathbf{0.451 \pm 0.016}$
	PLS	$\mathbf{0.293 \pm 0.016}$	$\mathbf{0.353 \pm 0.009}$	0.429 ± 0.020	0.434 ± 0.064
	NPLS	0.0402 ± 0.012	0.0698 ± 0.002	0.2 ± 0.011	0.224 ± 0.025
A, 20081224	QPFS	0.262 ± 0.012	0.291 ± 0.035	0.332 ± 0.089	0.4 ± 0.079
	NQPFS	0.196 ± 0.005	0.216 ± 0.003	0.298 ± 0.075	$\mathbf{0.407 \pm 0.086}$
	PLS	$\mathbf{0.276 \pm 0.006}$	$\mathbf{0.331 \pm 0.010}$	$\mathbf{0.349 \pm 0.086}$	0.401 ± 0.090
	NPLS	0.0894 ± 0.010	0.129 ± 0.008	0.169 ± 0.065	0.206 ± 0.082
A, 20090121	QPFS	0.222 ± 0.006	0.242 ± 0.011	0.344 ± 0.027	0.368 ± 0.026
	NQPFS	0.18 ± 0.011	0.224 ± 0.011	$\mathbf{0.362 \pm 0.018}$	$\mathbf{0.371 \pm 0.036}$
	PLS	$\mathbf{0.248 \pm 0.012}$	$\mathbf{0.297 \pm 0.016}$	0.341 ± 0.022	0.369 ± 0.029
	NPLS	0.0262 ± 0.011	0.0641 ± 0.008	0.151 ± 0.014	0.172 ± 0.021
A, 20090611	QPFS	$\mathbf{0.3 \pm 0.017}$	0.321 ± 0.022	0.345 ± 0.011	$\mathbf{0.357 \pm 0.011}$
	NQPFS	0.251 ± 0.005	0.256 ± 0.005	0.326 ± 0.004	0.355 ± 0.005
	PLS	0.282 ± 0.010	$\mathbf{0.321 \pm 0.009}$	$\mathbf{0.359 \pm 0.008}$	0.357 ± 0.011
	NPLS	0.108 ± 0.006	0.184 ± 0.007	0.27 ± 0.010	0.27 ± 0.008
Average	QPFS	$\mathbf{0.294 \pm 0.022}$	0.322 ± 0.026	0.381 ± 0.045	$\mathbf{0.43 \pm 0.043}$
	NQPFS	0.22 ± 0.015	0.244 ± 0.014	0.377 ± 0.032	0.425 ± 0.038
	PLS	0.285 ± 0.014	$\mathbf{0.335 \pm 0.022}$	$\mathbf{0.407 \pm 0.038}$	0.414 ± 0.046
	NPLS	0.0705 ± 0.012	0.135 ± 0.009	0.257 ± 0.031	0.285 ± 0.038

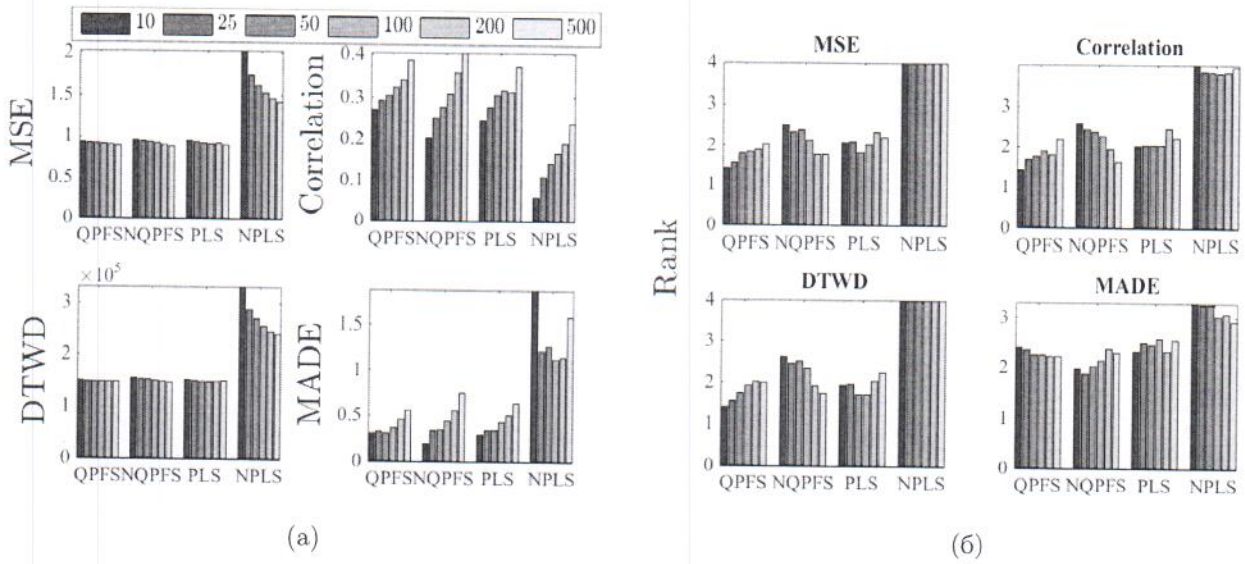


Рис. 5.8. (a) Средние значения рассматриваемых критериев качества. (b) Средние ранги сравниваемых алгоритмов (меньше – лучше).

размерности исходной выборки. Были рассмотрены выборки $\underline{\mathbf{D}} \in \mathbb{R}^{5000 \times n \times n}$, где значение n менялось от 10 до 50 с шагом 5. Методы выбирали 100, 625 или 2500 признаков из n^2 исходных признаков. На рисунке 5.9(a) зависимость времени от n для QPFS и NQPFS построена без учета оценки параметров модели. Так как методы QPFS и NQPFS ранжируют все признаки одновременно, время не зависит от количества выбираемых признаков. Основная нагрузка в методе QPFS приходится на вычисление матрицы сходства размером $n^2 \times n^2$ и проверку ее адекватности. При этом NQPFS оперирует двумя матрицами размера $n \times n$, поэтому наблюдается существенная разница во времени. На рисунок 5.9(b) время подсчитано с учетом оценки параметров линейной регрессии. Заметим, что использование метода NQPFS не менее эффективно, чем использование PLS регрессии.

5.2. Декодирование электрокортикограмм для прогнозирования движений пальцев

При экспериментальной оценке предложенного метода были рассмотрены данные из эксперимента по восстановлению жестов [85]. Выборка содержит временные ряды электрокортикограмм и координат пальцев. Были рассмотрены три типа экспериментов:

- 'base': испытуемые выполняли базовое движение,
- 'fingerflex': испытуемые двигали одним пальцем,
- 'freeform': испытуемые совершали произвольные движения пальцами.

Каждая запись велась в течение двух минут. Результаты восстановления движений перечислены в таблицах 5.2, 5.3, 5.444.

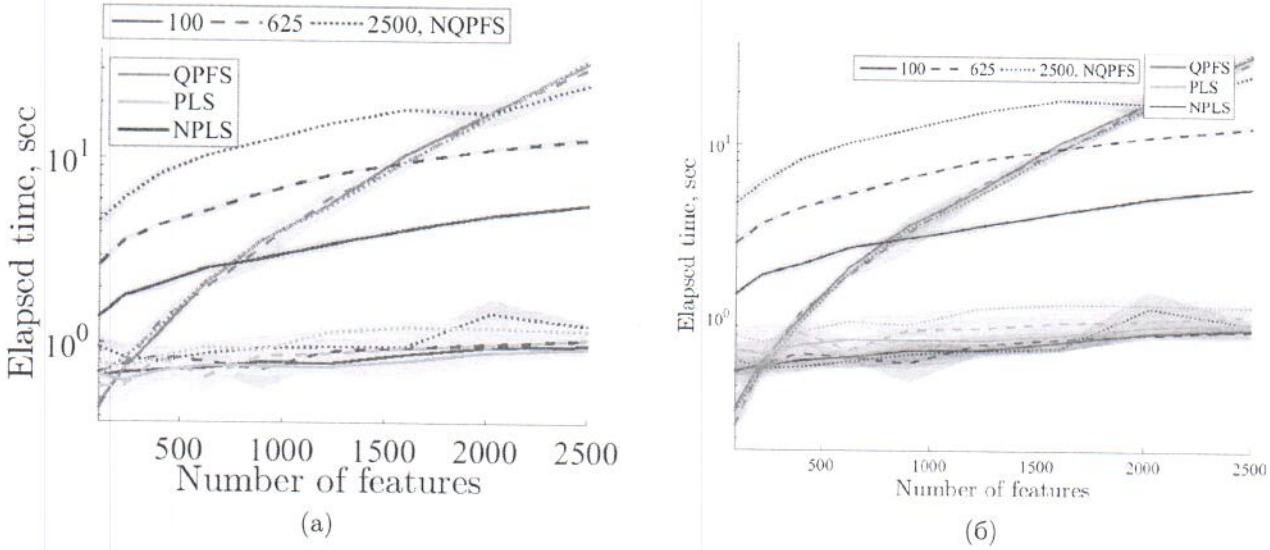


Рис. 5.9. Время работы алгоритмов (a) без учета оценки параметров модели (b) с учетом оценки параметров.

Таблица 5.2. Коэффициент корреляции между восстановленной и истинной траекториями для эксперимента 'base'.

Subject	Algorithm	$N = 10$	$N = 25$	$N = 50$	$N = 100$
wm_base	QPFS	0.141 ± 0.058	0.0979 ± 0.067	0.0863 ± 0.064	0.0669 ± 0.075
	NQPFS	$\mathbf{0.23 \pm 0.080}$	$\mathbf{0.227 \pm 0.087}$	0.204 ± 0.089	0.172 ± 0.084
	PLS	0.109 ± 0.068	0.0911 ± 0.068	0.107 ± 0.061	0.0702 ± 0.068
	NPLS	0.218 ± 0.103	0.212 ± 0.101	$\mathbf{0.226 \pm 0.113}$	$\mathbf{0.231 \pm 0.112}$
de_base	QPFS	$\mathbf{0.336 \pm 0.332}$	$\mathbf{0.325 \pm 0.306}$	$\mathbf{0.343 \pm 0.323}$	$\mathbf{0.363 \pm 0.310}$
	NQPFS	0.293 ± 0.310	0.232 ± 0.299	0.212 ± 0.257	0.248 ± 0.342
	PLS	0.118 ± 0.031	0.0453 ± 0.021	0.0383 ± 0.018	0.0978 ± 0.054
	NPLS	0.102 ± 0.032	0.104 ± 0.047	0.0643 ± 0.042	0.0961 ± 0.120
bp_base	QPFS	0.354 ± 0.317	$\mathbf{0.374 \pm 0.310}$	$\mathbf{0.35 \pm 0.319}$	$\mathbf{0.331 \pm 0.286}$
	NQPFS	$\mathbf{0.369 \pm 0.311}$	0.345 ± 0.335	0.233 ± 0.372	0.218 ± 0.377
	PLS	0.128 ± 0.051	0.187 ± 0.043	0.258 ± 0.050	0.32 ± 0.031
	NPLS	0.0422 ± 0.031	0.0983 ± 0.098	0.097 ± 0.118	0.101 ± 0.082
ca_base	QPFS	0.0235 ± 0.017	0.0282 ± 0.022	0.0486 ± 0.046	0.175 ± 0.088
	NQPFS	$\mathbf{0.186 \pm 0.244}$	$\mathbf{0.129 \pm 0.111}$	$\mathbf{0.264 \pm 0.351}$	$\mathbf{0.205 \pm 0.224}$
	PLS	0.0526 ± 0.032	0.0627 ± 0.022	0.0655 ± 0.070	0.0582 ± 0.026
	NPLS	0.0836 ± 0.031	0.116 ± 0.036	0.135 ± 0.028	0.107 ± 0.054
cc_base	QPFS	0.259 ± 0.269	0.232 ± 0.211	0.189 ± 0.217	0.128 ± 0.234
	NQPFS	$\mathbf{0.267 \pm 0.183}$	$\mathbf{0.462 \pm 0.138}$	$\mathbf{0.423 \pm 0.129}$	$\mathbf{0.39 \pm 0.102}$
	PLS	0.0893 ± 0.058	0.057 ± 0.022	0.0482 ± 0.034	0.122 ± 0.037
	NPLS	0.25 ± 0.111	0.315 ± 0.075	0.328 ± 0.097	0.312 ± 0.132

Таблица 5.3. Коэффициент корреляции между восстановленной и истинной траекториями для эксперимента 'fingers'.

Subject	Algorithm	$N = 10$	$N = 25$	$N = 50$	$N = 100$
wm_fingerflex	QPFS	0.407 ± 0.029	0.392 ± 0.051	0.339 ± 0.154	0.36 ± 0.092
	NQPFS	$\mathbf{0.424 \pm 0.009}$	0.388 ± 0.068	0.387 ± 0.068	0.404 ± 0.038
	PLS	0.415 ± 0.019	$\mathbf{0.488 \pm 0.019}$	$\mathbf{0.571 \pm 0.016}$	$\mathbf{0.582 \pm 0.030}$
	NPLS	0.395 ± 0.021	0.394 ± 0.017	0.401 ± 0.047	0.348 ± 0.111
bp_fingerflex	QPFS	0.235 ± 0.034	0.239 ± 0.027	0.242 ± 0.026	0.238 ± 0.021
	NQPFS	0.0287 ± 0.020	0.0405 ± 0.019	0.0587 ± 0.015	0.0702 ± 0.018
	PLS	$\mathbf{0.237 \pm 0.007}$	$\mathbf{0.282 \pm 0.006}$	$\mathbf{0.318 \pm 0.014}$	$\mathbf{0.366 \pm 0.013}$
	NPLS	0.0632 ± 0.010	0.0329 ± 0.011	0.0452 ± 0.012	0.0811 ± 0.030
ca_fingerflex	QPFS	0.0811 ± 0.031	0.0457 ± 0.031	0.0883 ± 0.030	0.103 ± 0.045
	NQPFS	0.0824 ± 0.020	0.08 ± 0.038	0.0796 ± 0.041	0.11 ± 0.017
	PLS	$\mathbf{0.263 \pm 0.012}$	$\mathbf{0.284 \pm 0.020}$	$\mathbf{0.284 \pm 0.017}$	$\mathbf{0.274 \pm 0.014}$
	NPLS	0.0963 ± 0.021	0.0128 ± 0.007	0.0225 ± 0.013	0.0202 ± 0.015

Таблица 5.4. Коэффициент корреляции между восстановленной и истинной траекториями для эксперимента 'freeform'.

Subject	Algorithm	$N = 10$	$N = 25$	$N = 50$	$N = 100$
wm_freeform	QPFS	0.0626 ± 0.032	0.108 ± 0.041	0.102 ± 0.032	0.121 ± 0.059
	NQPFS	0.345 ± 0.032	0.302 ± 0.060	0.277 ± 0.032	0.314 ± 0.042
	PLS	$\mathbf{0.392 \pm 0.196}$	$\mathbf{0.333 \pm 0.214}$	0.343 ± 0.189	0.406 ± 0.182
	NPLS	0.323 ± 0.152	0.328 ± 0.199	$\mathbf{0.409 \pm 0.192}$	$\mathbf{0.437 \pm 0.210}$
de_freeform	QPFS	$\mathbf{0.32 \pm 0.149}$	$\mathbf{0.339 \pm 0.163}$	$\mathbf{0.355 \pm 0.159}$	0.366 ± 0.169
	NQPFS	0.31 ± 0.059	0.286 ± 0.059	0.228 ± 0.118	0.204 ± 0.117
	PLS	0.297 ± 0.040	0.333 ± 0.009	0.338 ± 0.025	$\mathbf{0.4 \pm 0.054}$
	NPLS	0.243 ± 0.047	0.277 ± 0.034	0.308 ± 0.054	0.357 ± 0.059
bp_freeform	QPFS	0.179 ± 0.086	0.194 ± 0.086	0.127 ± 0.072	0.0352 ± 0.032
	NQPFS	$\mathbf{0.316 \pm 0.032}$	$\mathbf{0.368 \pm 0.045}$	0.101 ± 0.108	0.061 ± 0.027
	PLS	0.238 ± 0.044	0.274 ± 0.038	$\mathbf{0.284 \pm 0.051}$	$\mathbf{0.269 \pm 0.045}$
	NPLS	0.0302 ± 0.022	0.0453 ± 0.020	0.0806 ± 0.031	0.0961 ± 0.028
ca_freeform	QPFS	0.13 ± 0.072	0.101 ± 0.029	0.108 ± 0.045	$\mathbf{0.211 \pm 0.175}$
	NQPFS	$\mathbf{0.252 \pm 0.157}$	0.226 ± 0.172	$\mathbf{0.185 \pm 0.189}$	0.195 ± 0.183
	PLS	0.153 ± 0.051	$\mathbf{0.233 \pm 0.052}$	0.183 ± 0.030	0.174 ± 0.055
	NPLS	0.0724 ± 0.033	0.0686 ± 0.025	0.0843 ± 0.026	0.106 ± 0.130
cc_freeform	QPFS	0.182 ± 0.042	0.194 ± 0.021	0.236 ± 0.058	0.246 ± 0.016
	NQPFS	$\mathbf{0.28 \pm 0.027}$	$\mathbf{0.304 \pm 0.038}$	$\mathbf{0.316 \pm 0.042}$	$\mathbf{0.264 \pm 0.111}$
	PLS	0.155 ± 0.025	0.153 ± 0.013	0.202 ± 0.036	0.24 ± 0.047
	NPLS	0.0415 ± 0.027	0.0403 ± 0.031	0.0478 ± 0.019	0.0503 ± 0.034

5.3. Обнаружение внутреннего плагиата

В данном разделе приведены результаты анализа предложенного метода снижения размерности с сохранением локальной структуры близости между объектами при решении задачи обнаружения внутреннего плагиата.

Задача обнаружения внутреннего плагиата состоит в поиске заимствованных частей документа без использования внешних источников. При решении задачи исследуемый текст некоторым образом разбивается на сегменты. Каждому сегменту соответствует его вектор признаков. Сегмент считается минимальной единицей заимствования. Он считает либо полностью заимствованным, либо полностью оригинальным. Тогда задача обнаружения внутреннего плагиата является задачей классификации, где объектами являются векторы признаков сегментов, а классами — метки заимствования или оригинальности.

Рассматривается набор документов. Каждый документ рассматривается как последовательность сегментов s_i, каждый из которых описывается вектором признаков $\mathbf{x}$. В данной работе в качестве сегментов рассматриваются предложения.

Способы разбиения на сегменты, как и способы вычисления вектора признаков, являются предметом отдельного исследования. Подходы [143, 163, 123], продемонстрировали на конкурсе PAN-2011 [155] наилучшее качество решения задачи обнаружения внутреннего плагиата. Они включают разбиение документа на абзацы, предложения, блоки слов или символов. В них используются признаки, основанные на частотных профилях сегментов. Такие признаки имеют размерность, пропорциональную числу слов в документе, сильно разрежены и не всегда информативны. В данной работе предполагается, что объекты с таким признаковым описанием подчиняются гипотезе многообразия. Это означает, что метрически близкие объекты могут быть геодезически далекими и дает нам возможность применить методы снижения размерности для улучшения качества классификации.

Каждому сегменту s_i поставлена в соответствие метка класса $y_i \in \{0, 1\} : y_i = 1$, если s_i — заимствованный сегмент 1, иначе $y_i = 0$. Задача распознавания внутреннего плагиата ставится как задача восстановления меток y_i по документу.

5.3.1. Иллюстрация вложения реальных данных

Для демонстрации работы алгоритма на реальных данных из предоставленного корпуса [155] `part1` выделен один из документов. Выделенные из него объекты были разделены на обучающую и тестовую выборки. Каждой из них соответствуют непрерывные части текста. Это разделение необходимо для демонстрации работы предложенной модификации и не учитывается при применении оригинального t-SNE. На рис. 5.10 приведен результат применения оригинального непараметрического метода t-SNE к объектам, выделенным из

выбранного документа. На нем видно, что объекты, соответствующие заимствованным частям текста, имеют очаги концентрации в исходном пространстве, что свидетельствует об информативности выбранных признаков.

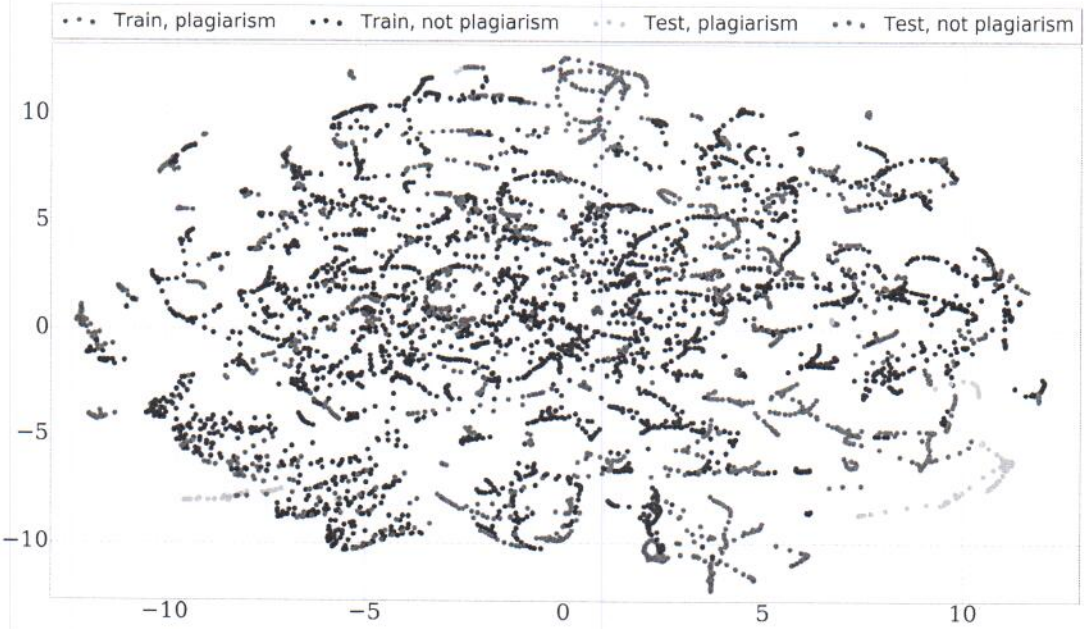
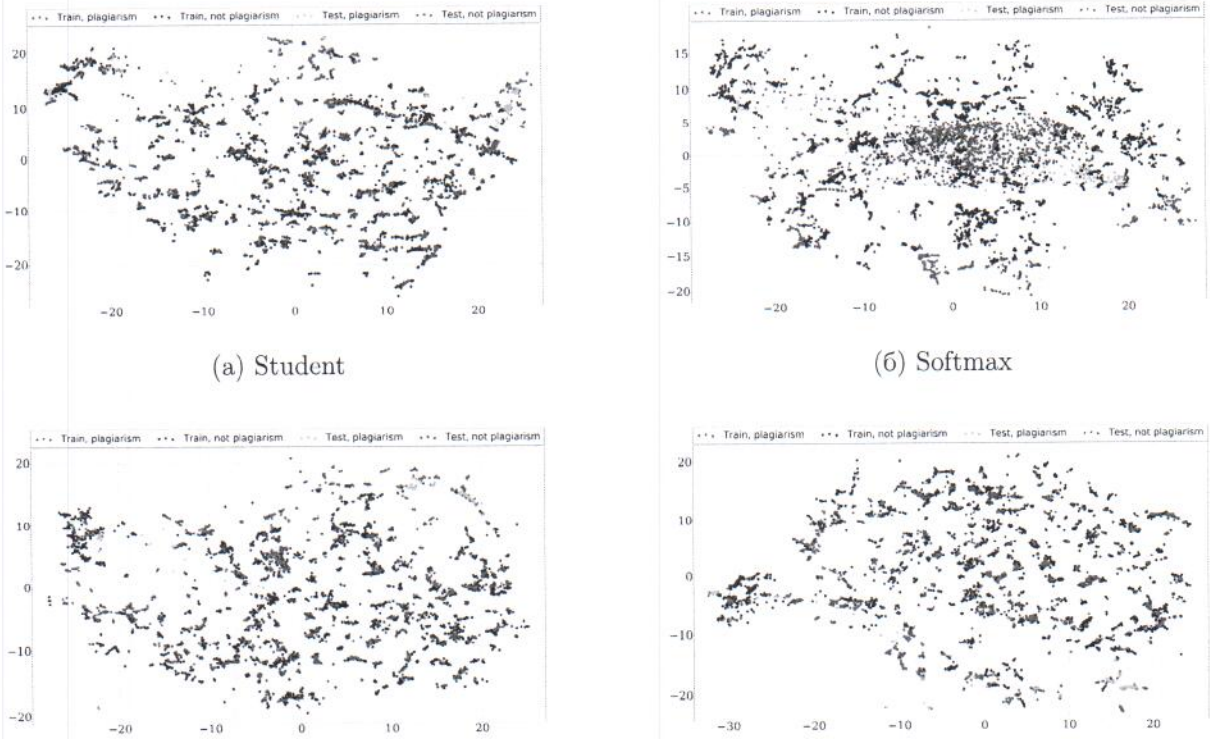


Рис. 5.10. Визуализация документа с использованием оригинального алгоритма t-SNE.

На рис. 5.11, 5.12, изображены результаты вложения данных выбранного документа с использованием предложенного алгоритма при различных методах начальной инициализации и при различных значениях веса μ. При выполнении вложений были зафиксированы параметры $S_s = 500$ и $S_b = 200$. В эксперименте данные из выбранного документа были разделены на обучающую и тестовую выборки. В тестовую часть попали образы предложений, которые образовывали в исходном тексте непрерывную цепочку. Обучающая часть вкладывалась с учетом ее разметки, а тестовая — без учета.

Из полученных графиков можно сделать вывод, что предложенная модификация при больших значениях веса μ принимает на себя часть ответственности за классификацию. Она склонна разделять и кластеризовать тестовую выборку по целевому признаку. Таким образом, исходя из описанных выше свойств t-SNE, любой построенный в результирующем пространстве классификатор получает свойство классификатора ближайших соседей с адаптивной константой, подстраиваемой под локальную геометрию выборки. Следует отметить также, что при больших значениях μ минимизация целевой функции (2.4) требует больше шагов градиентного алгоритма. Таким образом, этот параметр следует выбирать с оглядкой на время работы программы.



(а) Student

(б) Softmax

(в) PCA

(г) Random

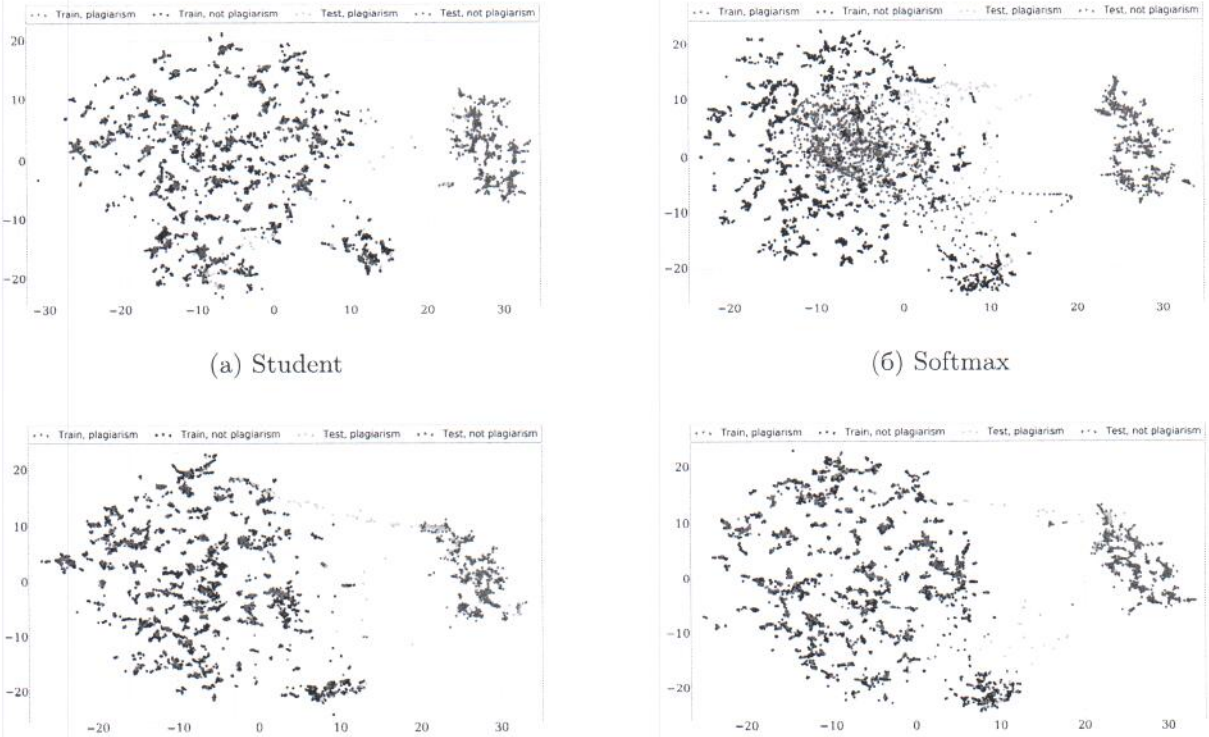
Рис. 5.11. Демонстрация вложения, выполненного предлагаемым методом при параметре $\mu = 0$.

5.4. Выделение периодов для интерпретируемой сегментации биомедицинских временных рядов

В данном разделе описаны результаты тестирования предложенного метода сегментирования квазипериодических временных рядов. Код для проведения вычислительных экспериментов доступен по адресу [?].

5.4.1. Устойчивость результатов выделения периодов к уровню шума

Устойчивость предложенного метода к уровню искажения периодического сигнала исследовалась на синтетических временных рядах с различными $\mathbf{s}$, f и a квазипериодического временного ряда X на синтетических данных. При проведении экспериментов предполагалось, что изменения формы периода и фундаментальной частоты достаточно малы, что временной ряд может быть рассмотрен как преобразованный периодический временной ряд. Точный уровень допустимой величины изменений трудноопределим, так как само понятие квазипериодического ряда слабо формализовано. Пределы изменений, в рамках которых преобразованных временной ряд еще может рассматриваться как квазипериодический, исследуются экспериментально. Для моделирования изме-



(а) Student

(б) Softmax

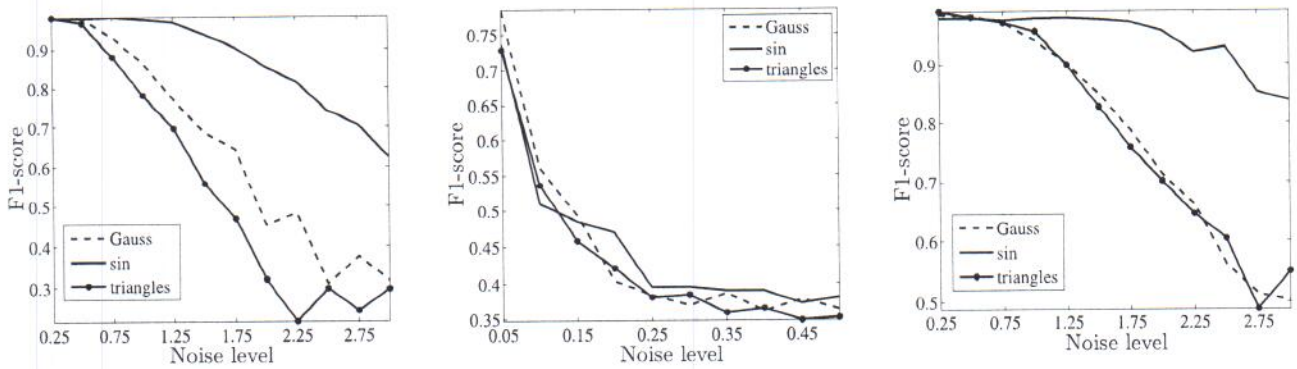
(в) PCA

(г) Random

Рис. 5.12. Демонстрация вложения, выполненного предлагаемым методом при параметре $\mu = 10$.

нений формы и периода используется зашумление амплитуды и частоты исходного сигнала. Дисперсия шума измеряет уровень искажения сигнала. Согласно (3.1), каждый синтетический квазипериодический временной ряд определяется формой $\mathbf{s}$ периода, масштабирующей функцией f и модификацией формы a. В качестве базовых форм были использованы период синуса, гауссова функция и период треугольной формы. Квазипериодические временные ряды X, использованные в экспериментах, были получены из периодического временного ряда $\tilde{X}$, $\tilde{x}(i+T) = s_i$, $1 \leq i \leq T$ с периодом T и формой $\mathbf{s}$ путем применения к нему $\tilde{X}$ следующих типов зашумления:

1. Аддитивный шум
$$x(i+T) = \tilde{x}(i) + \varepsilon, \tag{5.4}$$
где $\varepsilon \sim \mathcal{N}(0, \sigma)$. Уровень шума измерялся как $\sigma_\varepsilon / \max \tilde{X}$.
2. Зашумление аргумента
$$x(i+T) = \tilde{x}(i+\varepsilon), \tag{5.5}$$
где $\varepsilon \sim \mathcal{N}(0, \sigma)$. Уровень шума измерялся как σ_ε / T.
3. Смешанный тип зашумления:
$$x(i+T) = x_2(i) \cdot (1+\varepsilon), \tag{5.6}$$



(а) Type 1 (б) Type 2 (в) Type 3

Рис. 5.13. Зависимости метрики F_1 от уровня зашумления.

где ряд X_2 получен путем применения зашумления типа (5.5) к периодическому временному ряду X; $\varepsilon \sim \mathcal{N}(0, \sigma)$. Уровень шума измерялся как $\sigma_\varepsilon / \max \tilde{X}$.

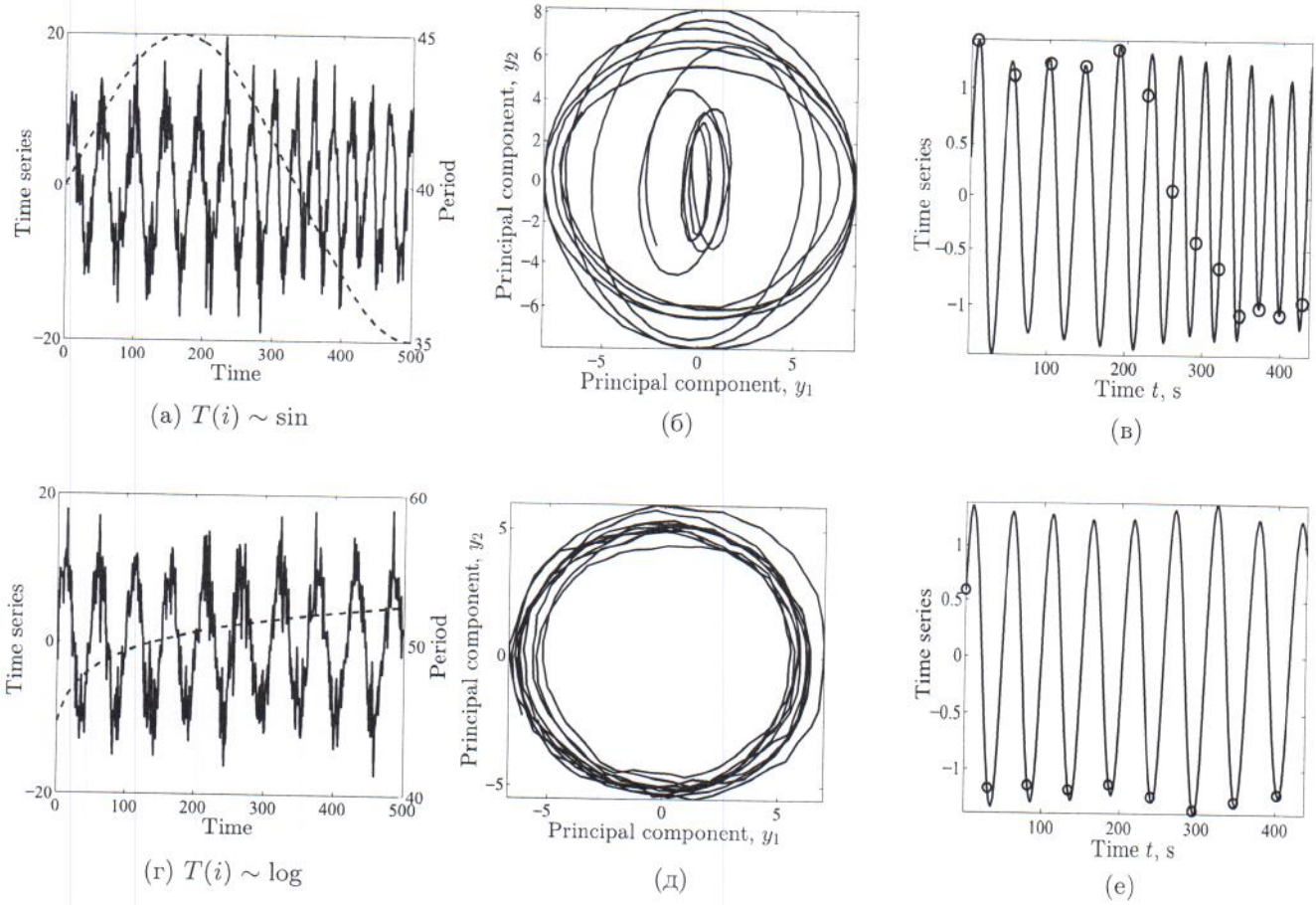
Рисунки 5.13 демонстрируют зависимости метрики F_1 от уровня шума. Каждая кривая на одном рисунке соответствует некоторой исходной форме периода $\mathbf{s}$; каждый рисунок — одному из типов зашумления (5.4), (5.5) (5.6). Каждый периодический временной ряд имел период $T = 50$.

Проведенные эксперименты позволяют оценить допустимый уровень трансформаций: качество $F_1 \gtrsim 0.7$ достигается при $\sigma_\varepsilon \lesssim 1.5 \max(X)$ для типов 1 и 3; и $\sigma_\varepsilon \lesssim 2T$ для типа 2.

5.4.2. Повышение устойчивости к изменениям периода с помощью скользящего окна

Рисунок 5.14 демонстрирует, как различные модуляции частоты влияют на результат сегментирования. На рисунках 5.14(a) и 5.14(d) отложены временные ряды, для которых период $T(i)$ — это синусоидальная и логарифмическая функции. Временной ряд отложен сплошной линией, значение мгновенного периода $T(i)$ – пунктирной линией. На рисунках 5.14(b) и 5.14(e) изображены траектории выбранных пар главных компонент; на рисунках 5.14(c) и 5.14(f) — результаты выделения периодов: временные ряды, восстановленные по выбранным парам главных компонент и выделенные границы периодов $x(i_t)$.

В первом случае 5.14(a) изменения периода более стремительно, и радиус траектории выбранной пары главных компонент, изображенной на рис. 5.14(b) меняется в более широких пределах по сравнению с радиусом соответствующей траектории для второго случая 5.14(e). Слишком большая вариабельность радиуса приводит к смещению в выделенных концах периодов $x(i_t)$, $ß_t \in \mathcal{I}_\mathcal{S}$ в случае синусоидальной функции $T(i)$ 5.14(c). Чтобы расширить пределы применимости предлагаемого метода был использован метод скользящего окна [122].



(а) $T(i) \sim \sin$ (б) (в)

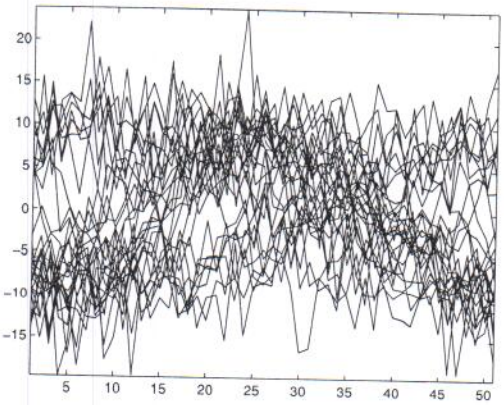
(г) $T(i) \sim \log$ (д) (е)

Рис. 5.14. Примеры сегментирования квазипериодических временных рядов с частотной модуляцией исходного периодического сигнала.

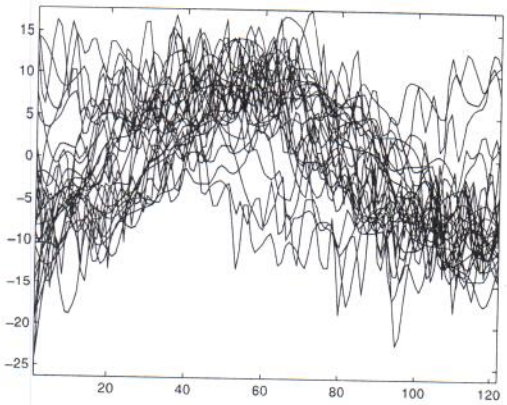
Исследуемый временной ряд разбивается на более короткие отрезки, в пределах которых период близок к постоянному и траектория главных компонент имеет более регулярную структуру. На рисунках 5.15(a) и 5.15(b) отложены выделенные периоды для квазипериодического временного ряда с синусоидальной модуляцией 5.14(a) без использования и с использованием скользящего окна. Выделенные периоды были приведены к одной длине. Среди периодов, выделенных с использованием скользящего окна 5.15(b), наблюдается меньший разброс. Значение метрики F_1 также улучшилось. Здесь для вычисления F_1 было использовано значение $\epsilon = 0$.

В вычислительных экспериментах анализ с помощью скользящего окна применялся, если длина временного ряда превышала 10 предполагаемых длин периода. В качестве оценки нулевого приближения длины периода T использовалась величина $T_0 = 1/w_{min}$. При этом число m отсчетов временного ряда должно составлять хотя бы $3T_0$. Процедура выделения периодов описывается следующим алгоритмом:

1. Для отрезка временного ряда $X = \{x(1), \ldots, x(10T_0)\}$ вычисляется новое приближение T_0 длины периода T и выделить набор $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ главных компонент с $\lambda_1 \geq \cdots \geq \lambda_N$. Параметр N траекторной матрицы $\mathbf{H}$



(а) "before": $F_1 = 0.4997$



(б) "after": $F_1 = 0.6414$

Рис. 5.15. Результаты выделения периодов для частотно-модулированного временного ряда с $T(i) \sim \sin(2\pi i/m)$ без использования скользящего окна (a) и с его использованием (b).

выбирался равным примерно $1.75T$. При выборе больших значений повышения качества выделения периодов не наблюдалось. Выбрав слишком маленькое значение N, можно упустить периодику $T > N$. Увеличение N повышает вычислительную сложность метода.

2. Выбирается пара главных компонент $(\mathbf{y}_j, \mathbf{y}_{j+1})$, соответствующих искомой периодике.
3. Множество $\mathcal{I}_S$ границ периодов определяется путем рассечения траектории выбранной пары $(\mathbf{y}_j, \mathbf{y}_{j+1})$ осью симметрии.
4. Новая оценка $T_{PC} = m/(k_j - 1)$ длины периода используется для сдвига скользящего окна.

Если длина временного ряда превышает $10T_0$, окно сдвигается на T_{PC}, и процедура повторяется для временного ряда $X = \{x(T_{PC}), \ldots, x(T_{PC} + 10T_{PC})\}$. После каждого сдвига скользящее окно покрывает новый отрезок временного ряда, для которого выделяются границы периодов и вычисляется новая оценка T_{PC}. Процедура повторяется, пока все m отсчетов временного ряда X не будут покрыты.

5.4.3. Альтернативные методы выделения периода

В данном разделе описаны результаты сравнения предложенного метода и методов выделения периодов, применяемых к периодическим временным рядам. При сравнении рассмотрены синтетические и биомедицинские временные ряды.

Метод наименьших квадратов. Для выделения периодов вычисляется

оценка коэффициентов модели

$$x(i) \approx \sum_{q=1}^{Q} A_q \cos(2\pi w_q i) + B_q \sin(2\pi w_q i)$$

методом наименьших квадратов

$$\{\mathbf{w}_{\text{LSE}},\ \mathbf{A},\ \mathbf{B}\} = \arg\min_{\mathbf{w},\,\mathbf{A},\,\mathbf{B}} \sum_{i=1}^{m} \Bigg(x(i) - \sum_{q=1}^{Q} \big(A_q \cos(2\pi w_q i) + B_q \sin(2\pi w_q i) \big) \Bigg)^2 . \tag{5.7}$$

В работе [167] показано, что при $0 < w < 0.5$ задача (5.7) асимптотически эквивалентна задаче

$$\{w_{\text{LSE}},\ A,\ B\} = \arg\min_{w,\,A,\,B} \left\{ \sum_{i=1}^{m} x^2(i) - 2\sum_{i=1}^{m} x(i)\big(A\cos(2\pi w i) + B\sin(2\pi w i)\big) + \frac{2}{m}(A^2 + B^2) \right\}, \tag{5.8}$$

что позволяет использовать

$$\hat{A}(w) = \frac{2}{m}\sum_{i=1}^{m} x(i)\cos(2\pi w i), \quad \hat{B}(w) = \frac{2}{m}\sum_{i=1}^{m} x \sin(2\pi w i)$$

для оценок $\hat{A}$ и $\hat{B}$ коэффициентов A и B.

Максимизация периодограмы. Переписав функционал из задачи (5.8) в виде

$$\sum_{i=1}^{m} x^2(i) - 2\sum_{i=1}^{m}\big(A\cos(2\pi w i) + B\sin(2\pi w i)\big) + \frac{2}{m}(A^2 + B^2) = \\ = \sum_{i=1}^{m} x^2(i) - \frac{2}{m}\left| x(i) e^{-\mathrm{i}2\pi w i} \right|,$$

получаем, что задача (5.8) эквивалентна максимизации периодограмы [114] $\frac{2}{m}\left| x(i) e^{-\mathrm{i}wi} \right|$. Поэтому в данной работе рассматриваются только оценки, полученные с помощью метода наименьших квадратов согласно (5.8).

Оценка максимальной кросс-корреляции. Временной ряд разбивается на сегменты $S_t = \{x(i_{T(t-1)+1}, \ldots, x(i_{T\cdot t}))\}$ одинаковой длины T для всех значений T в некотором заданном диапазоне $[T_{\min}, T_{\max}]$. Для каждого T вычисляется

линейный средний коэффициент корреляции Пирсона между последовательными сегментами

$$\mathrm{Corr}(X,T) = \frac{1}{S-1}\sum_{t=1}^{S-1}\rho(S_t, S_{t+1}).$$

Оценка максимальной кросскорреляции определяется как

$$T_{Corr} = \arg\max_{T} \mathrm{Corr}(X,T).$$

В экспериментах были выбраны значения $|T_{\max}| = m/4$ и $T_{\min} = 3$.

5.4.4. Сравнение результатов предложенного метода выделения периодов с альтернативными

Синтетические временные ряды для сравнения генерировались согласно (5.4), (5.5) и (5.6) с уровнем шума, равным 1 для всех типов временных рядов.

Выборка биомедицинских временных рядов записи ECG и временные ряды кровяного давления [111] и акселерометрические временные ряды. Акселерометрические временные ряды измерялись смартфоном при ходьбе, легком беге и беге вприпрыжку с частотой сэмплирования в 40Гц. Для анализа использовались временные ряды амплитуды ускорения:

$$x(i) = \sqrt{a_x^2(i) + a_y^2(i) + a_z^2(i)}, \quad i = 1, \ldots, m.$$

Подробное описание временных рядов ECG и кровяного давления можно найти в [151] и [152].

Таблица 5.5. Структура выборки.

Synthetic time series: total $N_{ts} = 240$				
	N_{ts}	m	S	$100\sigma_T/T$
Type 1	80	1500	30	0
Type 2	80	1500	30	7.2
Type 3	80	1500	30	6.8
Real-life time series: total $N_{ts} = 560$				
	N_{ts}	m	S	$100\sigma_T/T$
Walking	10	3000	18	5.7
Jogging	12	3000	20	5.2
Leaping	8	3000	15	6.9
ECG	350	6000	70	5.6
Blood pressure	180	7500	60	4.2

В таблице 5.5 приведены сводные данные по использованным временным рядам, включающие число N_{ts} временных рядов в каждом наборе, среднюю

длину m временного ряда, среднее число S периодов и показатель $100\sigma_T/T$ разброса значений длины периода T, где

$$\sigma_T = \frac{1}{S}\sum_{t=1}^{S}\left|(i^*_{t+1} - i^*_t) - (i^*_t - i^*_{t-1})\right|.$$

В отличие от дисперсии, σ_T измеряет вариабельность в длине последовательных сегментов. Значения $100\sigma_T/T$ для реальных данных примерно одинаковы.

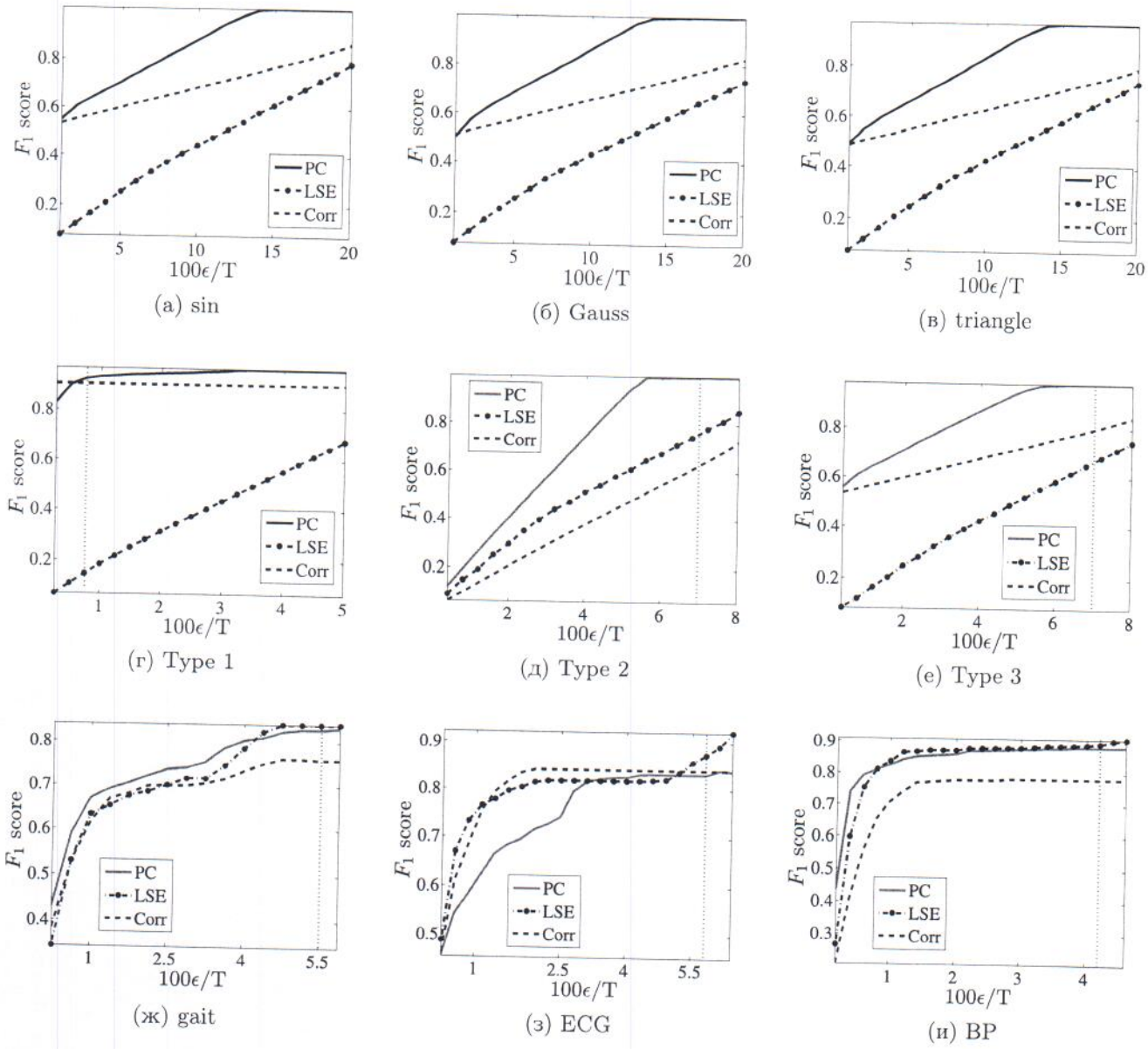
В таблице 5.6 приведены значения метрики качества F_1

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{and Recall} = \frac{TP}{TP + FN}$$

для сравниваемых методов на каждом из наборов данных. Чем выше значение F_1, тем более точно выделены периоды. Значения в таблице 5.6 получены усреднением по всем временным рядам из конкретного набора. Результаты предлагаемого метода обозначены как "PC", а альтернативные — как "LSE" для оценки методом наименьших квадратов и "Corr" метода максимальной кросскорреляции.

Сравнение результатов выделения периодов требует знания истинных границ периодов, которые для реальных временных рядов могут быть не определены, либо определены с некоторой погрешностью. Значения F_1 зависят от параметра ϵ, который определяет допустимую погрешность в отклонении i_t от i^*_t, при которой i_t все еще считается правильным концом периода. Рисунок 5.16 иллюстрирует как метрика F_1 увеличивается с ростом ϵ. На рисунках 5.16(a–f) показаны зависимости $F_1(\epsilon)$ для синтетических временных рядов с различной базовой формой сегмента (5.16(a, b, c)) и типами зашумленности (5.16 (d, e, f)). Зависимости на рисунках 5.16(g, h, i) соответствуют акселерометрическим временным рядам, временным рядам ЭКГ и кровяного давления. По горизонтали отложены значения параметра ϵ, выраженные в $100/T$ для соотнесения со значениями $100\sigma_T/T$, приведенными в таблице 5.5.

При увеличении ϵ больше точек i_t сегментации считаются правильно распознанными. При этом число FP ложных положительных срабатываний алгоритма не меняется, а число FN ложных отрицательных срабатываний может только уменьшиться настолько, насколько увеличивает число TP правильных положительных срабатываний. Поэтому сумма $TP + FN + FP$ постоянна. Быстрая сходимость $F_1(\epsilon)$ к 1 свидетельствует о низких значениях $FN + FP$. Поведение $F_1(\epsilon)$, близкое к линейному, свидетельствует о том, что значение $TP + FN + FP$ велико по сравнению с TP. Таким образом, алгоритмы с большей площадью под кривой $F_1(\epsilon)$ более предпочтительны. Рисунок 5.16 демонстрирует, что качество алгоритмов сильнее зависит от типа зашумленности, чем от формы сигнала. Поэтому в таблице 5.6 результаты для синтетических временных рядов агрегированы по формам базового сегмента.



(а) sin (б) Gauss (в) triangle

(г) Type 1 (д) Type 2 (е) Type 3

(ж) gait (з) ECG (и) BP

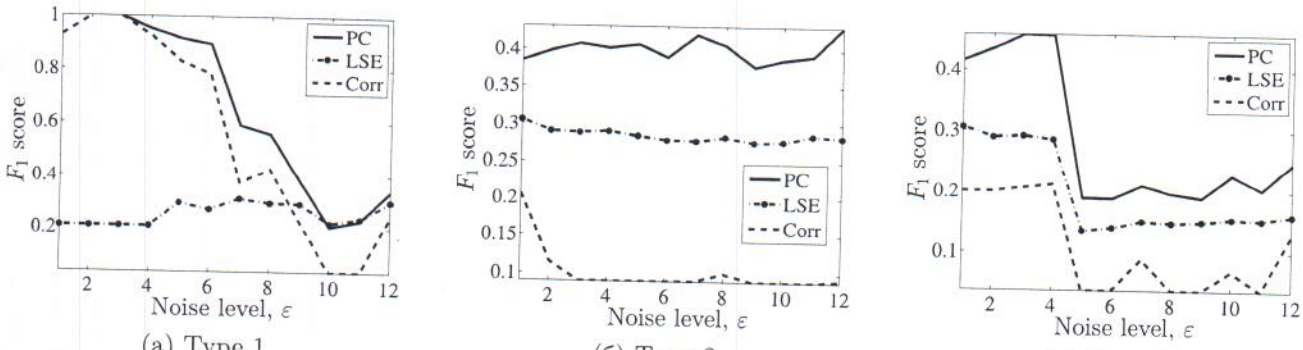
Рис. 5.16. Зависимость метрики F_1 от величины ϵ допустимой погрешности.

Конкретные значения ϵ, для которых приведены результаты 5.6 выбраны с учетом частоты сэмплирования, так что $\epsilon \approx \sigma_T$. При таких условиях алгоритмы, выделяющие периоды постоянной длины, должны производить адекватные результаты при наличии неточности в определении i_t^*, но будут давать плохие результаты при выделении периодов с частотной модуляцией. Выбранные значения ϵ отмечены вертикальными линиями на рисунках 5.16 (d-g).

На рисунке 5.17 изображены зависимости метрики F_1 от уровня шума 5.4.1. для предложенного ("PC") и альтернативных ("LSE", "Corr") методов.

Таблица 5.6. Усредненные метрики качества сравниваемых методов: F_1, точность и полнота.

	F_1			Точность			Полнота		
	PC	LSE	Corr	PC	LSE	Corr	PC	LSE	Corr
Type1	**0.9220**	0.1485	0.9039	**0.9623**	0.7861	0.6120	**0.9561**	0.7125	0.8966
Type2	**0.9953**	0.7825	0.6495	**0.9022**	0.7306	0.6070	0.9012	0.7094	**0.9930**
Type3	**0.9803**	0.6508	0.7891	**0.9466**	0.7167	0.6941	0.9273	0.6396	**0.9569**
gait	0.8287	**0.8400**	0.7759	0.8014	**0.9311**	0.5978	0.7710	0.6649	**0.7510**
ECG	0.8328	**0.8745**	0.8417	0.8169	**0.9500**	0.8577	0.9292	0.8167	**1.0000**
BP	0.8836	**0.8958**	0.7820	0.9622	**0.9722**	0.9336	0.9001	0.8225	**0.9492**



(а) Type 1 (б) Type 2 (в) Type 3

Рис. 5.17. Зависимость качества F_1 от уровня шума, использованного при генерации синтетических временных рядов.

5.5. Оценка объема выборки в задаче прогнозирования

Предложенный в работе критерий для решения задачи двух выборок применялся для решения задачи построения агрегированных прогнозов объемов железнодорожных грузоперевозок. Для получения агрегированных прогнозов требуется кластеризовать временные ряды таким образом, чтобы распределения временных рядов внутри кластера совпадали. При решении задачи кластеризации требуется оценить близость между временными рядами, исходя из их эмпирических распределений. Решена задача кластеризации временных рядов на основе матрицы парных расстояний между ними. Данные о перевозках

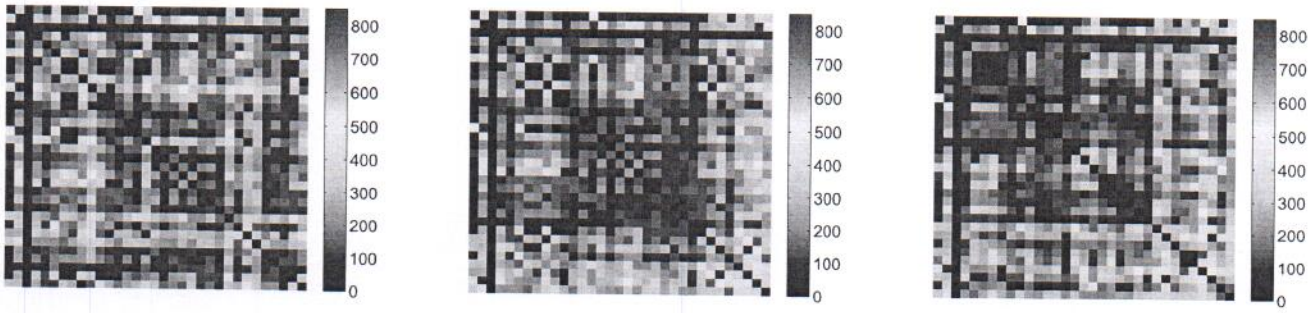


Рис. 5.18. Симметризованные матрицы попарных расстояний Кульбака–Лейблера между временными рядами для различных групп грузов: до кластеризации и после для $K = 5$ и $K = 10$ кластеров.

включают даты отправления, даты прибытия, станции внутри железнодорожной ветки и группы грузов. Для исследования были выбраны временные ряды с весами вагонов, нагруженных различными группами грузов, агрегированные по станциям. Вначале из рассмотрения были исключены ряды, содержащие менее 50 отсчетов времени. Матрицы (4.6) D симметризованных расстояний Кульбака–Лейблера для набора исследуемых временных рядов изображены на рисунке 5.18. Решая задачу кластеризации (4.3), мы стремимся привести матрицу D к блочному виду. В таблице 5.7 (слева) показано, как именно перевозимые группы грузов были разбиты на кластеры для случая $K = 5$. Ряды, содержащие менее 50-ти отсчетов, последовательно присоединялись к каждому из рядов, содержавших более 50-ти отсчетов. Затем для объединенного ряда и исходного ряда длины более 50 проверялась гипотеза о принадлежности рядов одному распределению. В результате было выполнено слияние рядов «Поваренная соль» и «Продукты промышленного потребления», а также «Флюсы» и «Шлаки гранулированные» с временным рядом «Нефть и нефтепродукты». Затем снова была выполнена кластеризация, результаты которой занесены в таблицу 5.7 (справа).

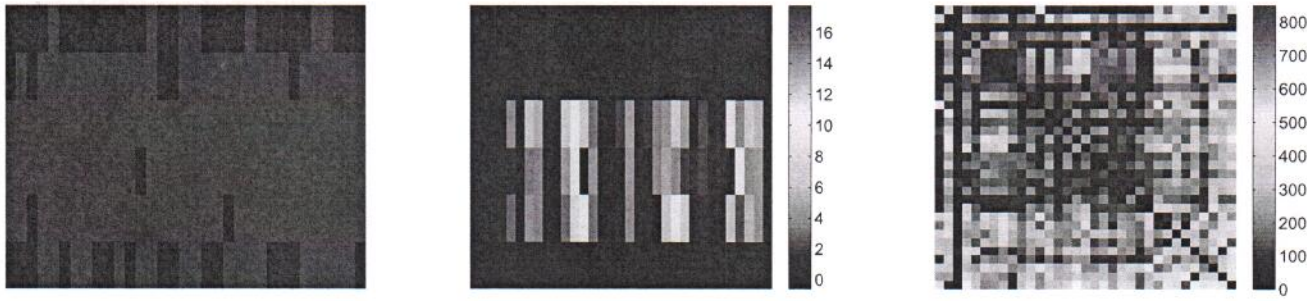


Рис. 5.19. Результат проверки гипотезы о принадлежности исходного временного ряда и временного ряда, получаемого присоединением к нему некоторого более короткого ряда, к одному распределению. Синий — гипотеза отвергается, красный - принимается.

Таблица 5.7. Группы грузов.

1 — Каменный уголь, 2 — Кокс, 3 — Нефть и нефтепродукты, 7 — Руда железная и марганцевая, 8 — Руда цветная и серное сырье, 9 — Черные металлы, 10 — Метизы и оборудование, 11 — Металлические конструкции, 14 — Сельскохозяйственные машины, 15 — Автомобили, 16 — Цветные металлы, изделия из них и лом цв. металлов, 17 — Химические и минеральные удобрения, 18 — Химикаты и сода, 20 — Промышленное сырье и формовочные материалы, 22 — Огнеупоры, 23 — Цемент, 25 — Сахар, 26 — Мясо и масло животное, 27 — Рыба, 28 — Картофель, овощи и фрукты, 36 — Комбикорма, 38 — Жмыхи, 39 — Бумага, 42 — Грузы в контейнерах	1 — Каменный уголь, 2 — Кокс, 3 — Нефть и нефтепродукты, **6 —Флюсы**, **21 — Шлаки гранулированные**, 7 — Руда железная и марганцевая, 8 — Руда цветная и серное сырье, 10 — Метизы и оборудование, 11 — Металлические конструкции, 12 — Метизы, 14 — Сельскохозяйственные машины, 15 — Автомобили, 16 — Цветные металлы, изделия из них и лом цв. металлов, 17 — Химические и минеральные удобрения, 18 — Химикаты и сода, 20 — Промышленное сырье и формовочные материалы, 22 — Огнеупоры, 23 — Цемент, 25 — Сахар, 26 — Мясо и масло животное, 27 — Рыба, 28 — Картофель, овощи и фрукты, 36 — Комбикорма, 38 — Жмыхи, 39 — Бумага, 42 — Грузы в контейнерах
43 — Остальные и сборные грузы, 19 — Строительные грузы	9 — Черные металлы, 12 — Метизы
31 — Промышленные товары народного потребления, 34 — Зерно	43 — Остальные и сборные грузы, 19 — Строительные грузы, 35 — Продукты перемола
13 — Лом черных металлов	31 — Промышленные товары народного потребления, **29 — Поваренная соль**, 34 — Зерно
35 — Продукты перемола	13 — Лом черных металлов

Заключение

Предложены методы выбора признаков и построения выборки оптимального состава при решении задачи выбора модели прогнозирования, классификации и декодирования временных рядов.

Во введении обоснована актуальность диссертационной работы, сформулированы цели и методы исследования, поставлены основные задачи, обоснована научная новизна, теоретическая и практическая значимость полученных результатов, приведено краткое содержание работы по главам.

В главе 1 приводится формальная постановка задач декодирования электрокортикорамм и задачи выбора признаков. Вводятся основные определения и обозначения, функции качества восстановления траектории.

В главах 1 и 2 описаны предлагаемые в работе методы выбора признаков и снижения размерности. Сформулированы теоремы о свойствах предложенных алгоритмов.

В главе 1 предложена модификация фильтрационного метода QPFS [13]. Фильтрационные методы ранжируют признаки вне зависимости от модели, поэтому они более эффективны чем методы, требующие настройки модели. Фильтрационные методы вычислительно эффективны [15] за счет разделения этапов выбора признаков и построения модели. С другой стороны, так как признаки оцениваются независимо, фильтрационные методы склонны к выбору коррелирующих признаков. Особенностью метода QPFS является способность учитывать отношения между признаками, которой зачастую лишены другие фильтрационные методы. Выбор признаков с помощью квадратичного программирования — это гибкий и эффективный подход, который позволяет выбирать наиболее релевантные признаки при наличии значительного числа мультикоррелирующих признаков. Метод QPFS основан на решении задачи квадратичного программирования, одновременно минимизирующей сходство между признаками, задаваемое квадратичной формой, и максимизирующей релевантность признаков, задаваемую линейным членом. Метод QPFS позволяет сбалансировать сходство между признаками и релевантность выбранных признаков задаче регрессии, не перебирая всех возможных комбинаций признаков. Предложен метод выбора признаков, учитывающий многоиндексную структуру данных. Проанализирована сложность предложенного алгоритма. Доказана сходимость метода.

Предложен метод, который может быть использован для построения прогнозирующих моделей при разработке НКИ. Декодирование кортикальной активности человеческого мозга является центральной проблемой нейроинженерии. Рассматривается проблема реконструкции трехмерных движений из кортикальной активности. Для построения адекватной, но вычислительно простой модели, предлагается тензорная формулировка метода QPFS. Предложенная модификация предназначена для многоиндексных представлений данных. Многоиндексная структура используется для формулировки модифицированного метода QPFS, применимого даже для больших размерах выборки. Предложенная модификация метода QPFS применяется для построения регрессионной модели для восстановления траектории руки по электрокортикографическим данным. Качество линейной регрессии на основе выбранных QPFS признаков сравнимо с качеством регрессии PLS. В проведенных экспериментах многоиндексная модификация QPFS продемонстрировала результаты, сравнимые по качеству с исходным QPFS. При этом было показано, что предложенная модификация вычислительно более эффективна за счет использования многоиндексной структуры.

В главе 2 предложен метод вложения признаков в пространство меньшей размерности, сохраняющий локальную структуру близости между объектами в исходном пространстве. Решена проблема не просмотренных объектов при выполнении вложений новых объектов. Предложен способ учета разметки при выполнении вложения для решения задачи классификации классификации. Проанализирована зависимость качества классификации с применением описанного метода от его параметров, экспериментально обосновано использование поэтапного обучающего вложения. Продемонстрирована устойчивость алгоритма к введенным параметрам размера начальной части S_s и максимального размера блоков S_b, что облегчает его использование на практике. Проанализировано признаковое пространство задачи внутреннего плагиата. Проиллюстрированы свойства предложенного алгоритма относительно данных задачи внутреннего плагиата.

В главе 3 предложен метод выделения периодов квазипериодических временных рядов. Метод основан на анализе фазовых траекторий главных компонент ряда, соответствующих периодической составляющей ряда.

В главе 4 предложен критерий информативности объекта для оценки достаточного объема выборки. Предложен способ оценки достаточного объема выборки. Объем выборки называется достаточным, если вероятностное распределение параметров модели, оцениваемое по выборке данного объема, устойчиво и перестает меняться при расширении выборки. Введен критерий значимости изменения функции распределения параметров модели. Применимость предложенного способа оценки исследована в случаях различных вероятностных моделей данных. Приведены рекомендации по практическому использованию разработанного метода оценки объема выборки.

В главе 5 продемонстрировано применение предложенных методов к прикладным задачам сегментирования квазипериодических временных рядов, декодирования электрокортикограмм, задачи агрегирования и прогнозирования временных рядов грузовых железнодорожных перевозок.

Список обозначений

D		регрессионная выборка
$\mathbf{X}$	$=$	матрица плана
$[\mathbf{x_1}^\intercal, \ldots, \mathbf{x}_m^\intercal]^T$		
$\mathbf{x}_m \in \mathbf{R}^n$		вектор-строка матрицы $\mathbf{X}$, используется для обозначения признаков объекта
$\boldsymbol{\chi}\mathbb{R}^M$		вектор-столбец матрицы плана
$\underline{\mathbf{X}} \in \mathbb{R}^{n_1 \times n_D}$		D-индексная матрица
$\mathbf{y} \in \mathbb{R}^M, \mathbf{Y} \in \mathbb{R}^{M \times N_y}$		вектор или матрица целевых переменных
$\mathbf{w}, \mathcal{W}$		параметры линейного классификатора
$\mathbf{a} \circ \mathbf{b} \in \mathbb{R}^{n_1 \times n_2}$		тензорное произведение векторов $\mathbf{a} \in \mathbb{R}^{n_1}$ и $\mathbf{b} \in \mathbb{R}^{n_2}$: $[\mathbf{a} \circ \mathbf{b}]_{ij} = a_i b_j$
$\underline{\mathbf{A}} \times_1 \mathbf{B} \in \mathbb{R}^{m \times n_2 \times n_3}$		скалярное умножение тензора $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ на матрицу $\mathbf{B} \in \mathbb{R}^{m \times n_1}$: $[\underline{\mathbf{A}} \times_1 \mathbf{B}]_{ijk} = \sum_{i'} a_{i'jk} b_{ii'}$
$\underline{\mathbf{A}} * \underline{\mathbf{B}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$		поэлементное произведение многоиндексных матриц $\underline{\mathbf{A}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ и $\underline{\mathbf{B}} \in \mathbb{R}^{n_1 \times n_2 \times n_3}$: $[\underline{\mathbf{A}} * \underline{\mathbf{B}}]_{ijk} = a_{ijk} b_{ikj}$
$\mathbf{H}$		траекторная матрица временного ряда
$D_{\mathrm{KL}}(p\|\|p')$	$\equiv$	дивергенция Кульбака-Лейблера
$\mathrm{KL}(p\|\|p')$		
$\chi^2(p\|\|p')$		χ^2-дивергенция
U^ξ		критическая область случайной величины ξ

Список иллюстраций

Список таблиц

Литература

1. K. Vorontsov. Combinatorial probability and the tightness of generalization bounds. Pattern Recognition and Image Analysis, 18(2):243–259, 2008.
2. George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In ICML, volume 94, pages 121–129, 1994.
3. Il-Gyo Chong, Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems, 78(1–2):103 – 112, 2005.
4. Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B, 58:267–288, 1994.
5. D. Madigan, G. Rideway. Discussion of least square regression. least angle regression. The Annals of Statistics, 32(2):465–469, 2004.
6. N. I. Rashwan M. El-Dereny. Solving multicollinearity problem using ridge regression models. Int. Journal of Contemp. Math. Sciences, 6:585 — 600, 2011.
7. Hui Zou, Trevor Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B, 67:301–320, 2005.
8. Frank E. Harrell. Regression Modeling Strategies: With Applications to Linear M Springer, 2001.
9. Laurent Bougrain, Nanying Liang. Band-specific features improve finger flexion prediction from ECoG. Jornadas Argentinas sobre Interfaces Cerebro Computadora – JAICC 2009, 2009.
10. David T. Bundy, Mrinal Pahwa, Nicholas Szrama, Eric C. Leuthardt. Decoding three-dimensional reaching movements using electrocorticographic signals in humans. Journal Of Neural Engineering, 13:026021, 2016.
11. E. Sellers, T. Vaughan, J. Wolpaw. A brain-computer interface for long-term independent home use. Amyotrophic lateral sclerosis, 11:449–455, 2010.
12. K. Shimoda, Y. Nagasaka, Z.C. Chao, N. Fujii. Decoding continuous three-dimensional hand trajectories from epidural electrocorticographic signals in Japanese macaques. Journal of Neural Engeneering, (9):036015, 2012. http://neurotycho.org/food-tracking-task.
13. Aleksandr Katrutsa, Vadim Strijov. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Systems with Applications, 76(1–11), 2017.
14. Huan Liu, Hiroshi Motoda, Rudy Setiono, Zheng Zhao. Feature selection: An ever evolving frontier in data mining. In JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining., 2010.
15. A.M. Katrutsa, V.V. Strijov. Stresstest procedure for feature selection algorithms. Chemometrics and Intelligent Laboratory Systems, 142:172–183, 2015.

16. Bokai Cao, Lifang He, Xiangnan Kong, Philip S. Yu, Zhifeng Hao, Ann B. Ragin. Tensor-based multi-view feature selection with applications to brain diseases. Proceedings. IEEE International Conference on Data Mining, pages 40–49, 2014.
17. Aaron Smalter, Jun Huan, Gerald Lushington. Feature selection in the tensor product feature space. Proceedings. IEEE International Conference on Data Mining, pages 1004–1009, 2009.
18. Jie Li, Liqing Zhang. Regularized tensor discriminant analysis for single trial EEG classification in BCI. Pattern Recognition Letters, 2009.
19. Tae-Kyun Kim, Shu-Fai Wong, Roberto Cipolla. Tensor canonical correlation analysis for action classification. IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07, 2007.
20. Andrey Eliseyev, Cecile Moro, Thomas Costecalde, Napoleon Torres, Sadok Gharbi, Corinne Mestais, Alim Louis Benabid, Tatiana Aksenova. Iterative N-way partial least squares for a binary self-paced brain-computer interface in freely moving animals. Journal of Neural Engineering, 8(4):046012, 2011.
21. Andrey Eliseyev and Tetiana Aksenova. Penalized multi-way partial least squares for smooth trajectory decoding from lectrocorticographic (ECoG) recording. volume 11, page e0154878, 2016.
22. Charles Fefferman, Sanjoy Mitter, Hariharan Narayanan. Testing the manifold hypothesis. Journal of the American Mathematical Society, 29(4):983–1049, 2013.
23. Laurens van der Maaten. Visualizing data using t-SNE. Journal of Machine Learning Research, 9:2579–2605, 2008.
24. H. Narayanan, S. Mitter. Sample complexity of testing the manifold hypothesis. Advances in neural information processing systems, pages 1786–1794, 2010.
25. Julian Brooke, Adam Hammond, Graeme Hirst. Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. Proceedings of the 1st NAACL-HLT Workshop on Computational Linguistics for Literature, pages 26–35, 2012. Association for Computational LinguisticsStroudsburg, PA, USA.
26. Julian Brooke, Graeme Hirst. Paragraph clustering for intrinsic plagiarism detection using a stylistic vector-space model with extrinsic features. Notebook for PAN at CLEF 2012, 2012.
27. A. N. Gorban, B. Kegl, D. Wunsch, A. Y. Zinovyev. Principal manifolds for data visualisation and dimension reduction. Lecture Notes in Computational Science and Engineering, page 58, 2008.
28. J. Tenenbaum, V. Silva, J. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(5500):2319–2323, 2000.
29. M. Belkin, P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. MIT Press, 14(14):585–591, 2001.

30. S. Roweis, L. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(5500):2323–2326, 2000.
31. D. Donoho, C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Proceedings of the National Academy of Sciences, 100(10):5591–5596, 2003.
32. Z. Zhang, H. Zha. Principal manifolds and nonlinear dimension reduction via local tangent space alignment. SIAM Journal on Scientific Computing, 26(1):313–338, 2005.
33. K. Weinberger, L. Saul. Unsupervised learning of image manifolds by semidefinite programming. International Journal of Computer Vision, 70(1):77–90, 2006.
34. Changyou Chen, Junping Zhang, R. Fleischer. Distance approximating dimension reduction of riemannian manifolds. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40(1):208–217, 2010.
35. Laurens van der Maaten. Learning a parametric embedding by preserving local structure. Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, PMLR, 5:384–391, 2009.
36. Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. Journal of Machine Learning Research, 15(1):3221–3245, 2014.
37. J. Kubánek, K. J. Miller, J. G. Ojemann, J. R. Wolpaw, G. Schalk. Decoding flexion of individual fingers using electrocorticographic signals in humans. Journal Of Neural Engineering, 6(6):066001, 2009.
38. C. Chen, D. Shin, H. Watanabe, Y. Nakanishi, H. Kambara. Prediction of hand trajectory from electrocorticography signals in primary motor cortex. PLoS ONE 8(12), page e83534, 2013.
39. Rémi Flamary, Alain Rakotomamonjy. Decoding finger movements from ecog signals using switching linear models. Frontiers in Neuroscience, 2012.
40. Hansheng Wang, Shain-Chung Chow. Sample size calculation for comparing proportions. Wiley Encyclopedia of Clinical Trials, 2007.
41. S. G. Self, R. H. Mauritsen. Power/sample size calculations for generalized linear models. Biometrics, 44(1):79–86, 1988.
42. S. G. Self, R. H. Mauritsen, J. Ohara. Power calculations for likelihood ratio tests in generalized linear models. Biometrics, 48(1):31–39, 1992.
43. G. Shieh. On power and sample size calculations for likelihood ratio tests in generalized linear models. Biometrics, 56(4):1192–1196, 2000.
44. E. Demidenko. Sample size determination for logistic regression revisited. Statist. Med., 26:3385–3397, 2007.
45. P. J. Bickel, K. A. Doksum. Mathematical Statistics (2nd edn). Prentice-Hall: Upper Saddle River, N. J., 2001.
46. C. R. Rao. Linear Statistical Inference and its Applications (2nd edn). Wiley: New York, 1973.

47. R. J. Serfling. Approximation Theorems of Mathematical Statistics. Wiley: New York, 1980.
48. K. Kelley, E. Maxwell. Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. Psychological Method, 8(3):305–321, 2003.
49. Maher Qumsiyeh. Using the bootstrap for estimating the sample size in statistical experiments. Journal of modern applied statistical methods, 12(1):45–53, 2013.
50. D. V. Lindley. The choice of sample size. The Statistician, 46:129–138, 1997.
51. D. B. Rubin, H. S. Stern. Sample size determination using posterior predictive distributions. Sankhya : The Indian Journal of Statistics Special Issue on Bayesian Analysis, 60:161–175, 1998.
52. Fei Wang, Alan E. Gelfand. A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. Statistical Science, 17(2):193–208, 2002.
53. L. Joseph, D. B. Wolfson, R. du Berger. Sample size calculations for binomial proportions via highest posterior density intervals. Statistician, 44:143–154, 1995.
54. L. Joseph, R. du Berger, P. Be'lisle. Bayesian and mixed bayesian likelihood criteria for sample size determination. Statististical Medicine, 16:769–781, 1997.
55. Qibin Zhao, Guoxu Zhou, TΓjlay Adali, Andrzej Cichocki. Kernelization of tensor-based models for multiway data analysis. IEEE Signal Processing Magazine, 30(4):137–148, 2013.
56. Yijun Wang, Shangkai Gao, Xiaornog Gao. Common spatial pattern method for channel selection in motor imagery based brain-computer interface. Engineering in Medicine and Biology 27th Annual Conference, (5392–5395), 2005.
57. Andrzej Cichocki. Tensor decompositions: A new concept in brain data analysis? arXiv:1305.0395, 2013.
58. A. L. Gibbs, F. E. Su. On choosing and bounding probability metrics. International Statistical Review, 70(3):419–435, 2002.
59. U. Keich, N. Nagarajan. A fast and numerically robust method for exact multinomial goodness-of-fit test. Journal of Computational and Graphical Statistics, 15(15):779–802, 2006.
60. S. M. Ali, S. D. Silvey. A general class of coefficients of divergence of a distribution from another. Journal of Royal Statistical Soiety. Series B (Methodoligical), 1(28):131–142, 1966.
61. S. M. Ali, S. D. Silvey. A general class of coefficients of divergence of a distribution from another. Journal of Royal Statistical Soiety. Series B (Methodoligical), 1(28):131–142, 1966.
62. А. П. Мотренко, В. В. Стрижов. Построение агрегированных прогнозов объемов железнодорожных грузоперевозок. Информатика и ее применения, 8(2):86–97, 2014.

63. A. Motrenko. Sample size estimation in classification problems. International Digital Processing Conference, pages 34–35, 2016.
64. А. П. Мотренко. Оценка объема выборки в задачах классификации. In Математические методы распознавания образове, pages 28–29, 2015.
65. A. Motrenko. Bayesian sample size estimation for patient classification survey. 20th Conference of the International Federation of Operational Research, pages 99–99, 2014.
66. A. Motrenko. Small cvd sample set classification: generative versus discriminative. 26th European conference on operation research, pages 90–90, 2013.
67. A. Motrenko. Multiclass classification of cardio-vascular disease patients with sample size estimation. 25th European conference on operation research, pages 172–172, 2012.
68. А.П. Мотренко, В.В. Стрижов. Многоклассовая логистическая регрессия. Известия ТулГУ, (1):153–162, 2012.
69. A. Motrenko, V. Strijov, G.-W. Weber. Bayesian sample size estimation for logistic regression. Journal of Computational and Applied Mathematics, 255:743–752, 2014.
70. A. Motrenko, V. Strijov. Extracting fundamental periods to segment biomedical signals. Journal of Biomedical and Health Informatics, 20(6):1466–1476, 2016.
71. A. Motrenko, V. Strijov. Multi-way feature selection for ECoG-based brain-computer interface. Expert Systems with Applications, 2018.
72. Y. Bengio, A. Courville, P. Vincent. Representation learning: A review and new perspectives. arXiv:1206.5538, 2012.
73. S. Xiang, X. Shen, J. Ye. Efficient nonconvex sparse group feature selection via continuous and discrete optimization. Artificial Intelligence, 224:28–50, 2015.
74. James C. Bezdek, Richard J. Hathaway. Convergence of alternating optimization. Neural, Parallel and Scientific Computations, 111(4):351–368, 2003.
75. L. Bottou. Stochastic gradient descent tricks. Neural networks: Tricks of the trade, pages 421–436, 2012.
76. André Gensler, Bernhard Sick. Novel criteria to measure performance of time series segmentation techniques. Proceedings of the LWA 2014 Workshops: KDML, IR, FGWM, 2014.
77. N. Golyandina, V. Nekrutkin, A. Zhigljavsky. Analysis of Time Series Structure: SSA and related techniques. Chapman and Hall/CRC, 2001.
78. Th. Alexandrov, N. Golyandina. Automatic extraction and forecast of time series cyclic components within the framework of ssa. Proceedings of the 5th Workshop on Simulation, pages 45–50, 2005.
79. R. Vautard, M. Ghil. Singular singular spectrum analysis in nonlinear dynemics, with applications to paleoclimatic time series. Nonlinear Phenomena, 35(3):395–424, 1989.

80. А.Н. Каркищенко, В. Б. Мнухин. Восстановление симметричности точек на изображениях объектов с отражательной симметрией. Машинное обучение и анализ данных, 1(5):621–631, 2013.
81. A. N. Karkishchenko, V. B. Mnukhin. Reflective symmetrization of feature points in images. 11th Conference (International) on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013) Proceedings, pages 209–212, 2013.
82. I. Csiszar, P. Shields. Information theory and statistics: A tutorial. Foundations and Trend in Communications and Information Theory, 4:417–528, 2004.
83. A. Motrenko. Feature Selection for decoding ECoG time series. https://github.com/Anastasia874/ECoGFeatureSelection
84. R. Bro. Multiway calibration. Multilinear PLS. Chemometrics, 10(1):47–61, 1996.
85. Kai Joshua Miller. A library of human electrocorticographic data and analyses. Stanford Digital Repository. Available at: https://purl.stanford.edu/zk881ps0522, 2016.
86. http://svn.code.sf.net/p/mlalgorithms/code/Group874/Motrenko2014TSsegmnet
87. C. J. Adcock. Sample size determination: A review. 46:261–283, 1997.
88. S. Amari, N. Murata, K. R. Muller, M. Finke, H.H. Yang. Asymptotic statistical theory of overtraining and cross-validation. IEEE Transactions on Neural Networks, 8(5):985–996, 1997.
89. S. Amiri, R. Fazel-Rezai, V. Asadpour. A review of hybrid brain-computer interface systems. Advances in Human-Computer Interaction, 2013:187024, 2013.
90. J. Baek, G. Lee, W. Park, B. J. Yun. Accelerometer signal processing for user activity detection, in knowledge-based intelligent information and engineering systems. Knowledge-Based Intelligent Information and Engineering Systems. Lecture Notes in Computer Science, 3215:610–617, 2004.
91. L. Bao, S. Intille. Activity recognition from user-annotated acceleration data. In: Proceedings of the international conference on pervasive computing and communications, 2004.
92. J. Barbic, A. Safonova, Jia-Yu Pan, C. Faloutsos. Segmenting motion capture data into distinct behaviors. GI '04 Proceedings of Graphics Interface 2004, pages 185–194, 2004.
93. P. Barralon, N. Vuillerme, N. Noury. Walk detection with a kinematic sensor: frequency and wavelet comparison. Proceedings of the 28th IEEE EMBS annual international conferenc, (1):1711–1714, 2006.
94. P. Beaudoin, S. Coros, M. van de Panne, P. Poulin. Motion-motif graphs. Proceedings of the 2008 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pages 117–126, 2008.
95. C. M. Bishop. Pattern recognition and machine learning. Springer, 2006.

96. C. M. Bishop. A new framework for machine learning. In J.M. Zurada et al., editor, WCCI 2008 Plenary/Invited Lectures, LNCS 5050, 2008.
97. C. M. Bishop, J. Lasserre. Generative or discriminative? getting the best of both worlds. Bayesian Statistics, 8:3–23, 2007.
98. G. Bouchard, B. Triggs. The tradeoff between generative and discriminative classifiers. In J. Antoch, editor, Proc. of COMPSTAT'04, 16th Symposium of IASC, volume 16. Physica-Verlag, 2004.
99. J. B. J. Bussmann, Y. M. van de Laar, M. P. Neeleman, H. J. Stam. Ambulatory accelerometry to quantify motor behaviour in patients after failed back surgery: a validation study. Pain, 74(2-3):153–161, 1998.
100. Z.C. Chao, Y. Nagasaka, N. Fujii. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. Frontiers in Neuroengineering, 3:3, 2010.
101. A. Charpentier, J.-D. Fermanian, O. Scaillet. The estimation of copulas: theory and practice. 2006.
102. H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann. Math. Statist., 4(23):493–655, 1952.
103. Morris H. DeGroot. Optimal Statistical Decisions. McGraw-Hill Book Company, New York, 1970.
104. B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. Journal of the American Statistical Association, 352(70):892–898, 1975.
105. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani. Least angle regression. newblock Ann. Statist, pages 407–499, 2004.
106. B. S. Everitt. The Cambridge Dictionary of Statistics. Cambridge University Press, 2002.
107. T. Fawcet. Roc graphs: notes and practical considerations for researchers. ReCALL 31(HPL-2003-4), pages 1–38, 2004.
108. A. Fink. The survey handbook. Thousand Oaks, CA: Sage Publications., 1995.
109. I. K. Fodor. A survey of dimension reduction techniques. Lawrence Livermore National Laboratory Reports, 9:1–18, 2002.
110. J. Friedman, R. Hastie, T. andTibshirani. Additve logistic regression: a statistical way of boosting. The Annals of Statistics, 28(2):337–407, 2000.
111. A. L. Goldberger, L.A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation, 101:23:e215–e220, 2000.
112. C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37:424–432, 1969.
113. J. Hair, R. Anderson, R. Tatham, W. Black. Multivariate data analysis (4th ed.). Upper Saddle River, NJ: Prentice Hall., 1995.

114. E. J. Hannan. The estimation of frequency. Journal of Applied Probability, 10:510–519, 1973.

115. T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Springer, 2001.

116. B. Heidecker, E. K. Kasper, I.S. Wittstein, H. C. Champion, E. Breton, S. D. Russell, M. M. Kittleson, K. L. Baughman, J. M. Hare. Transcriptomic biomarkers for individual risk assessment in new-onset heart failure. Circulation, 118(3):238–246, 2008.

117. D. Hosmer, S. Lemeshow. Applied logistic regression. N. Y.: Wiley, 2000.

118. Y. Ichimaru, G. B. Moody. Development of the polysomnographic database on CD-ROM. Psychiatry and Clinical Neurosciences, 53:175–177, 1999.

119. A. Irpino, R. Verde, Y. Lechevallier. Dynamic clustering of histograms using wasserstein metric. COMPSTAT, pages 869–876, 2006.

120. Tony Jebara. Discriminative, Generative and Imitative Learning. PhD thesis, Media Laboratory, MIT, 2001.

121. O. C. Jenkins, M. J. Mataric. Deriving action and behavior primitives from human motion data. In In International Conference on Intelligent Robots and Systems, pages 2551–2556, 2002.

122. P. P. Kanjilal, J. Bhattacharya, G. Saha. Robust method for periodicity detection and characterization of irregular cyclical seriesin terms of embedded periodic components. Physical review E, 59(4):4013–4025, 1999.

123. Mike Kestemont, Kim Luyckx, Walter Daelemans. Intrinsic plagiarism detection using character trigram distance scores. Notebook for PAN at CLEF 2011, 2011.

124. R. B. Kline. Principles and Practice of Structural Equation Modeling. New York: Guilford., 2005.

125. Tamara G. Kolda, Brett W. Bader. Tensor decompositions and applications. SIAM REVIEW, 51(3):455–500, 2009.

126. A. N. Kolmogorov. On the approximation of distributions of sums of independent summands by infinitely divisible distributions. Contributions to statistics, pages 158–174, 1965.

127. S.M. Kosslyn, M. Behrmann, M. Jeannerod. The cognitive neuroscience of mental imagery. Neuropsychologia, 33:1335–1344, 1995.

128. F. R. Kschischang, B. J. Frey, H.-A. Loeliger. Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory, 47:498–519, 2001.

129. S. Kullback. Information Theory and Statistics. New York: Wiley, 1959.

130. J. A. Lasserre, C. M. Bishop, T. P. Minka. Principled hybrids of generative and discriminative models. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'06), pages 87–94, 2006.

131. E. C. Leuthardt, K. J. Miller, G. Schalk, R. P. N. Rao, J. G. Ojemann. Electrocorticography-based brain computer interface — the Seattle experience. IEEE Transactions on Neural Systems and Rehabilitation, 14(2):194–198, 2006.

132. Q. Li, J. Racine. Nonparametric estimation of distributions with categorical and continuous data. Journal of Multivariate Analysis, 86(2):266–292, 2003.

133. P. Liang, M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudo-likelihood estimators. In In Proceedings of the 25th International Conference on Machine Learning (ICML), 2008.

134. J. Lin, E. Keogh, S. Lonardi, B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. Proceedings of the 8th SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pages 2–11, 2003.

135. D. V. Lindley. On a measure of the information provided by an experiment. 27(4):986–1005, 1956.

136. D. J. C. MacKay. Information theory, inference, and learning algorithms. Cambridge University Press, 2003.

137. C. Mallows. A note on asymptotic joint normality. Annals of Mathematical Statistics, 42(2):508–515, 1972.

138. Andrew McCallum, Chris Pal, Greg Druck, Xuerui Wang. Multi-conditional learning: Generative-discriminative training for clustering and classification. AAAI'06 Proceedings of the 21st national conference on Artificial intelligence, 1:433–439, 2006.

139. P. McCullagh, J. A. Nelder. Generalized Linear Models. 2nd edition. Chapman and Hall, London., 1989.

140. C. E. McCulloch. Joint modeling of mixed outcome types using latent variables. Statistical Methods in Medical Research, 17:53–73, 2008.

141. D. J. McFarland, L. A. Miner, T.M.Vaughan, J. R.Wolpaw. Mu and beta rhythm topographies during motor imagery and actual movements. Brain Topography, 12(3):177–186, 2000.

142. J. del R. Millán, F. Renken, J. Mouri no, W. Gerstner. Brain-actuated interaction. Artificial Intelligence, 159:241–259, 2004.

143. Markus Muhr, Roman Kern, Mario Zechner, Michael Granitzer. External and intrinsic plagiarism detection using a cross-lingual retrieval and segmentation system. Lab Report for PAN at CLEF 2010, 2010.

144. B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Bula, P. Robert. Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly. IEEE Transactions on Biomedical Engineering, 50(6):711–723, 2003.

145. Hariharan Narayanan, Sanjoy Mitter. Sample complexity of testing the manifold hypothesis. NIPS 2010, pages 1786–1794, 2010.

146. R. B. Nelsen. An introduction to copulas. Springer, 2006.

147. A. Y. Ng, M. I. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in Neural Information Processing Systems, 14:841–848, 2002.

148. Luis Fernando Nicolas-Alonso, Jaime Gomez-Gil. Brain computer interfaces, a review. Sensors (Basel), 12(2):1211–1279, 2012.

149. Giovanni Parmigiani, Donald A. Berry. Applications of Lindley Information Measure to the Design of Clinical Experiments, pages 329–348. Wiley, New York, 1994.

150. J. Pearl. Graphs, causality and structural equation models. Sociological Methods and Research, 27(2):226–284, 1998.

151. T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, J. H. Peter. The Apnea-ECG database. Computers in Cardiology, 255–258, 2000.

152. Y. Ichimaru, G. B. Moody. Development of the polysomnographic database on CD-ROM. Psychiatry and Clinical Neurosciences, 53:175-177, 1999.

153. F. Perez-Cruz. Kullback-leibler divergence estimation of continuous distributions. IEEE International Symposium on Information Theory, 2008.

154. M. Potthast, B. Stein, A. Barrón-Cedeño, P. Rosso. An evaluation framework for plagiarism detection. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 997–1005, 2010.

155. Martin Potthast, Andreas Eiselt, Alberto Barr'on-Cedeno, Benno Stein, Paolo Rosso. Overview of the 3rd international competition on plagiarism detection. Working Notes Papers of the CLEF 2011 Evaluation Labs, 2011.

156. L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–285, 1989.

157. R. Raina, Y. Shen, A. Y. Ng, A. McCallum. Classication with hybrid generative-discriminative models. Advances in Neural Information Processing Systems, 16:545–552, 2003.

158. J. Rennie, L. Shih, J. Teevan, D. Karger. Tackling the poor assumptions of naive bayes classifiers. Proceedings of the Twentieth International Conference on Machine Learning (ICML), 2003.

159. Irene Rodriguez-Lujan, Ramon Huerta, Charles Elkan, Carlos Santa Cruz. Quadratic programming feature selection. Journal of Machine Learning Research, 11:1491–1516, 2010.

160. B. Rosner. Fundamentals of biostatistics. Duxbury Press, 1999.

161. Y. Dan Rubinstein, Trevor Hastie. Discriminative vs informative learning. In Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining, pages 49–53, 1997.

162. E. Said, D. A. Dickey. Testing for unit roots in autoregressive moving average models of unknown order. Biometrika, 71(1):599–607, 1984.

163. Efstathios Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. PAN'09, pages 38–46, 2009.

164. G. Sugihara, R.M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. Nature, 344-6268:734–741, 1990.

165. Marcin Olof Szummer. Learning from Partially Labeled Data. PhD thesis, Department of Electrical Engineering and Computer Science, MIT, 2002.

166. F. Takens. Detecting strange attractors in turbulence. Dynamical Systems and Turbulence, Lecture Notes in Mathematics, 898:366–381, 1981.

167. A. M. Walker. On the estimation of a harmonic component in a time series with stationary independent residuals. Biometrika, 58:21–36, 1971.

168. J. R. Wolpaw, D. J. McFarland. Multichannel eeg-based brain-computer communication. Electroencephalography and Clinical Neurophysiology, 90(6):444–449, 1994.

169. J. R. Wolpaw, D. J. McFarland, G. W. Neat, C. A. Forneris. An eeg-based brain-computer interface for cursor control. Electroencephalography and clinical neurophysiology, 78(3):252–259, 1991.

170. J.-H Xue, D.M. Titterington. Comment on "discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes". Neural Processing Letters, 28(3):169–187, 2008.

171. F. Zhou, F. De la Torre. Aligned cluster analysis for temporal segmentation of human motion. In IEEE Conference on Automatic Face and Gestures Recognition, 2008.