

ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ  
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ  
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук  
Образовательная программа «Прикладная математика и информатика»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

**Исследовательский проект на тему:**

**Задача линейной регрессии со структурированными векторами  
признаков**

**Выполнил студент:**

группы #БПМИ207, 4 курса

Пархоменко Юлия Владимировна

**Принял руководитель ВКР:**

Пучкин Никита Андреевич

Старший научный сотрудник

Факультет компьютерных наук НИУ ВШЭ

Москва 2024

# Содержание

<b>1</b>	<b>Введение</b>	<b>5</b>
<b>2</b>	<b>Обзор литературы</b>	<b>6</b>
<b>3</b>	<b>Теоретическое описание методов</b>	<b>8</b>
3.1	Оценка матриц ковариаций . . . . .	9
3.1.1	Постановка задачи . . . . .	9
3.1.2	Оценка произведения Кронекера . . . . .	10
3.1.3	Оценка суммы произведений Кронекера . . . . .	11
3.2	Применение Барицентров Вассерштейна . . . . .	12
3.2.1	Постановка задачи . . . . .	12
3.2.2	Добавление регуляризации . . . . .	13
<b>4</b>	<b>Эксперименты</b>	<b>14</b>
4.1	Оценка структурных матриц ковариаций . . . . .	14
4.1.1	Данные . . . . .	14
4.1.2	Оценка матрицы ковариаций . . . . .	14
4.1.3	Оценка параметров модели . . . . .	18
4.1.4	Выводы . . . . .	21
4.2	Применение Барицентров Вассерштейна . . . . .	22
<b>5</b>	<b>Заключение</b>	<b>23</b>
	<b>Список литературы</b>	<b>25</b>

## Аннотация

Модель регрессии матричного следа используется в квантовой томографии, восстановлении матриц и многих других областях знаний. В исследовании рассматривается проблема решения данной задачи, которая является обобщением хорошо известной и исследованной линейной регрессии, в случае, когда признаки записываются в виде матрицы. В работе матрица ковариаций задаётся в виде сумм произведений Кронекера, рассматриваются некоторые её оценки, учитывающие структуру, и проверяется качество индуцированной оценки параметров модели. Другим направлением развития работы является структуризация самой матрицы признаков, которая позволяет свести исходную задачу к задаче оптимального транспорта и решить её на основе барицентров Бюреса-Вассерштейна. В работе проводятся численные эксперименты с целью проверить эффективность энтропийной регуляризации в такой постановке.

## Abstract

The trace regression model is used in quantum tomography, matrix reconstruction and many other fields of knowledge. The study considers the problem of solving this problem, which is a generalization of the well-known and studied linear regression, in the case when the features are written in the form of a matrix. In this paper, the covariance matrix is given as sums of Kronecker products, some of its estimates that take into account the structure are considered, and the quality of the induced estimate of the model parameters is checked. Another direction of development of the work is the structuring of the feature matrix itself, which allows us to reduce the initial problem to the problem of optimal transport and solve it on the basis of Bures-Wasserstein barycenters. Numerical experiments are carried out in order to test the effectiveness of entropy regularization in such a formulation.

## Ключевые слова

линейная регрессия, произведение Кронекера, структурные векторы признаков, оценка матрицы ковариаций, задача оптимального транспорта

# 1 Введение

Наше исследование сосредоточено на модели регрессии матричного следа (англ. название *trace regression*) являющейся обобщением линейной регрессии, фундаментальной проблемы статистического моделирования и машинного обучения.

Данное обобщение является широко используемым методом моделирования зависимостей, особенно в сценариях, где результатом работы является матрица ([5], [1]). Стандартная и хорошо изученная модель линейной регрессии расширяется на случаи, когда входными данными для предсказания целевой переменной являются матрицы, а не векторы. В такой формулировке записываются различные задачи восстановления или заполнения матриц, а также модель применима во многих других областях, таких как финансы, медицина и обработка изображений, где повсеместно возникают матричные данные.

В нашей работе рассматриваются модели с дополнительной структурой у признаков. Одним из направлений исследования является задание структуры у матрицы ковариаций для распределения признаков переменных, записанных в виде вектора и соответственно восстановление решения исходной задачи через построение оценки этой матрицы ковариаций. Для задания структуры существует множество вариантов ограничений на множество матриц, включая как классические ограничения на разреженность и малоранговость, так и более специфические структуры: матрицы Тёплица, циркулянтные матрицы, Ганкелевы матрицы. Работа опирается на статьи [7] и [10], описывающие оценки матриц ковариаций в виде сумм произведений Кронекера двух меньших матриц. Такой подход к заданию структуры позволяет значительно сократить число параметров модели, интерпретировать результат с точки зрения разделения значений и признаков осей. Будет экспериментально исследоваться качество работы данных оценок, а затем качество предсказаний параметров модели построенных по ним.

Другое направление состоит в наложении структуры непосредственно на матрицу признаков. Здесь наша работа опирается на связь поставленной задачи с задачей оптимального транспорта и на соответствующие статьи [9], [6] и [2]. Подход ограничен множеством возможных матриц признаков, а именно тем, что матрица признаков должна быть одноранговой и неотрицательно определённой. Благодаря таким требованиям исходная задача сводится к задаче оптимального транспорта и решается нахождением барицентра Вассерштейна. В работе делается попытка извлечь пользу из разработанной теории для задачи оптимального транспорта с целью применить её к рассматриваемой нами модели. Для этого переносится метод регуляризации энтропии, значительно улучшающий сходимость градиентного спуска при поиске барицентра Вассерштейна.

## 2 Обзор литературы

Как уже упоминалось ранее, рассматриваемая проблема является обобщением линейной регрессии на случай, когда независимые переменные записаны в виде матриц, а не векторов. Существуют и активно развиваются различные модификации данной постановки задачи. Например в [5] исследуется ещё более общий случай, где в качестве целевых переменных выступают матрицы, а не числа. В [1] используется аналогичная постановка с матричными значениями и дополнительными предположениями о структуре параметров модели. Развиваются и другие направления модификаций, связанные с добавлением нестандартных норм для регуляризации, как в [4]. Наша работа сосредоточена на рассмотрении задачи при условии наличия некоторой структуры у признаковых данных.

Структуру можно задавать для матрицы ковариаций, предполагая нормальность распределения признаков. По известной матрице ковариаций можно восстановить ответ на задачу, то есть матрицу параметров исходной модели,

используя набор данных. Таким образом, алгоритм заключается в оценке матрицы ковариаций с учётом её структуры. В предыдущем разделе отмечалось, что наша работа опирается на статьи [7] и [10]. В первой делается предположение, что структура - одно произведение Кронекера двух матриц, и строится соответствующая оценка. Во второй используется структура уже в виде суммы нескольких произведений Кронекера и строится более общая оценка, получаемая через поиск малорангового приближения предварительно найденной выборочной оценки. Различные алгоритмы поиска приближения дают соответственно различные вариации оценки, как в [8] и [11].

Другой вариант задания структуры использует малоранговое разложение для матрицы признаков. Работа основана на идее сведения задачи регрессии к задаче оптимального транспорта. Неформально данная задача заключается в поиске оптимального способа перевозки однородного товара из одного набора точек в другой с заданными объёмами у каждой точки. Часто рассматривается постановка, когда сумма объёмов в двух наборах совпадает, то есть нужно переместить весь товар полностью. С учётом данного предположения можно отнормировать объёмы, разделив на их сумму и таким образом формально ввести вероятностные распределения, а задачу определить как минимизацию некоторой функции, зависящей от двух распределений. К качеству результата предъявляется план перевозок, представляющий собой в дискретном случае матрицу со значениями для всех возможных пар исходных и конечных точек. Минимум оптимизируемой функции соответствует расстоянию Вассерштейна между двумя введёнными распределениями. Для случая нормальных распределений можно преобразовать формулу для данного расстояния и выписать его в конкретном виде для двух матриц ковариаций, то есть по сути для любых двух неотрицательно определённых матриц. Такое расстояние уже называется расстоянием Бюреса-Вассерштейна и позволяет построить связь между регрессионной задачей и задачей оптимального транспорта [9]. После сведения одной задачи к другой, используется метод градиентного спуска для поиска барицен-

тра Вассерштейна [2] по которому строится решения исходной задачи.

В нашей работе рассматривается известная модификация задачи оптимального транспорта, заключающаяся в добавлении энтропийной регуляризации [3] к минимизируемой функции. Данный подход значительно улучшает сходимость градиентного спуска и позволяет контролировать форму итогового решения. Мы используем регуляризацию энтропии, а также  $L_2$ -регуляризацию применительно к решению задачи регрессии и проверим работоспособность такого подхода после сведения к оптимальному транспорту.

### 3 Теоретическое описание методов

В данной работе рассматриваются две вариации модели линейной регрессии. Обе модели можно записать в общем виде (в англоязычной литературе это соответствует модели trace regression) как:

$$y_i = \text{Tr}(X_i^T \theta) + \varepsilon_i \quad (1)$$

где  $\theta, X_i \in \mathbb{R}^{p \times q}$  и  $\varepsilon_i$  - некоторый случайный шум,  $\text{Tr}$  - след матрицы квадратной матрицы, равный сумме элементов на главной диагонали. Модель задаёт зависимость переменной  $y$  от нескольких переменных, записанных в виде матрицы  $X$  и имеет несколько параметров, записанных в матрицу  $\theta$ . Задача заключается в построении такой модели на основе имеющихся данных, то есть в оценке матрицы  $\theta$  по набору данных  $X_i$  и  $y_i$ . Далее описываются два рассматриваемых в данной работе подхода к решению задачи и соответственно две возникающие вариации модели. В первом подходе используется сведение к стандартной задаче линейной регрессии с помощью векторизации матриц, а затем исследуются методы оценки матрицы ковариаций, по которой уже восстанавливается искомый параметр модели. Во втором устанавливается связь с задачей оптимального транспорта, и исходная задача решается с помощью методов из несколько



другой области.

## 3.1 Оценка матриц ковариаций

### 3.1.1 Постановка задачи

Как было упомянуто ранее, описанная модель (1) сводится к оригинальной модели линейной регрессии при помощи операции векторизации. Это операция, преобразующая входную матрицу в вектор путём записи её столбцов друг за другом. Формально, для матрицы  $A \in \mathbb{R}^{m \times n}$  получим:

$$vec(A) = [A_{1,1}, \dots, A_{m,1}, A_{1,2}, \dots, A_{m,2}, \dots, A_{1,n}, \dots, A_{m,n}]^T$$

Теперь перепишем формулу (1) в виде:

$$y_i = Tr(X_i^T \theta) + \varepsilon_i = vec(X_i)^T vec(\theta) + \varepsilon_i \quad (2)$$

что уже является классической моделью линейной регрессии с коэффициентами  $vec(\theta)$ . Для оценки параметров воспользуемся одним из базовых методов решения таких задач - методом наименьших квадратов. Суть его работы заключается в нахождении параметров, которые минимизируют сумму квадратов разности предсказанных значений зависимой переменной и их действительных значений. Также к минимизируемой функции добавляется регуляризация, например частым выбором является  $L_2$ -регуляризация. Итоговая задача оптимизации записывается в виде:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \left( \frac{1}{n} \sum_{i=1}^n (y_i - vec(X_i)^T vec(\theta))^2 + \lambda \|\theta\|_F^2 \right) \quad (3)$$

где  $\|\theta\|_F = \sqrt{\sum_{i=1}^p \sum_{j=1}^q \theta_{i,j}^2} = \|vec(\theta)\|_F$  - норма Фробениуса и  $\lambda$  - коэффициент регуляризации.

Данная задача имеет аналитическое решение, записываемое в виде:

$$vec(\hat{\theta}) = (\hat{\Sigma} + \lambda I_n)^{-1} \left( \frac{1}{n} \sum_{i=1}^n vec(X_i) y_i \right), \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n vec(X_i) vec(X_i)^T \quad (4)$$

Пусть элементы матрицы  $X_i$  задаются многомерным нормальным распределением с нулевым смещением, то есть  $vec(X_1) \sim \mathcal{N}(0, \Sigma)$ , тогда получается, что  $\hat{\Sigma}$  в формуле (4) является выборочной оценкой матрицы ковариаций  $\Sigma$ . В следующих разделах рассматриваются другие методы оценки матрицы ковариаций, по которым уже получаются решения исходной задачи (1) используя формулу (4). Обе оценки основаны на предположении наличия некоторой структуры у  $\Sigma$ , а именно - в виде суммы произведения Кронекера двух неотрицательно определённых матриц:

$$\Sigma = \sum_{i=1}^K A_i A_i^T \otimes B_i B_i^T, \quad A_i \in \mathbb{R}^{q \times q}, B_i \in \mathbb{R}^{p \times p} \quad (5)$$

где  $\otimes$  - произведение Кронекера, определённое как  $C \otimes D = \begin{bmatrix} c_{11}D & \cdots & c_{1n}D \\ \vdots & \ddots & \vdots \\ c_{m1}D & \cdots & c_{mn}D \end{bmatrix}$ ,

$C \in \mathbb{R}^{m \times n}$ . В формуле (5) в качестве неотрицательно определённых матриц выступают  $A_i A_i^T$  и  $B_i B_i^T$  и согласно свойствам произведения Кронекера результирующая матрица ковариаций также получается неотрицательно определённой, что обеспечивает корректность предположения наличия данной структуры.

### 3.1.2 Оценка произведения Кронекера

Для начала приведём оценку для более простого случая, когда в формуле (5)  $K = 1$  и соответственно можно записать  $\Sigma = A A^T \otimes B B^T$ . Далее можно воспользоваться преобразованием  $X_i = B Z_i A^T$ , где  $Z_i \in \mathbb{R}^{p \times q}$ ,  $vec(Z_1) \sim$

$\mathcal{N}(0, I_{pq})$ . Тогда в статье [7] предлагается оценка, основанная на том, что

$$\mathbb{E}(\Phi_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i^T X_i\right) = \text{Tr}(BB^T)AA^T$$

$$\mathbb{E}(\Psi_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i X_i^T\right) = \text{Tr}(AA^T)BB^T$$

Наконец, оценив  $\text{Tr}(AA^T)\text{Tr}(BB^T)$  как  $\frac{1}{n} \sum_{i=1}^n \|X_i\|_F^2$  можно получить итоговую оценку на матрицу ковариаций:

$$\tilde{\Sigma} = (\Phi_n \otimes \Psi_n) / \frac{1}{n} \sum_{i=1}^n \|X_i\|_F^2 \quad (6)$$

Таким образом, при предположении о наличии структуры у матрицы ковариаций в виде произведения Кронекера двух меньших матриц можно записать оценку, которая в отличие от стандартной оценки непосредственно строит матрицу с такой структурой.

### 3.1.3 Оценка суммы произведений Кронекера

Для более общего случая  $K > 1$  предыдущий метод не обобщается, однако здесь можно рассмотреть несколько другие подходы. В работе будет использоваться оценка матрицы ковариаций, рассматриваемая в статье [10]. Идея метода заключается в том, чтобы использовать некоторый обратимый оператор  $\mathcal{R} : \mathbb{R}^{pq \times pq} \rightarrow \mathbb{R}^{q^2 \times p^2}$  для которого будет верно  $\mathcal{R}(AA^T \otimes BB^T) = \text{vec}(AA^T)\text{vec}(BB^T)^T$ , тогда для структурной матрицы ковариаций (5) можно записать:

$$\mathcal{R}(\Sigma) = \sum_{i=1}^K \mathcal{R}(A_i A_i^T \otimes B_i B_i^T) = \sum_{i=1}^K \text{vec}(A_i A_i^T) \text{vec}(B_i B_i^T)^T \quad (7)$$

Из формулы (7) следует что ранг матрицы  $\mathcal{R}(\Sigma)$  равен  $K$ , поэтому алгоритм заключается в нахождении  $K$ -рангового приближения матрицы  $\mathcal{R}(\hat{\Sigma})$  и после-

дующем обратном преобразовании  $\mathcal{R}^{-1}$  для получения оценки матрицы ковариаций  $\Sigma$ . Для нахождения малорангового приближения матрицы в данной работе используется алгоритм поиска сингулярного разложения (SVD), которое обрезается до  $K$  наибольших сингулярных чисел. Существуют и другие вариации округления SVD применительно к данной оценке, например в работе [11] несколько переписывается задача оптимизации и используются все сингулярные числа, превосходящие заданный параметр, из-за чего однако полученная матрица может не иметь ранг в точности равный  $K$ . Таким образом, данная оценка строится как приближение стандартной оценки матрицы ковариаций, но с учётом структуры может обеспечивать лучшее качество относительно исходной матрицы  $\Sigma$ .

## 3.2 Применение Барицентров Вассерштейна

### 3.2.1 Постановка задачи

Здесь описанная модель (1) используется при условиях  $p = q$  и  $X_i = x_i x_i^T$ , где  $x_i \in \mathbb{R}^p$ , то есть в качестве  $X_i$  могут использоваться только одноранговые неотрицательно определённые матрицы. Более того, в качестве матрицы параметров  $\theta$  также используется неотрицательно определённая матрица. Это значительное ограничение, однако только в таком случае сведения данной задачи к задаче оптимального транспорта, предложенное в [9]. Также рассматривается модель без шума, а значит  $\varepsilon_i = 0$  для любого  $i$  в упомянутой формуле.

В работе [9] подробно описывается способ сведения нашей задачи к задаче оптимального транспорта и возникающая интуиция подхода. Здесь мы кратко опишем детали алгоритма для нахождения непосредственно матрицы параметров  $\theta$ . Для нахождения минимизируется функция:

$$F(\theta) = \frac{1}{2n} \sum_{i=1}^n d_{BW}^2(X_i, \theta) \quad (8)$$

Как упоминалось ранее, оптимизация данной функции производится с помощью градиентного спуска. Учитывая свойства матрицы  $\theta$ , а именно неотрицательную определённую можно записать  $\theta = UU^T$ . Это позволяет более вычислительно эффективно реализовать шаг градиентного спуска и записать его в виде:

$$U_{k+1} = (1 - \nu_k)U_k + \nu_k \frac{1}{n} \sum_{i=1}^n \frac{x_i x_i^T U_k}{\|U_k^T x_i\|} \quad (9)$$

где  $\nu_k$  - вес текущего шага, обычно полагаемый равным 1.

### 3.2.2 Добавление регуляризации

Метод предложенный в [3] заключается в добавлении к минимизируемой функции энтропии транспортного плана (как многомерного распределения) с некоторым коэффициентом. В нашем случае это соответствует добавлению к функции (8) энтропии нормального распределения с матрицей ковариаций  $\theta$ . По итогу к шагу градиентного спуска (9) необходимо прибавить величину:

$$\nu_k \cdot \lambda (U_k U_k^T + I_p)^{-1} U_k \quad (10)$$

где  $\lambda$  - коэффициент регуляризации и  $= 10^{-3}$  - фиксированное небольшое число, необходимое для возможности обращения матрицы. Аналогично в работе предлагается добавить  $L_2$ -регуляризацию матрицы  $\theta$  в формуле (8) и это соответствует следующей добавке к шага градиентного спуска:

$$-\nu_k \cdot \lambda U_k U_k^T U_k \quad (11)$$

В работе планируется экспериментально проверить данные способы регуляризации.

## 4 Эксперименты

Предложенные методы были реализованы на языке программирования python с использованием библиотек numpy и scipy для проведения численных экспериментов. Анализ проводился с использованием синтетических данных.

### 4.1 Оценка структурных матриц ковариаций

#### 4.1.1 Данные

В данной постановке задачи используется модель (2), соответственно нужно задать параметры распределения  $X_i$ , параметры модели, а затем сгенерировать данные. В качестве элементов матрицы  $\theta$  возьмём независимые стандартные нормальные распределения. Матрица ковариаций  $\Sigma$  задаётся формулой (5), элементы матриц  $A_i, B_i$  также независимые стандартные нормальные распределения. Обе матрицы параметров  $\theta$  и  $\Sigma$  поэлементно делятся на их нормы Фробениуса, чтобы обеспечить более удобное сравнение методов при разных размерах матриц. Набор данных генерируется сначала сэмплированием  $vec(X_i) \sim \mathcal{N}(0, \Sigma)$ , а затем вычислением  $y_i$  по формуле (2). В качестве шума  $\varepsilon_i$  используется стандартное нормальное распределение, домноженное на некоторый коэффициент  $c_{noise}$ . При  $c_{noise} = 0$  шум отсутствует, и при наличии достаточного количества данных стандартная оценка позволяет найти точное решение (не лучше машинной точности в случае программной реализации) путём решения системы уравнений, поэтому в данном разделе будет рассматриваться ненулевой шум.

#### 4.1.2 Оценка матрицы ковариаций

Прежде всего, необходимо экспериментально проверить качество предсказаний матрицы ковариаций, которая в дальнейшем будет использоваться для предсказания искомых параметров  $\theta$ . Влияние случайного шума и коэффици-

ента  $c_{noise}$  здесь отсутствует, так как значения  $y_i$  не используются в оценках матрицы ковариаций. В качестве ошибки будет использоваться Фробениусова норма разности между матрицей ковариаций и её оценкой  $\|\hat{\Sigma} - \Sigma\|_F$ , которая также равна  $\frac{\|\hat{\Sigma} - \Sigma\|_F}{\|\Sigma\|}$  из-за нормировки при генерации параметров.

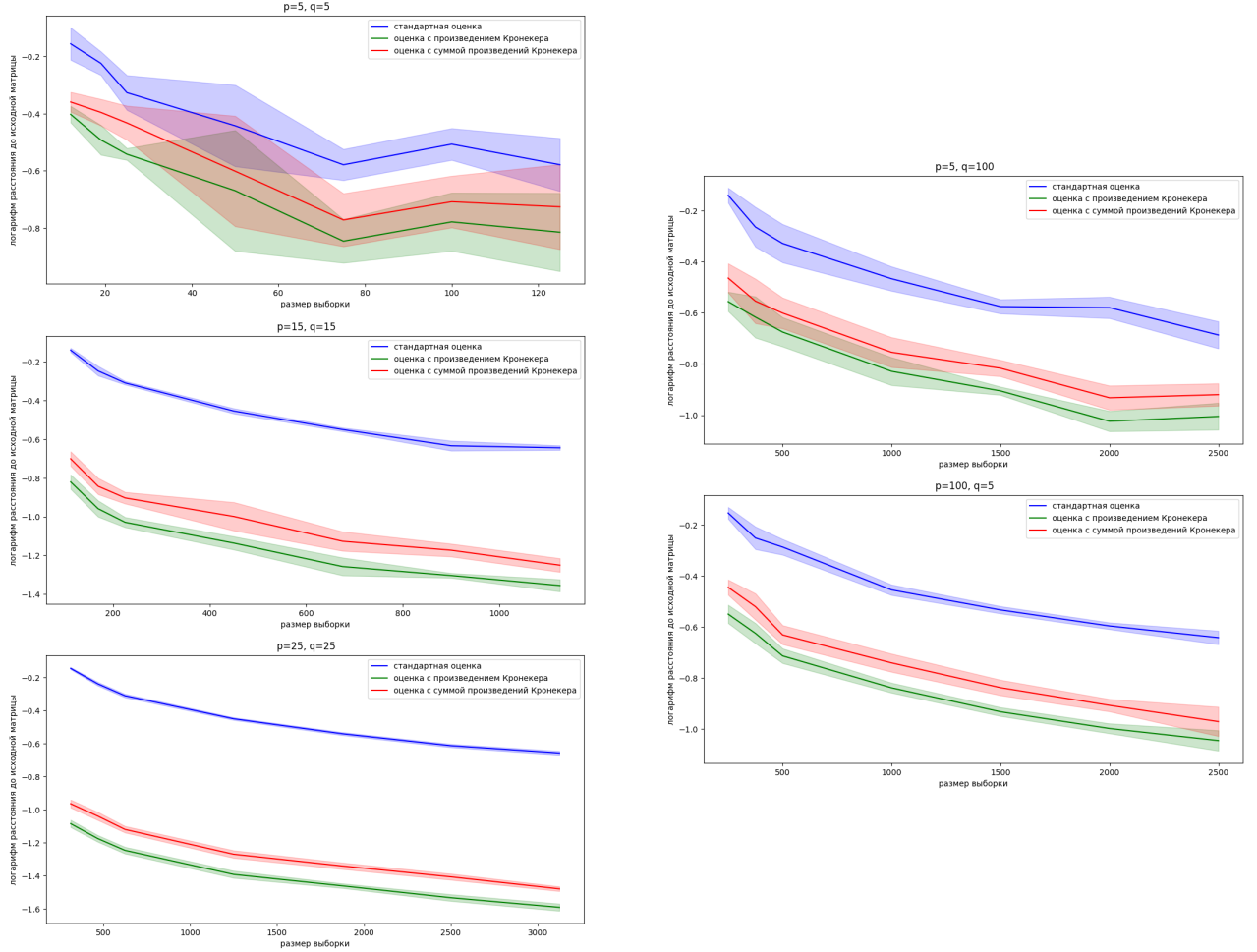


Рис. 4.1: Зависимость десятичного логарифма расстояния Фробениуса между ковариационной матрицей и её оценкой от объема выборки. Матрица ковариаций задана по формуле (5) при  $K = 1$ . Отображается среднее значение ошибки по 5 запускам, а также дисперсия.

Сначала проверим работу разных методов оценок при  $K = 1$ , то есть когда  $\Sigma = AA^T \otimes BB^T$ . На Рис. 4.1 приведены графики зависимости ошибок от размера выборки для различных размеров матриц  $p$  и  $q$ . Отображены размеры выборки от  $0.5 \cdot p \cdot q$  до  $5 \cdot p \cdot q$ . Можно видеть что при любых размерах матриц оба метода оценки матрицы ковариаций, использующие структуру, работают лучше стандартной оценки, а также метод, предназначенный исключительно

для оценки случаев с  $K = 1$ , работает несколько лучше более общего метода для произвольных  $K$ . При увеличении размерностей разница в качестве в сравнении со стандартной оценкой значительно увеличивается и может достигать одного порядка на приведённых графиках.

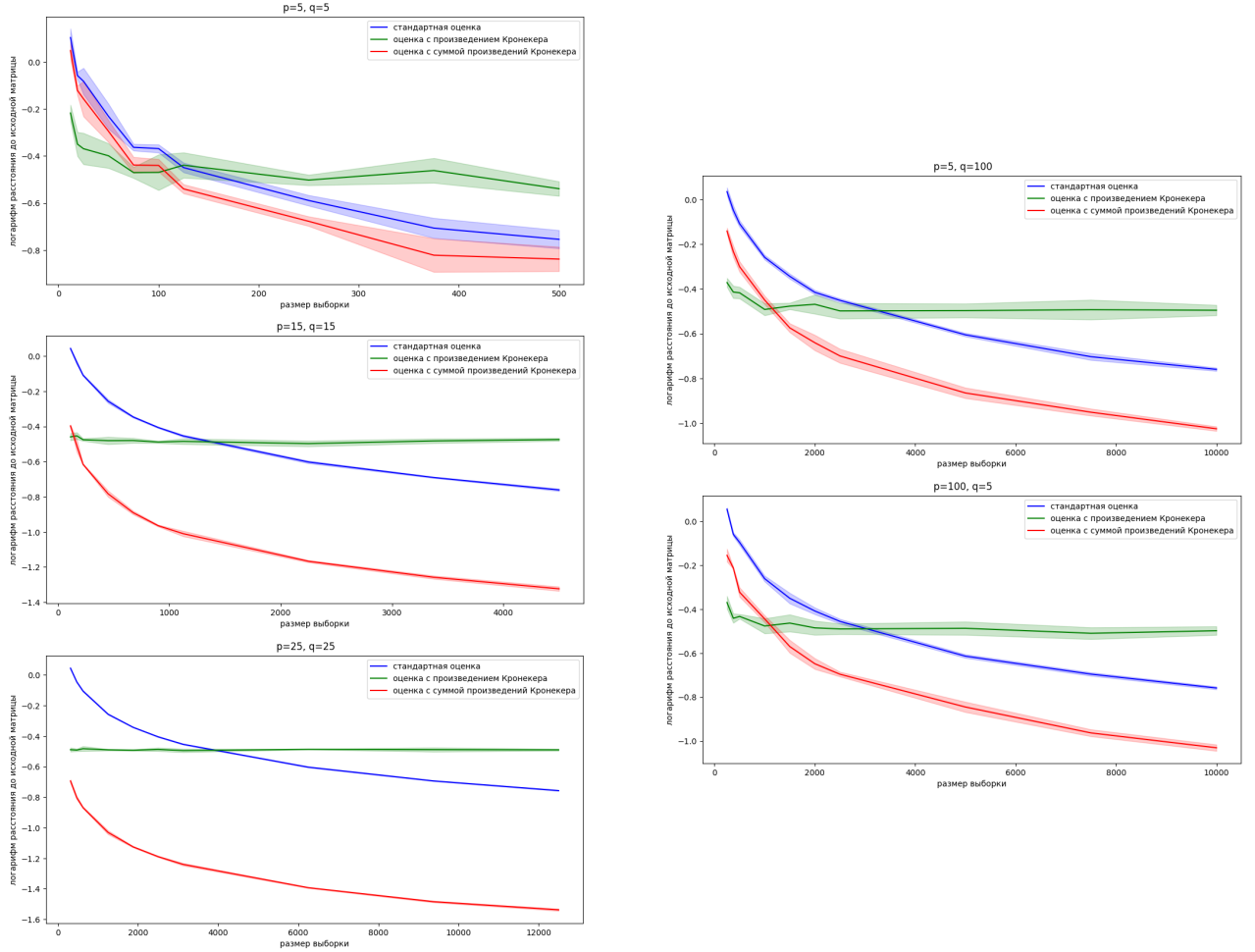


Рис. 4.2: Зависимость десятичного логарифма расстояния Фробениуса между ковариационной матрицей и ее оценкой от объема выборки. Матрица ковариаций задана по формуле (5) при  $K = 5$ . Отображается среднее значение ошибки по 5 запускам, а также дисперсия.

Теперь приведём результаты для  $K > 1$ . Так, на Рис. 4.2 приведены аналогичные предыдущим графики, однако отображены размеры выборки от  $0.5 \cdot p \cdot q$  до  $20 \cdot p \cdot q$ , чтобы более полно отразить тенденции при росте объёма данных. Для любых размеров  $p$  и  $q$  можно заметить, что оценка, предполагающая  $K = 1$ , теперь упирается в некоторый порог качества при росте размера выборки. Для других двух оценок такого не происходит: они показывают схо-



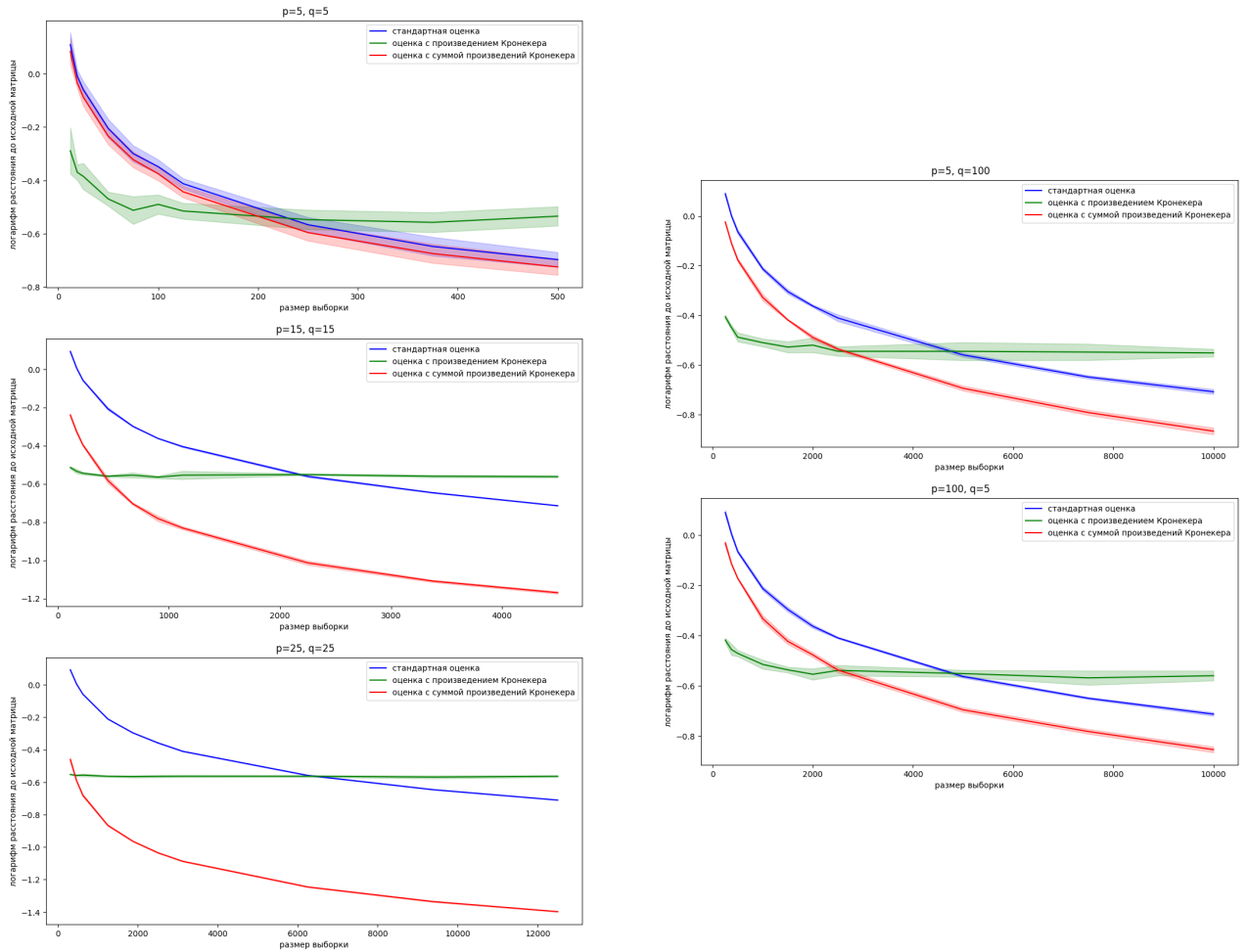


Рис. 4.3: Зависимость десятичного логарифма расстояния Фробениуса между ковариационной матрицей и ее оценкой от объема выборки. Матрица ковариаций задана по формуле (5) при  $K = 10$ . Отображается среднее значение ошибки по 5 запускам, а также дисперсия.

жие тенденции падения ошибки, однако видно, что метод с оценкой суммы произведений Кронекера стабильно достигает лучших результатов, чем стандартная оценка. Такие же выводы получаются при экспериментах с большими значениями  $K$ . Например, на Рис. 4.3 представлены результаты для  $K = 10$ , и результаты получаются аналогичные. Также дополнительно подтверждается, что преимущество перед стандартным методом оценки наблюдается при росте размерностей, так как для небольших матриц  $25 \times 25$  ковариаций при  $p = q = 5$  графики ошибок очень близки, и находить SVD вычислительно невыгодно при незначительном улучшении качества.

### 4.1.3 Оценка параметров модели

Последним этапом решения исходной задачи является восстановление параметров модели по формуле (4), используя полученную оценку матрицы ковариаций. Соответственно, в данном разделе будет производиться анализ качества предсказаний исходных параметров модели  $\theta$ . Для корректного восстановления  $\theta$  нужно использовать коэффициент регуляризации  $\lambda > 0$ , чтобы обеспечить обратимость матрицы  $(\Sigma + \lambda I_n)$  для различных оценок. Заметим, что подбор данного коэффициента в некоторых случаях может значительно изменить результат, особенно когда получается ошибка  $\|\hat{\theta} - \theta\|_F > 1 = \|0_{p \times q} - \theta\|_F$  ( $0_{p \times q}$  - матрица из нулей размера  $p \times q$ ), однако в таких случаях улучшение будет достигнуто за счёт приближения матрицы к нулевой, а не за счет улучшения предсказания. Учитывая это, во всех экспериментах будет использоваться фиксированный небольшой коэффициент  $\lambda = 10^{-6}$ . В качестве ошибки будет аналогично предыдущему разделу использоваться Фробениусова норма разности между матрицей параметров модели и её оценкой  $\|\hat{\theta} - \theta\|_F = \frac{\|\hat{\theta} - \theta\|_F}{\|\theta\|_F}$  (так как  $\|\theta\|_F = 1$ ). Для сравнения в эксперименты также добавляется оценка матрицы  $\theta$  полученная на основе настоящей матрицы ковариаций  $\Sigma$  по формуле (4).

В случае с оценкой параметров модели есть проблемы, так как при нулевом шуме ( $c_{noise} = 0$ ) нахождение  $\theta$  на основе стандартной оценки по сути решает задачу минимизации (3). Предположения о структуре матрицы ковариаций не улучшат предсказание, так как для нахождения  $\theta$  не нужна информация про распределение  $X_i$  и, соответственно, про какие-либо его параметры. При добавлении шума идеального решения уже не существует, и знания о структуре матрицы ковариаций могут приводить к уменьшению ошибки. Так, на графиках Рис. 4.4 приведены результаты экспериментов аналогичных Рис. 4.2, но теперь по построенным ранее оценкам матрицы  $\sigma$  строятся оценки матрицы  $\theta$ . Отображены размеры выборки от  $0.5 \cdot p \cdot q$  до  $5 \cdot p \cdot q$ . Как видно из графиков, стандартная оценка практически всегда оказывается лучше всех других оценок,

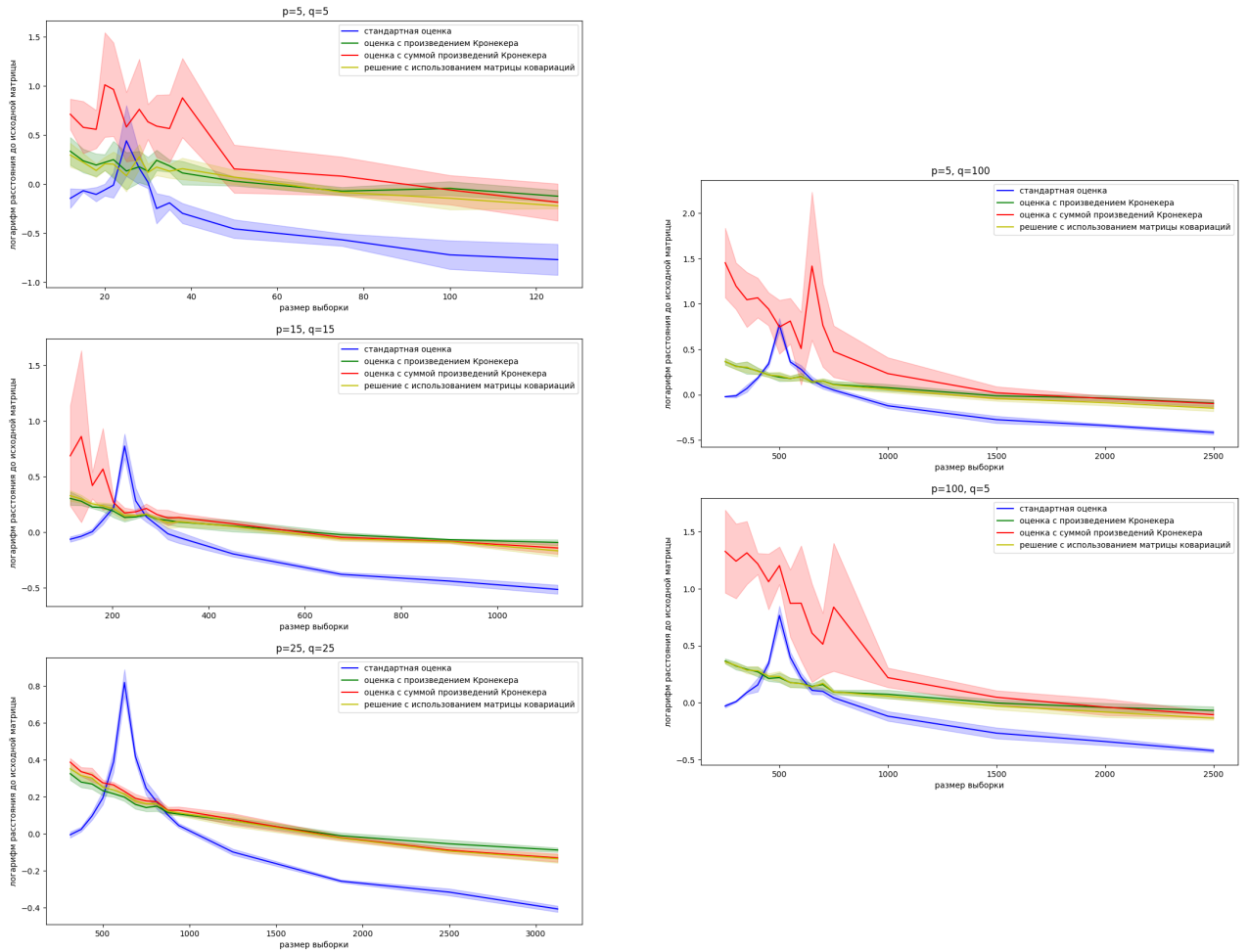


Рис. 4.4: Зависимость десятичного логарифма расстояния Фробениуса между матрицей параметров модели  $\theta$  и ее оценкой от объема выборки. Матрица ковариаций задана по формуле (5) при  $K = 5$ . Для всех графиков  $c_{noise} = 0.1$ . Отображается среднее значение ошибки по 5 запускам, а также дисперсия.

включая построение  $\theta$ , подставляя в формулу (4) настоящую матрицу  $\Sigma$ , вместо её оценки. Также оценка для произвольного числа слагаемых  $K$  оказывается менее стабильной и хуже работающей с точки зрения качества по сравнению с оценкой для предположения  $K = 1$  даже при условии, что действительная матрица ковариаций сгенерирована с  $K = 5$ . Данный эффект не наблюдается для достаточно больших и совпадающих размеров матриц ( $p = q = 25$  на графике), так как все оценки кроме стандартной оказываются очень близки друг к другу. Наконец, можно наблюдать деградацию стандартной оценки в окрестности объёма выборки равного  $n = p \cdot q$ .

Для дальнейшего анализа преимуществ подходов с оценкой структурных

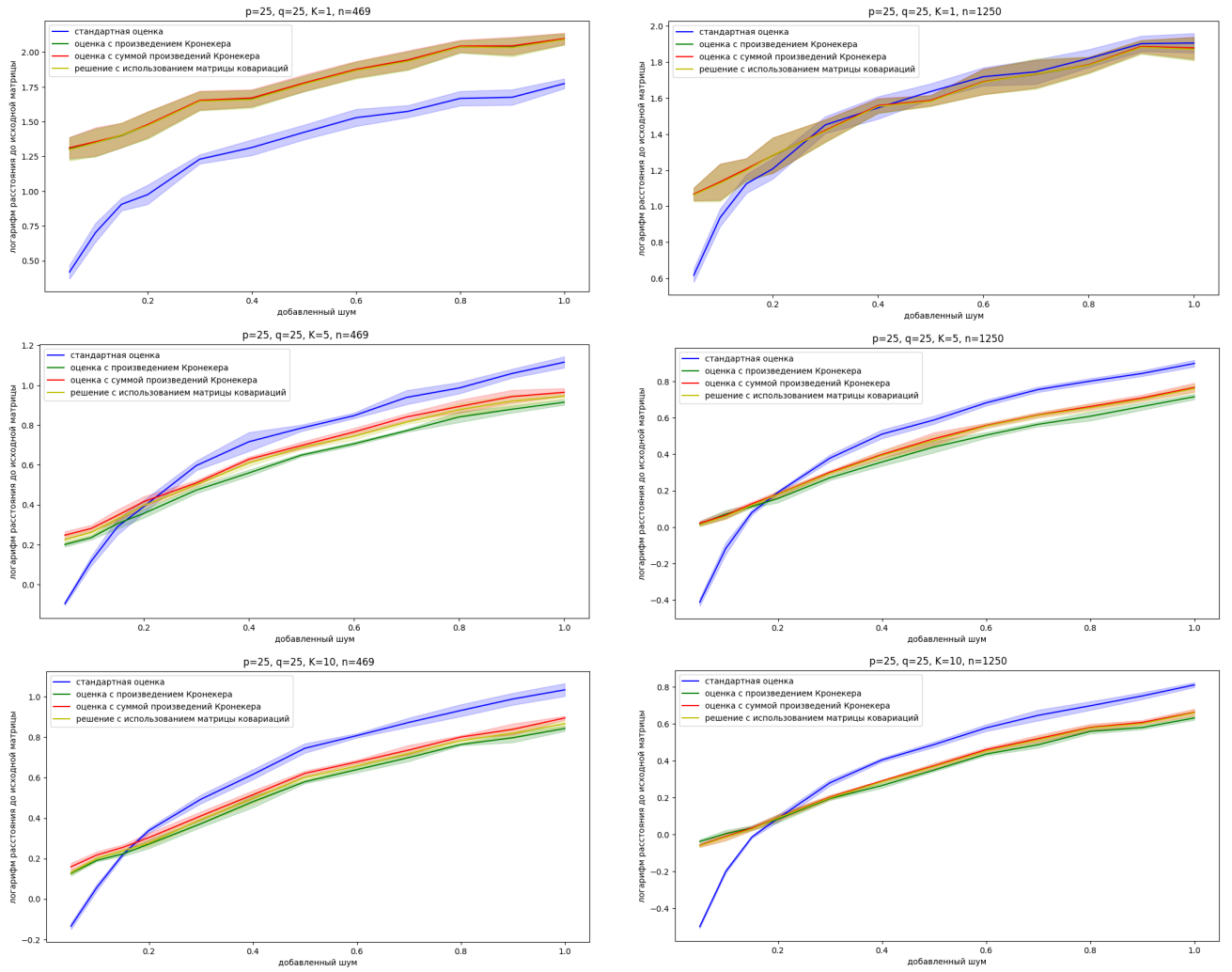


Рис. 4.5: Зависимость десятичного логарифма расстояния Фробениуса между матрицей параметров модели  $\theta$  и ее оценкой от добавленного шума. Матрица ковариаций задана по формуле (5) для различных  $K$ . Объем выборки  $0.75 \cdot p \cdot q$  в левом столбце и  $2 \cdot p \cdot q$  в правом. Отображается среднее значение ошибки по 5 запускам, а также дисперсия.

матриц построим графики зависимости ошибок от уровня шума. Учитывая выводы полученные выше, целесообразным будет исследовать работу алгоритмов на достаточно больших матрицах и рассматривать объёмы обучающей выборки близкие к  $p \cdot q$ . На графиках Рис. 4.5 представлены результаты для различных значений  $K$  (1, 5 и 10), а также различных объёмов выборок (округлённое  $0.75 \cdot p \cdot q$  и  $2 \cdot p \cdot q$ ). Отображается уровень добавленного шума от  $c_{noise} = 0.05$  до  $c_{noise} = 1$ . Можно видеть, что для матрицы ковариаций, сгенерированной с параметром  $K = 1$ , стандартная оценка оказывается не хуже или же намного лучше других оценок при всяких уровнях шума. При  $K \geq 5$  ситуация меняется,

и видно, что ошибка стандартной оценки растёт быстрее ошибок других оценок, поэтому после определённого уровня добавленного шума стандартная оценка становится хуже всех остальных, представленных на графиках. Обе оценки, предполагающие наличие структуры, очень близки по качеству друг к другу и также очень близки к оценке, полученной на основе настоящей матрицы ковариаций путём подстановки в формулу (4). Также стоит отметить, что для  $K > 1$  (то есть  $K = 5$  и  $K = 10$  на графиках) немного лучше оказывается именно оценка с предположением структуры в виде одного произведения Кронекера (то есть  $K = 1$ ), а значит, улучшение в оценке матрицы ковариаций не гарантирует улучшение оценки параметров модели.

#### 4.1.4 Выводы

Два рассмотренных метода обеспечивают лучшее качество оценки матрицы ковариаций при условии наличия у неё структуры в форме (5). Соответственно, при  $K = 1$  лучше работает первый рассмотренный метод с оценкой в виде произведения Кронекера, а при  $K > 1$  преимущество уже у более общего метода с оценкой в виде суммы произведений Кронекера. С оценкой матрицы параметров  $\theta$ , построенной по оценке матрицы ковариаций, уже нет стабильного улучшения и только при наличии достаточного уровня шума в данных, а также соответствующего объёма выборки, близкого к  $p \cdot q$ , оценки параметров модели превосходят стандартную оценку и показывают более предсказуемое качество при изменении количества данных. С другой стороны, оба метода за счёт лучшей оценки матрицы ковариаций строят оценку параметров модели близкую к полученной, согласно настоящей матрице ковариаций, поэтому данное свойство можно использовать, когда в задаче важен ответ с учётом предположения о структуре распределения данных и необходима устойчивость алгоритма к смещению выборки. Описанные эффекты более выражены, соответственно, при больших размерах матриц  $p$  и  $q$ , что делает подход перспективным для рассмотрения в работе с большими размерностями. Наконец, эксперименты по-

казывают необходимость баланса между качеством предсказания матрицы ковариаций и качеством предсказания параметров  $\theta$ , поскольку исходная матрица  $\theta$  обеспечивает оценку хуже стандартной оценки, а значит, уменьшение ошибки оценки матрицы ковариаций приводит к ухудшению ошибки для матрицы параметров.

## 4.2 Применение Барицентров Вассерштейна

Эксперименты проводятся на синтетических данных. Данные  $x_i$  семплируются из стандартного нормального распределения независимо. Также каждый элемент матрицы  $U \in \mathbb{R}^{p \times r}$  генерируется из стандартного распределения независимо. Итоговая матрица параметров  $\theta = UU^T$ . Целевая переменная  $y_i$  находится по формуле (1), где  $X_i = x_i x_i^T$  и  $\varepsilon_i = 0$ . Для инициализации используется также  $\theta_0 = U_0 U_0^T$ , и все элементы матрицы  $U_0 \in \mathbb{R}^{p \times r}$  независимые из стандартного нормального распределения. Устанавливается  $\lambda = 10^{-6}$  во всех запусках.

На графиках Рис. 4.6 представлены результаты нескольких экспериментов. Везде используется  $r = 4$ , размерность данных  $p$  либо 32 либо 64, и объём выборки принимает три значения:  $3 \cdot p \cdot r$ ,  $5 \cdot p \cdot r$ ,  $10 \cdot p \cdot r$ . В качестве алгоритмов используется базовый градиентный спуск, предложенный в [9], а также две модификации с добавлением, соответственно, энтропийной и  $L_2$ -регуляризаций, и все три алгоритма начинают работу из одинаковой начальной точки. По итогу можно видеть, что добавление регуляризаций не ускоряет сходимость, как это происходит в задаче оптимального транспорта. Кроме того, качество методов практически совпадает до некоторого порога, после чего стандартный метод продолжает уменьшать ошибку, а методы с регуляризациями выходят на плато. Можно предположить, что это происходит из-за того, что на таком уровне точности излишняя оптимизация под дополнительную функцию потерь ухудшает качество основной минимизируемой функции. Во время экспериментов

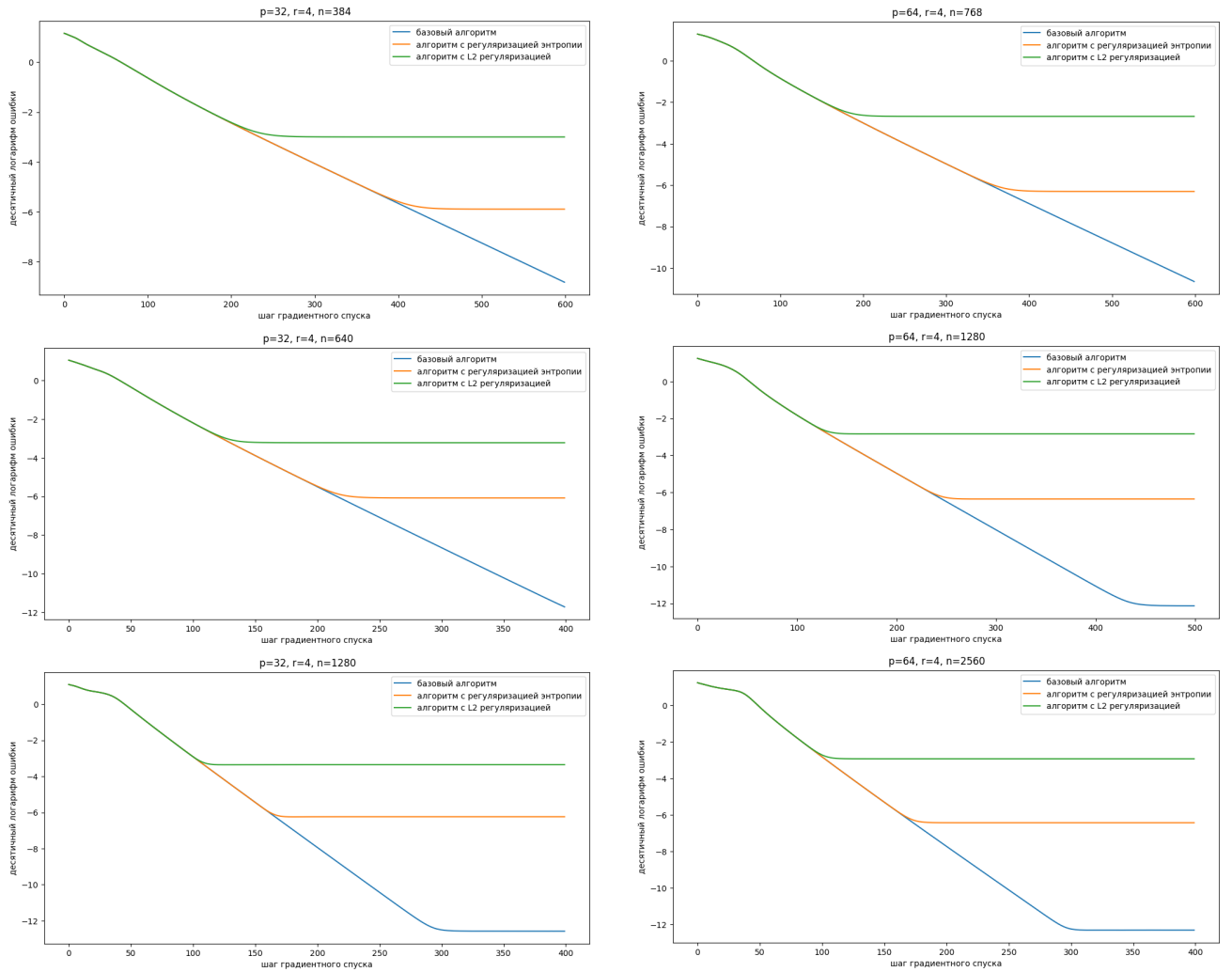


Рис. 4.6: Зависимость десятичного логарифма расстояния Фробениуса между корнями матрицы параметров модели  $\theta$  и ее оценкой от шага градиентного спуска.

удавалось найти примеры параметров генерации данных, при которых подход с энтропийной регуляризацией давал небольшое улучшение, однако это слишком нестабильное и незначительное улучшение, поэтому по итогу можно сделать вывод, что регуляризация из задачи оптимального транспорта не помогает в решения регрессионной задачи.

## 5 Заключение

В работе были рассмотрены различные подходы к структуризации векторов признаков: в одном случае мы добавляем дополнительную структуру у

признаков, накладывая ограничение на матрицы ковариаций их распределения, задающуюся в виде суммы произведений Кронекера, а в другом - накладываем структуру на матрицу признаков, чтобы свести задачу к задаче оптимального транспорта. Как показали результаты экспериментов, представление матрицы ковариаций в виде суммы произведений Кронекера действительно позволяет построить оценки, которые прежде всего намного лучше оценивают матрицу ковариаций по сравнению со стандартной выборочной оценкой, а также при определённых параметрах данных (когда объём данных сравним с  $p \cdot q$  и присутствует достаточный уровень шума в зависимой переменной) работают стабильнее и лучше с точки зрения предсказания исходных параметров модели. Второй подход со сведением к задаче оптимального транспорта и последующим решением с использованием расстояния Бюреса-Вассерштейна и регуляризацией энтропии не дал положительных результатов и соответственно метод регуляризации не помогают ускорить сходимость градиентного спуска в такой постановке.

В качестве дальнейших направлений можно выделить изучение рассмотренных подходов для более общих задач, таких как частичная регрессия матричного следа [5], работу над новыми оценками матрицы ковариаций при других её структурах, а также возможное развитие связи поставленной задачи с задачей оптимального транспорта с точки зрения поиска других более подходящих методов регуляризации.



## Список литературы

- [1] Yin-Jen Chen и Minh Tang. “Regression for matrix-valued data via Kronecker products factorization”. В: *arXiv preprint arXiv:2404.19220* (2024).
- [2] Sinho Chewi, Tyler Maunu, Philippe Rigollet и Austin J Stromme. “Gradient descent algorithms for Bures-Wasserstein barycenters”. В: *Conference on Learning Theory*. PMLR. 2020, с. 1276—1304.
- [3] Pavel Dvurechensky, Alexander Gasnikov и Alexey Kroshnin. “Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm”. В: *International conference on machine learning*. PMLR. 2018, с. 1367—1376.
- [4] Jianqing Fan, Wenyan Gong и Ziwei Zhu. “Generalized high-dimensional trace regression via nuclear norm regularization”. В: *Journal of econometrics* 212.1 (2019), с. 177—202.
- [5] Hachem Kadri, Stéphane Ayache, Riikka Huusari, Alain Rakotomamonjy и Ralaivola Liva. “Partial trace regression and low-rank kraus decomposition”. В: *International Conference on Machine Learning*. PMLR. 2020, с. 5031—5041.
- [6] Alexey Kroshnin, Vladimir Spokoiny и Alexandra Suvorikova. “Statistical inference for Bures–Wasserstein barycenters”. В: *The Annals of Applied Probability* 31.3 (2021), с. 1264—1298.
- [7] Chenlei Leng и Guangming Pan. “Covariance estimation via sparse Kronecker structures”. В: *Bernoulli* 24 (нояб. 2018), с. 3833—3863. DOI: [10.3150/17-BEJ980](https://doi.org/10.3150/17-BEJ980).
- [8] T Masak, S Sarkar и VM Panaretos. “Separable expansions for covariance estimation via the partial inner product”. В: *Biometrika* 110.1 (2023), с. 225—247.

- [9] Tyler Maunu, Thibaut Le Gouic и Philippe Rigollet. “Bures-Wasserstein Barycenter and Low-Rank Matrix Recovery”. В: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, с. 8183—8210.
- [10] Nikita Puchkin и Maxim Rakhuba. “Dimension-free Structured Covariance Estimation”. В: *arXiv preprint arXiv:2402.10032* (2024).
- [11] Theodoros Tsiligkaridis и Alfred Hero. “Covariance Estimation in High Dimensions Via Kronecker Product Expansions”. В: *Signal Processing, IEEE Transactions on* 61 (нояб. 2013), с. 5347—5360. DOI: [10.1109/TSP.2013.2279355](https://doi.org/10.1109/TSP.2013.2279355).