САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

На правах рукописи

АНАНЬЕВСКАЯ Полина Валерьевна

Исследование конечно-линейных статистических моделей. Оптимизация и избыточность

05.13.18 – Математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата физико-математических наук

Санкт-Петербург – 2013

Работа выполнена на кафедре статистического моделирования математико-механического факультета Санкт-Петербургского государственного университета.

Научный руководитель:	кандидат физико-математических наук, доцент Алексеева Нина Петровна,
Официальные оппоненты:	доктор физико-математических наук, профессор Егоров Владимир Алексеевич (Санкт-Петербургский государственный электротехнический университет “ЛЭТИ”)
	доктор технических наук, доцент Буре Владимир Мансурович, (Санкт-Петербургский государственный университет, факультет ПМ-ПУ)
Ведущая организация:	Санкт-Петербургский государственный политехнический университет

Защита состоится 19 июня 2013 г. в 16.00 часов на заседании диссертационного совета *Д.212.232.50* по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук при Санкт-Петербургском государственном университете по адресу: 199034, Санкт-Петербург, В. О., Университетская наб., 7/9, Менделеевский Центр.

С диссертацией можно ознакомиться в Научной библиотеке им. М. Горького Санкт-Петербургского государственного университета, расположенной по адресу: 199034, Санкт-Петербург, Университетская наб., 7/9.

Автореферат разослан «____» _____ 2013 г.

Ученый секретарь
диссертационного совета,
доктор физ.-мат. наук, профессор *КУРБАТОВА Г. И.*

Общая характеристика работы

Актуальность темы. Описание некоторых систем в различных областях современной науки, таких как биология, медицина, физика и химия, зачастую содержит информацию, представленную в виде большого количества категориальных признаков. В качестве примеров таких признаков можно привести наличие или отсутствие симптомов различных заболеваний пациентов в медицине, кодирование нуклеотидных остатков при помощи двух или четырех битов в задачах генетики, описание свойств и взаимодействий молекул в химии, идентификацию различных семантических структур в лингвистике.

Для анализа данных такого рода используются разнообразные методы, суть которых можно свести к трем основным направлениям: агрегирование информации в структуры меньшей размерности, выявление наиболее связных композиций, прогнозирование итоговой характеристики. Предназначенные для решения таких задач линейные многомерные статистические методы могут применяться в этом случае только после соответствующего преобразования категориальных данных в числовые или порядковые, представленного в работах (Greenacre, 1984) и (Giffi, 1990).

Однако когда для шкалирования наблюдений используются более сложные логические формулы, более уместным оказывается привлечение алгебраических и комбинаторных методов, позволяющих преобразовывать категориальные данные на основе линейных комбинаций случайных величин над конечным полем в новые наборы признаков той же природы, но с более предпочтительными информационными свойствами. Подобные преобразования в работе (Cox, 1972) называются перестановочными трансформациями, но вводятся не линейно, а через систему логических отношений. В работе (Bloomfield, 1974) намечается переход от перестановочных трансформаций к сложению двух бинарных признаков по модулю два, которое использует-

ся для расширения вариантов лог-линейных моделей. В работах (Алексеева 2004, 2007, 2008) посвященных исследованию комбинаторной структуры бинарных признаков на основе конечных геометрий, линейные оболочки над конечным полем используются для обеспечения так называемого симптомно-синдромального подхода к решению задач клинической диагностики. Широкий спектр приложений привел к необходимости развития и формализации такого подхода, заключающегося в использовании линейных преобразований исходных признаков над конечным полем и называемого далее конечно-линейным подходом.

Цель работы. Основной целью работы является формализация статистических моделей, основанных на конечно-линейных методах, с учетом расширения на случай произвольного числа градаций, а также построение эффективного алгоритма оценки параметров таких моделей, по возможности адаптированного к параллельным вычислениям.

Основные положения и результаты, выносимые на защиту. Формализованы и расширены на случай произвольной характеристики поля конечно–линейные статистические модели редукции размерности набора признаков, взаимодействия двух или более наборов и классификации. При помощи моделирования исследована сходимость по вероятности оценок параметров к теоретическим их значениям в данных моделях. Разработан и адаптирован для параллельных вычислений на GPU специальный метод дискретной оптимизации для оценки параметров конечно–линейных моделей, опирающийся на алгебраические свойства многообразий Грассмана. Метод реализован в системе программ, в том числе и на основе параллельных вычислений с использованием технологии CUDA. Для задачи классификации найдена оценка функции распределения количества ошибок классификации в случае независимых и равномерно распределенных исходных признаков.

Методы исследования. В работе применяются методы статистическо-

го моделирования, теории вероятностей, комбинаторики и линейной алгебры. Программирование осуществлялось в пакетах R и SAGE, а так же на языке программирования C++ с использованием технологии CUDA.

Научная новизна. Все основные результаты диссертации являются новыми.

Теоретическая и практическая ценность. Конечно–линейные модели позволяют выявлять скрытую информацию в анализе категориальных данных и могут быть использованы как аналоги анализа главных компонент, канонического корреляционного анализа и регрессии в традиционной многомерной статистике. Построенный на основе алгебраических методов специальный алгоритм дискретной оптимизации позволяет в реальном времени решать задачи оценки параметров в данных моделях, а так же проверять избыточность классификационной модели. Адаптированность алгоритма к параллельным вычислениям позволяет ускорять вычисления за счет привлечения графических процессоров, как одного из активно развивающихся сегодня направлений в области параллельных вычислений. Предложенная методология и комплекс программ успешно использованы на практике для анализа экспериментальных данных.

Апробация работы. Основные результаты диссертации докладывались и обсуждались на семинаре кафедры статистического моделирования математико–механического факультета СПбГУ, а так же были представлены на конференциях: the 2nd International Conference on BioMedical Engineering and Informatics (BMEI'09), Tianjin, China, 17–19 October 2009; всероссийская научно-практическая конференция с международным участием «Алмазовские чтения 2011», ФЦСКЭ им. В. А. Алмазова, г. Санкт-Петербург, 19–21 мая 2011 г; V Международная научная конференция (заочная) «Системный анализ в медицине» (САМ 2011), г. Благовещенск, 25–26 мая 2011 г.

Публикации. По теме диссертации опубликованы статьи [1, 2] в журна-

лах по перечню ВАК и статьи [3–5] в сборниках трудов конференций. Статьи [2–5] написаны в соавторстве, в статье [4] автору принадлежит доказательство теоремы о базисных симптомах и импульсный алгоритм поиска оптимального синдрома второго порядка, в статьях [2, 3] — алгоритм дискретной оптимизации и его реализации для исследования связности цепочек РНК и выделения прогностических факторов у больных с глиомами, в статье [5] — раздел о информационном структурировании категориальных данных.

Структура и объем диссертации. Диссертация состоит из введения, 7 глав, заключения и библиографии. Общий объем диссертации 142 страницы, из них 128 страниц текста, включая 19 рисунков. Библиография включает 98 наименований на 11 страницах.

Содержание работы

Рассматривается конечное поле $\mathbb{F}_q$ и дискретная сигма-алгебра $2^{\mathbb{F}_q}$, состоящая из всевозможных подмножеств $\mathbb{F}_q$. Через $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_m)^T$ обозначен случайный вектор, состоящий из m случайных компонент, представляющих собой дискретные случайные величины со значениями в $(\mathbb{F}_q, 2^{\mathbb{F}_q})$, заданные на некотором вероятностном пространстве $(\Omega, \mathcal{F}, P)$. В приложениях наблюдаемые реализации вектора $\mathcal{X}$, состоящие из реализаций его одномерных компонент $\mathcal{X}_i$, называются набором категориальных признаков, каждый из которых принимает значения в поле $\mathbb{F}_q$.

Линейная комбинация $\mathcal{X}_\tau = a_1\mathcal{X}_1 + \ldots + a_m\mathcal{X}_m$, где a_i — элементы поля $\mathbb{F}_q$, а операции сложения и умножения производятся в поле $\mathbb{F}_q$, также является случайной величиной со значениями в $(\mathbb{F}_q, 2^{\mathbb{F}_q})$.

Любое линейное преобразование случайного вектора $\mathcal{X}$ в новый случайный вектор $\tilde{\mathcal{X}} = (\mathcal{X}_{\tau_1}, \mathcal{X}_{\tau_2}, \ldots, \mathcal{X}_{\tau_k})^T$, состоящий из k новых случайных компонент со значениями также в $(\mathbb{F}_q, 2^{\mathbb{F}_q})$, может быть задано состоящей из

элементов поля $\mathbb{F}_q$ матрицей $A = \{a_{ij}\}$, где $1 \le i \le k$ и $1 \le j \le m$. Каждая компонента вектора $\tilde{\mathcal{X}} = A\mathcal{X}$ задается равенством $\mathcal{X}_{\tau_i} = a_{i1}\mathcal{X}_1 + \ldots + a_{im}\mathcal{X}_m$.

Одной из главных задач диссертации является построение конечно–линейных статистических моделей как аналогов методов анализа главных компонент, канонического корреляционного анализа и регрессии. Параметром в каждой из трех моделей является матрица A над конечным полем $\mathbb{F}_q$, задающая линейное преобразование исходного вектора, удовлетворяющее некоторому критерию оптимальности.

В первой главе производится формализация конечно–линейных статистических моделей, в рамках которых ставятся специальные задачи дискретной оптимизации. Первый параграф посвящен обзору существующих вариантов построения аналогов трех многомерных статистических методов (анализ главных компонент, канонический корреляционный анализ и регрессия) как задач оптимизации различных функций, решение которых основано на известных алгоритмах сингулярного разложения, метода наименьших квадратов, векторного метода подгонки и других.

Во втором параграфе нами формулируются задачи дискретной оптимизации в рамках трех конечно–линейных моделей, предназначенных для анализа данных категориального типа и получивших названия редукции (снижения) размерности, канонической сцепленности (структурной зависимости), жестко–симметрической классификации. В основе данных моделей лежит другой подход к решению традиционных задач сжатия, связности и прогнозирования информации, рассматриваемых в первом параграфе. Для решения задач оптимизации в конечно–линейных моделях требуются собственные алгоритмы дискретной оптимизации.

Снижение размерности вектора $\mathcal{X}$ сводится к задаче поиска матрицы параметров $A = A(k, m)$, при заданных $k < m$, чтобы различие между распределениями исходного вектора $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_m)^T$ и вектора $\tilde{\mathcal{X}} = A\mathcal{X}$ было

минимальным. В качестве меры различия предлагается применять разность энтропий Шеннона (Shennon, 1948) $H(\mathcal{X})$ и $H(A\mathcal{X})$, при этом задача снижения размерности заключается в поиске хотя бы одной матрицы $A = A(k, m)$, $k < m$, при которой достигается минимум следующей функции:

$$\sigma(A) = H(\mathcal{X}) - H(A\mathcal{X}). \tag{1}$$

Выбор такой функции обусловлен широко используемым в работах (Akaike, 1973), (Burg, 1975), (Berger, 1996) принципом максимума энтропии. Принцип заключается в выборе такого распределения для описания эксперимента, которое удовлетворяет заданным ограничениям и имеет максимальную энтропию среди всех распределений, удовлетворяющих таким же ограничениям.

Модель канонической сцепленности предназначена для выявления k зависимостей между векторами $\mathcal{X}$ и $\mathcal{Y}$ с m_1 и m_2 случайными компонентами соответственно. В качестве меры зависимости $\mathcal{X}$ и $\mathcal{Y}$ предлагается использовать односторонний и двусторонний коэффициенты неопределенности Тейла (Theil, 1970)

$$R_0(\mathcal{X}, \mathcal{Y}) = \frac{I(\mathcal{X}, \mathcal{Y})}{H(\mathcal{Y})} \quad \text{и} \quad R(\mathcal{X}, \mathcal{Y}) = 2 \cdot \frac{I(\mathcal{X}, \mathcal{Y})}{H(\mathcal{X}) + H(\mathcal{Y})}$$

как нормальзованные версии совместной информации $I(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) + H(\mathcal{Y}) - H(\mathcal{X}, \mathcal{Y})$. Здесь через $H(\mathcal{X}, \mathcal{Y})$ обозначена энтропия распределения вектора $(\mathcal{X}_1, \ldots, \mathcal{X}_{m_1}, \mathcal{Y}_1, \ldots, \mathcal{Y}_{m_2})$. Оптимизационная задача состоит в поиске хотя бы одной пары матриц преобразований $A^{(1)}(k, m_1) = (A_1^{(1)}, \ldots, A_k^{(1)})^T$ для $\mathcal{X}$ и $A^{(2)}(k, m_2) = (A_1^{(2)}, \ldots, A_k^{(2)})^T$ для $\mathcal{Y}$, максимизирующих функции

$$\sigma(A_i^{(1)}, A_i^{(2)}) = R((A_i^{(1)})^T\mathcal{X}, (A_i^{(2)})^T\mathcal{Y}), \quad i = 1, \ldots, k. \tag{2}$$

при ограничениях на значения энтропии $H((A_1^{(1)}, \ldots, A_i^{(1)})^T\mathcal{X}) \geq h_{1i} \cdot H(\mathcal{X})$ и $H((A_1^{(2)}, \ldots, A_i^{(2)})^T\mathcal{Y}) \geq h_{2i} \cdot H(\mathcal{X})$, где h_{1i} и h_{2i} – управляющие параметры. Кроме того предлагается вариант обобщения этого подхода на случай l случайных векторов $\mathcal{X}^{(s)} = (\mathcal{X}_1^{(s)}, \ldots, \mathcal{X}_{m_s}^{(s)})$, $s = 1, \ldots, l$.

Задача жестко–симметрической классификации заключается в предсказании случайной величины $\mathcal{Y}$ по случайному вектору $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_m)^T$ на основе его линейного преобразования $A\mathcal{X}$ при помощи матрицы параметров A размера $1 \times m$ так, чтобы отличие между $\mathcal{Y}$ и $A\mathcal{X}$ было минимально. В качестве меры отличия в данном случае используется вероятность несовпадения двух случайных величин $A\mathcal{X}$ и $\mathcal{Y}$ с учетом перенумерации

$$\rho_1(A\mathcal{X}, \mathcal{Y}) = \min_f (1 - P(A\mathcal{X} = f(\mathcal{Y}))), \tag{3}$$

где минимизация производится по всем возможным биекциям $f\colon \mathbb{F}_q \to \mathbb{F}_q$, осуществляющим перенумерации множества значений $\mathcal{Y}$. Оптимизационная задача заключается в поиске хотя бы одной точки минимума функции на множестве матриц $A = A(1, m)$

$$\sigma(A) = \rho_1(A\mathcal{X}, \mathcal{Y}). \tag{4}$$

Во всех трех оптимизационных задачах искомая матрица A, на которой достигается глобальный экстремум, может быть не единствена. Это означает существование различных способов снижения размерности или несколько видов взаимосвязей или способов классификации. В таких задачах интерес представляет нахождение хотя бы одного решения $\hat{A}$. Его поиск производится на основе эмпирического распределения при помощи построенного в последующих главах специального алгоритма дискретной оптимизации.

Во второй главе предлагается вариант векторной параметризации многообразия Грассмана (грассманиана) над конечным полем, на основе которой в дальнейшем строится необходимый алгоритм дискретной оптимизации.

Рассмотрим пространство $V_m = (\mathbb{F}_q)^m$ с базисом, состоящим из m линейно независимых над полем $\mathbb{F}_q$ векторов $X_1, \ldots, X_m$. Любой набор линейно независимых векторов $(X_{\tau_1}, \ldots, X_{\tau_k})$ является базисом k-мерного подпространства линейных комбинаций $W_k = \langle X_{\tau_1}, \ldots, X_{\tau_k} \rangle$ пространства V_m. Каж-

дый вектор X_{τ_i} раскладывается по базису $X_1, \ldots, X_m$

$$X_{\tau_i} = a_{i1}X_1 + \ldots + a_{im}X_m, \tag{5}$$

где a_{ij} — элементы поля $\mathbb{F}_q$.

Грассманианом называется множество всех k-мерных подпространств W_k пространства V_m. Нами решается задача о быстром перечислении всех W_k (точек грассманиана), для чего используется векторная параметризация грассманиана, которая представляет собой универсальный способ выделения базисов $(X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k})$ всех W_k, таким образом, что никакие два базиса не задают одно и то же подпространство. Эта параметризация описывается в следующей доказанной диссертантом теореме. Зафиксируем полный флаг $\mathcal{F}$ на пространстве $V_m = (\mathbb{F}_q)^m$,

$$V_0 = \{0\} \subset V_1 = \langle X_1 \rangle \subset V_2 = \langle X_1, X_2 \rangle \subset \ldots \subset V_m = \langle X_1, \ldots, X_m \rangle .$$

Отношение линейного порядка $\prec$ на V_m согласовано с флагом $\mathcal{F}$, если для всех $i = 0, 1, \ldots, m$ и $v \in V_i$, $w \in V_m \setminus V_i$ имеет место $v \prec w$.

Теорема 1. *Отображение* $(X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k}) \longmapsto \langle X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k} \rangle$ *устанавливает биекцию между k-мерными подпространствами W_k при фиксированном $\mathcal{F}$ и наборами векторов $X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k} \in V_m$ такими, что*

1. *для $X_{\tau_i} = a_{i1}X_1 + \ldots + a_{im}X_m$ при $s_i < m$ имеет место*

   $(a_{i1}, \ldots, a_{im}) = (a_{i1}, \ldots, a_{is_i}, 1, 0, \ldots, 0)$

2. $X_{\tau_i} \prec X_{\tau_j}$ *при* $i < j$,

3. *для всех X_{τ_i} и X_{τ_j}, $i < j$, выполнено* $a_{j(s_i+1)} = 0$.

Теорема 1 доказывается для случая $q = 2$ во втором параграфе, причем доказательство основано на свойствах флагов. Общий случай рассматривается при помощи классического клеточного разложение в третьем параграфе.

В третьей главе показывается, как задачи поиска матриц преобразований A в конечно–линейных моделях сводятся к задачам оптимизации функций (1),(2),(4) на множестве подпространств $W_k \subset V_m$ над $\mathbb{F}_q$. Для решения задач такого рода существуют методы полного перебора, релаксации (Calvet, 2003), рекурсивный метод ветвей и границ (Land, Doig, 1960) и другие (Raghavan, Thompson, 1987). Зачастую наиболее эффективными оказываются методы, основанные на специфических особенностях рассматриваемых функций и объектов. В данном случае предлагается алгоритм оптимизации основанный на векторной параметризация грассманиана.

В первом параграфе на предмет согласованности с флагом исследованы три отношения линейного порядка: лексикографический, степенно–лексикографический (импульсный) и обобщенный порядок Грея (Guan, 1998). Упорядочивание $X_{\tau_i} \in V_m$ согласно обобщенному порядку Грея осуществляется путем упорядочивания наборов коэффициентов $(a_{i1}, \ldots, a_{im})$ в разложении (5) так, что каждый следующий набор $(a_{i1}, \ldots, a_{im})$ отличается от предыдущего $(a'_{i1}, \ldots, a'_{im})$ прибавлением 1 по модулю q ровно к одному a'_{ij}.

Теорема 2. *Лексикографический порядок и обобщенный порядок Грея согласованы с флагом $\mathcal{F}$ на пространстве $V_m = (\mathbb{F}_q)^m$, а степенно–лексикографический порядок согласован только при $m = 1$ и при $m = 2$, $q = 2$.*

Теорема 3. *Упорядочивание множества векторов пространства V_m над полем $\mathbb{F}_q$, соответствующее обобщенному порядку Грея с ограничением на старший разряд, согласованным с пунктом 1 теоремы 1, минимизирует количество используемой памяти и гарантирует одинаковое количество операций для формирования каждого следующего вектора.*

Во втором параграфе на основе теоремы 1 построен алгоритм быстрого перечисления точек грассманиана FGEA (Fast Grassmannian Enumeration

Algorithm). В соответствии с этим алгоритмом, в четвертом параграфе рассмотрены два подхода к задаче дискретной оптимизации некоторой функции μ, заданной на множестве подпространств пространства V_m: поиск оптимального W_k путем *последовательного* перебора всех W_k, *пошаговый* перебор k базисных векторов X_{τ_i}, где на i-том шаге максимизируется (минимизируется) функция $\mu(W_i)$, $i \leq k$. Особым требованием к μ при пошаговом переборе является свойство ее монотонного возрастания (или убывания) на множестве всех W_k, состоящее в том, что для любых $W_s \subset W_t \subset V_m$ имеет место $\mu(W_s) \leq \mu(W_t)$ (соответственно $\mu(W_t) \leq \mu(W_s)$).

Теорема 4. *В случае пошагового перебора алгоритм FGEA (S-FGEA) эффективнее метода полного перебора наборов* $(X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k})$ *более чем в* q^k *раз, а в случае последовательного перебора (F-FGEA) более, чем в* q^{k^2} *раз.*

В пятом параграфе описывается применение F-FGEA к задаче снижения размерности и задаче классификации, а так же применение S-FGEA к задаче поиска структурной зависимости. В обоих случаях применение основывается на взаимно однозначном соответствии между наборами векторов $(X_{\tau_1}, X_{\tau_2}, \ldots, X_{\tau_k})$ и матрицами коэффициентов $\hat{A}$, а также на инвариантности энтропии относительно обратимых линейных преобразований и монотонном возрастании при конкатенации случайных векторов.

В четвертой главе алгоритм FGEA в случае поля $\mathbb{F}_2$ адаптируется к параллельным вычислениям на графических процессорах Nvidia поколения Fermi с использованием технологии CUDA. В первом параграфе приводится описание особенностей архитектуры таких графических процессоров, отражающихся в методах работы с потоковой структурой и различными видами памяти. Во втором параграфе показано, каким образом реализуется алгоритм FGEA на CUDA с использованием минимального количества памяти, которое достигается за счет свойства порядка Грея из теоремы 3. Третий параграф

посвящен способу быстрого вычисления энтропии без составления гистограммы как еще одному преимуществу, имеющему место при перенесении алгоритма на графические процессоры. Оценка эффективности алгоритма FGEA на модельных выборках, представленная в четвертом параграфе, показала, что такой подход позволяет сократить время вычислений в сотни раз.

В пятой главе исследуются свойства оценок $\hat{A}$ матриц параметров A из (4) и (1) на модельных выборках. В первом параграфе изучаются свойства оценок в задаче снижения размерности.

В случае независимых $\mathcal{X}_i$ для $\hat{A}$ проверяется, что для $\forall\, \varepsilon > 0$ имеет место $P(|p_n - 1| > \varepsilon) \underset{n\to\infty}{\longrightarrow} 0$, где p_n обозначает вероятность совпадения матриц A и $\hat{A}$ для заданного объема выборки n. Для этого моделируется 100 значений оценки $\hat{p}_n$, вычисляемых в свою очередь по 100 испытаниям и демонстрируется сходимость $E\hat{p}_n \underset{n\to\infty}{\longrightarrow} 1$ и $E\left(\hat{p}_n - 1\right)^2 \underset{n\to\infty}{\longrightarrow} 0$.

Для независимых $\mathcal{X}_i$ были получены следующие результаты. Скорость сходимости инвариантна относительного линейного обратимого преобразования вектора $\mathcal{X}$ над полем $\mathbb{F}_q$, но зависит от распределения информационной нагрузки между компонентами $\mathcal{X}_i$. При этом самая высокая скорость сходимости наблюдается в случае матрицы A размера $k \times m$, где k совпадает с количеством s компонент $\mathcal{X}_i$, распределение которых вносит основной вклад в общую энтропию. С увеличением m при фиксированном $s < m$ наблюдается падение скорости сходимости. Представление о характере распределения информационной нагрузки можно получить по профилю энтропии, представляющего собой график значений энтропии всевозможных подпространств.

В случае зависимых $\mathcal{X}_i$ теоретическое вычисление A, а следовательно, и моделирование p_n, требует явного задания вида зависимости. Перебор всех видов зависимости не представляется возможным. В данном случае при помощи моделирования можно проверить лишь факт существования такой матрицы $\hat{A}$, что $P(\hat{A} = \underset{A}{\operatorname{argmax}}\, H(AX)) \underset{n\to\infty}{\longrightarrow} 1$ и $P(\hat{A}_j = \underset{A}{\operatorname{argmax}}\, H(AX)) \underset{n\to\infty}{\longrightarrow} 0$

для остальных возможных вариантов $\hat{A}_j$. Для зависимых компонент таким способом был установлен факт наличия сходимости. Отсутствие сходимости для зависимых компонент равно как и для независимых может наблюдаться при полимодальности.

Второй параграф посвящен вопросу сходимости по вероятности в задаче классификации. Помимо естественного уменьшения скорости сходимости при увеличении количества компонент m выявлена проблема неоправданно малого количества ошибок классификации при малом объеме выборки n и $\hat{A} \neq A$. Этот факт обусловлен возрастанием числа вариантов AX при увеличении m, которое в крайнем случае может покрыть все множество возможных Y.

В общем случае решение проблемы соотношения m и n является сложной задачей, однако для случая одинаково распределенных и независимых компонент $\mathcal{X}_i$ в рамках ее решения **в шестой главе** при заданных Y, n, m, q вычисляется оценка функции распределения $F_\eta(t) = P(\eta < t)$ случайной величины η, характеризующей количество ошибок классификации, где $\eta = \rho_2(X, Y) = \min\limits_{X_\tau \in \mathcal{L}(X)} \rho_1(X_\tau, Y)$. Здесь $\mathcal{L}(X)$ обозначает линейную оболочку множества столбцов матрицы $X = X(n, m)$, а Y – итоговый вектор длины n, координаты которого принимают $l \leq q$ различных значений над $\mathbb{F}_q$ и задают разбиение n элементов на l классов. Расстояние ρ_1 между векторами X_τ и Y определяется по аналогии с (3) $\rho_1(X_\tau, Y) = \min\limits_{Y^* \in \mathbb{Y}} \rho_0(X_\tau, Y^*)$, где $\rho_0(X_\tau, Y^*)$ вычисляется как минимальное количество замен, необходимое для получения вектора X_τ из вектора Y^*. В обозначении $\mathbb{Y}$ — множество всех векторов $Y^* \in \mathbb{Y}$, задающих то же разбиение ($Y \in \mathbb{Y}$). Вектора $Y, Y' \in \mathbb{F}_q^m$ задают одинаковое разбиение n элементов на l классов, если оба вектора состоят ровно из l различных элементов поля $\mathbb{F}_q$ и выполнено $\rho_1(Y, Y') = 0$. Для оценивания $F_\eta(t)$ диссертантом доказаны следующие теоремы.

Теорема 5. *Количество матриц $X = X(n, m)$ ($\#\mathbb{X}$) над $\mathbb{F}_q$ таких, что хотя бы один вектор, задающий такое же разбиение n элементов на l классов,*

что и Y*, инцидентен* $\mathcal{L}(X) = \langle X_1, \ldots, X_m \rangle$*, вычисляется по формуле:*

$$\#\mathbb{X} = \sum_{r=0}^{m} \left(\prod_{i=0}^{r-1} (q^m - q^i) \right) \sum_{k=\max(0,l+r-n,r-m+1)}^{\min(l,r)} A_q(l,k) \cdot q^{(l-k)(r-k)} \cdot \binom{n-l}{r-k}_q, \tag{6}$$

где $\binom{l}{0}_q = 1$*,* $A_q(1,0) = 1$*,* $A_q(l,0) = 0$ *при* $l > 1$*,* $A_q(l,l) = 1$*,*

$$A_q(l,k) = A_q(l-1,k-1) + A_q(l-1,k) \cdot (q^k - l + 1),$$

$$\binom{l}{k}_q = \prod_{i=0}^{k-1} \frac{1-q^{l-i}}{1-q^{i+1}}.$$

Обозначим элементы поля $\mathbb{F}_q$ числами $\{1, 2, \ldots, q\}$. Пусть координаты вектора $Y^{'}$ принимают ровно $r \leq q$ значений и $b(K) = \sum\limits_{i \notin K} \#\{k \,|\, y_k = i\}$.

Теорема 6. *Количество* R_r^b *векторов* $Y^{'}$ *таких, что они отличаются от вектора* $Y = (y_1, y_2, \ldots, y_n)$ *ровно в* b *координатах, вычисляется по формуле:*

$$R_r^b = \sum_{\substack{K \subset 1,\ldots,q \\ |K|=r}} \sum_{i=0}^{|K|} (-1)^i \sum_{\substack{K^{'} \subset K \\ \left|K^{'}\right| = |K|-i}} C_{n-b(K^{'})}^{b-b(K^{'})} \cdot r^{b(K^{'})} \cdot (r-1)^{(b-b(K^{'}))}. \tag{7}$$

Оценка искомой функции распределения приобретает следующий вид:

$$F_\eta(t) = P(\eta < t) \leq \sum_{r=1}^{q} \frac{\#\mathbb{X}}{q^{mn}} \cdot \sum_{j=0}^{t-1} R_r^j, \tag{8}$$

где $t \in \mathbb{Z}_+$, а $\#\mathbb{X}$ вычисляется по формуле (6), а R_r^j по формуле (7).

Теоретическая оценка из формулы (8) хорошо согласуется при $P(\eta < t) \leq 0.1$ с эмпирической оценкой функции распределения $F_\eta(t)$, полученной при помощи моделирования. Использование этого результата возможно в практических приложениях для проверки адекватности применения конечно–линейного метода классификации, хотя строгая формализация критерия требует отдельного исследования. Рассматриваемые конечно–линейные

модели были успешно применены при обработке реальных медицинских данных, примеры описаны **в седьмой главе**. В **заключении** подводятся итоги диссертационного исследования и формулируются основные результаты.

Список публикаций автора

1. **Ананьевская (Грачева) П. В. Метод дискретной оптимизации на основе параметризации грассманиана в многомерном структурировании дихотомических данных** // Вестн. С.-Петерб. ун-та. Сер. 1: Математика, Механика, Астрономия. 2011. Вып. 4. С. 28–37.

2. **Мартынов Б. В., Алексеева Н. П., Ананьевская (Грачева) П. В. и др. Прогностические факторы у больных с глиомами: симптомно-синдромальный анализ** // Вестн. Рос. Воен.-мед. акад. 2010. Т. 1, № 29. С. 7–14.

3. Alexeyeva N., Alexeyev A., Ananyevskaya (Gracheva) P. et al. Symptom and syndrom analysis of categorial series, logical principles and forms of logic // Proc. of the 3nd International Conference on BioMedical Engineering and Informatics (BMEI). 2010. P. 2603–2606.

4. Alexeyeva N., Smirnov I., Ananyevskaya (Gracheva) P., Martynov B. The finitely geometric symptom analysis in the glioma survival study // Proc. of the 2nd International Conference on BioMedical Engineering and Informatics (BMEI). 2009. P. 1–4.

5. Алексеева Н. П., Ананьевская (Грачева) П. В., Подхалюзина Е. М. Структурирование и систематизация факторов на основе конечных проективных подпространств. // Материалы V международной конференции «Системный анализ в медицине» (САМ), Благовещенск. 2011. С. 14–16.