УДК 519.234.3 DOI: 10.18698/1812-3368-2021-3-23-45

СРАВНИТЕЛЬНЫЙ АНАЛИЗ КАЧЕСТВА РОБАСТНЫХ МОДИФИКАЦИЙ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ ПРИ СЖАТИИ КОРРЕЛИРОВАННЫХ ДАННЫХ

В.Б. Горяинов[1] vb-goryainov@bmstu.ru
Е.Р. Горяинова[2] el-goryainova@mail.ru

[1] МГТУ им. Н.Э. Баумана, Москва, Российская Федерация
[2] НИУ ВШЭ, Москва, Российская Федерация

Аннотация

Одним из традиционных методов решения задачи снижения размерности многомерного вектора с коррелированными компонентами является метод главных компонент. Построение главных компонент проводится с использованием специального представления ковариационной или корреляционной матрицы наблюдаемых показателей. В классическом методе главных компонент в качестве оценок элементов корреляционной матрицы используются выборочные коэффициенты корреляции Пирсона. Эти оценки крайне чувствительны к засорению выборки и наличию аномальных наблюдений. Для робастификации метода главных компонент предложено заменить выборочные оценки корреляционных матриц известными робастными аналогами, к числу которых относятся ранговый коэффициент Спирмена, MCD-оценки, ортогонализованные оценки Гнанадесикана — Кетенринга (OGK) и оценки Олива — Хокинса. Цель работы состоит в проведении численного сравнительного анализа классического метода главных компонент и его робастных модификаций. Для этого проведено моделирование девятимерных векторов с известной структурой корреляционных матриц и введена специальная метрика, позволяющая оценивать качество сжатия данных. Обширный численный эксперимент показал, что наилучшее качество сжатия при нормальном распределении наблюдений имеет классический метод главных компонент. Когда наблюдения имеют распределение Стьюдента с тремя степенями свободы, а также при наличии в данных кластера выбросов, отдельных аномальных наблюдений или симметрич-

Ключевые слова

Робастный метод главных компонент, MCD-оценка, оценка типа Гнанадесикана — Кетенринга, оценка Олива — Хокинса

ных засорений, описываемых распределением Тьюки, наилучшее качество сжатия показывают оценки Гнанадесикана — Кетенринга и Олива — Хокинса модификации метода главных компонент. Качество классического метода главных компонент и ранговой модификации Спирмена в этих случаях снижается

Поступила 16.07.2020
Принята 29.01.2021

Введение. Многие объекты исследования описываются большим числом показателей. Это может приводить к тому, что среди собранных данных появляются показатели, которые характеризуют одно и то же свойство объекта и поэтому являются коррелированными. Статистический анализ таких массивов становится затруднительным и может приводить к неверным результатам. В связи с этим возникает задача по описанию исходных показателей с использованием небольшого числа обобщенных интегративных показателей, сохранив при этом как можно больше важной информации об объектах. Широко используемым методом снижения размерности многомерных показателей является метод главных компонент (МГК). Суть МГК — нахождение небольшого числа главных компонент, представляемых в виде линейных комбинаций исходных показателей, изменчивость которых в значительной степени описывает изменчивость совокупности исходных показателей. В настоящее время МГК активно применяется для анализа многомерных данных, которые часто встречаются в биохимии [1], компьютерном зрении [2], генетике [2], экономике [3], при распознавании образов [4], обработке изображений [5] и в других областях. Отметим, что МГК часто является первым этапом при проведении дискриминантного и кластерного анализа и построении линейных регрессионных моделей с мультиколлинеарными регрессорами.

Классический МГК, использующий выборочные оценки ковариационных и корреляционных матриц, достаточно чувствителен к наличию аномальных наблюдений. Поэтому снижение размерности пространства показателей с использованием классического МГК становится ненадежным, если в наблюдениях присутствуют выбросы. Одним из важнейших способов устранения этого недостатка (см., например, [6] и [7]) является замена выборочных оценок ковариационных матриц их робастными аналогами. Первой робастной оценкой коэффициента корреляции следует, по-видимому, считать ранговый коэффициент корреляции, предложенный в [8]. В настоящее время наиболее распространенными робастными оценками ковариационных матриц являются предложенные в [9] MCD-оценки; оценки Гнанадесикана — Кетенринга (OGK-оценки) [10, 11] и использующие технику концентрации оценки Олива — Хокинса [12].

Применяя в МГК различные способы оценивания корреляционных матриц, необходимо определить, какой из способов будет давать наилучший результат в той или иной ситуации. Для проведения сравнительного анализа робастных модификаций МГК предложены различные подходы. В [7] сравнительный анализ основан на изучении поведения функций влияния собственных чисел и собственных векторов оцененных корреляционных матриц, в [13] рассмотрены усредненные квадраты отклонений фишеровских преобразований оценок коэффициентов корреляции от их истинных значений, в [14] — усредненные квадраты отклонений между истинными собственными значениями и соответствующими скорректированными оценками этих значений, в [6] для каждого метода введена мера относительной ошибки прогноза собственных значений. Таким образом, нет единой метрики, позволяющей оценивать качество МГК. Следует отметить, что МГК направлен не только на уменьшение объемов информации. Компонентный анализ полагают успешным, если построенные главные компоненты дают исследователю возможность достаточно четкой интерпретации этих компонент в качестве новых обобщенных показателей. Поэтому в настоящей работе для оценивания эффективности МГК введен показатель качества, измеряющий схожесть редуцированной матрицы нагрузок с эталонной матрицей, описывающей корреляционную структуру взаимосвязей между исходными показателями и построенными главными компонентами. Для проведения численного сравнительного анализа моделируются девятимерные векторы, состоящие из трех трехмерных подвекторов, компоненты которых коррелированы между собой, но некоррелированы с компонентами других подвекторов. С использованием статистического моделирования будет показано, что классический МГК является наилучшим (в смысле наименьшего значения функционала качества) в случае, когда наблюдения имеют нормальное распределение, а в случаях распределений, имитирующих различные типы засорения данных, лучшими оказываются модификации, применяющие в качестве оценок корреляционных матриц ортогонализованные оценки Гнанадесикана — Кетенринга и оценки Олива — Хокинса.

Метод главных компонент. Пусть каждый из n наблюдаемых объектов характеризуется r-мерным случайным вектором коррелированных показателей $X = (X_1, \ldots, X_r)^{\mathrm{T}}$. Требуется найти некоррелированные показатели $f_1, \ldots, f_k$, $k < r$, вариация которых описывает максимальную долю вариации исходных показателей $X_1, \ldots, X_r$.

Согласно МГК [15, 16], новые показатели $f_1, \ldots, f_k$ принадлежат классу $\mathcal{F}$ линейных ортогональных нормированных комбинаций центриро-

ванных исходных показателей $X_1,\ldots,X_r$. Компоненты вектора $F = (f_1,\ldots,f_k)^{\mathrm{T}} \in \mathcal{F}$ представляются в виде

$$f_j = a_{j1}\overset{\circ}{X}_1 + \ldots + a_{jr}\overset{\circ}{X}_r, \quad j = 1,\ldots,k,$$

где

$$\overset{\circ}{X}_i = X_i - \mathrm{E}X_i, \qquad \sum_{i=1}^{r} a_{ji}^2 = 1, \qquad \sum_{i=1}^{r} a_{ji}a_{mi} = 0 \ \text{ при } \ j \neq i.$$

Принцип МГК заключается в нахождении таких $\tilde{f}_1,\ldots,\tilde{f}_k$ из указанного класса $\mathcal{F}$, при которых будет достигаться максимальное значение функционала

$$\mu_k(f_1,\ldots,f_k) = \frac{\mathrm{D}f_1 + \ldots + \mathrm{D}f_k}{\mathrm{D}X_1 + \ldots + \mathrm{D}X_r}.$$

Соответственно случайные величины $\tilde{f}_1,\ldots,\tilde{f}_k$ такие, что

$$\mu(\tilde{f}_1,\ldots,\tilde{f}_k) = \max_{(f_1,\ldots,f_k)\in\mathcal{F}} \mu(f_1,\ldots,f_k),$$

называют главными компонентами системы показателей $X_1,\ldots,X_r$. Таким образом, среди всех случайных величин $\tilde{f}_1,\ldots,\tilde{f}_k \in \mathcal{F}$ первая главная компонента $\tilde{f}_1$ вносит наибольший вклад в суммарную дисперсию $\sum_{i=1}^{r} \mathrm{D}X_i$, а l-я главная компонента $\tilde{f}_l$, $l = 2,\ldots,r$, некоррелированная с $\tilde{f}_1,\ldots,\tilde{f}_{l-1}$, будет обладать наибольшей дисперсией среди всех $f \in \mathcal{F}$ некоррелированных с $\tilde{f}_1,\ldots,\tilde{f}_{l-1}$.

Опишем алгоритм построения главных компонент (см., например, [16]). Пусть K_X — ковариационная матрица вектора X. Обозначим через $\lambda_1,\ldots,\lambda_r$ собственные значения матрицы K_X, а через $a_j = (a_{1j},\ldots,a_{rj})^{\mathrm{T}}$, $j = 1,\ldots,r$, — нормированные собственные векторы, соответствующие собственным числам $\lambda_1,\ldots,\lambda_r$. Отметим, что, будучи симметричной и неотрицательно определенной, матрица K_X имеет r вещественных неотрицательных собственных значений. Предположим, что $\lambda_1 > \ldots > \lambda_r$. Введем диагональную матрицу Λ размером $r \times r$ с собственными числами на диагонали

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_r \end{bmatrix}$$

и матрицу A, столбцами которой являются собственные векторы $a_1,\ldots,a_r$. Матрица A является ортогональной, и $A^{\mathrm{T}}K_X A=\Lambda$, $\operatorname{tr}K_X=\operatorname{tr}\Lambda$.

Вектор главных компонент $F=(\tilde{f}_1,\ldots,\tilde{f}_r)^{\mathrm{T}}$ будет иметь вид

$$F=A^{\mathrm{T}}\overset{\circ}{X}, \tag{1}$$

где $\overset{\circ}{X}=(\overset{\circ}{X}_1,\ldots,\overset{\circ}{X}_r)^{\mathrm{T}}$. Отметим, что при таком построении ковариационная матрица K_F вектора главных компонент F будет равна

$$K_F=\mathrm{E}\left(FF^{\mathrm{T}}\right)=\mathrm{E}\left(A^{\mathrm{T}}\overset{\circ}{X}\overset{\circ}{X}^{\mathrm{T}}A\right)=A^{\mathrm{T}}K_X A=\Lambda.$$

Из последнего равенства следует, что построенные главные компоненты $\tilde{f}_1,\ldots,\tilde{f}_r$ некоррелированы, и $\mathrm{D}\tilde{f}_i=\lambda_i$, $i=1,\ldots,r$, а $\sum\limits_{i=1}^{r}\mathrm{D}\tilde{f}_i=\operatorname{tr}\Lambda=\sum\limits_{i=1}^{r}\lambda_i$. Поскольку

$$\sum_{i=1}^{r}\mathrm{D}X_i=\operatorname{tr}K_X=\operatorname{tr}\Lambda=\sum_{i=1}^{r}\lambda_i,$$

сумма дисперсий исходных показателей $X_1,\ldots,X_r$ полностью исчерпывается суммарной дисперсией главных компонент $\tilde{f}_1,\ldots,\tilde{f}_r$. В силу упорядоченности по убыванию собственных значений $\lambda_1>\ldots>\lambda_r$ оказывается, что каждая последующая главная компонента вносит меньший вклад в суммарную дисперсию исходных показателей, чем предыдущие главные компоненты. Таким образом, при описании исходных показателей можно пренебречь последними главными компонентами $\tilde{f}_{k+1},\ldots,\tilde{f}_r$, так как эти компоненты несут в себе малую часть информации об изменчивости показателей $X_1,\ldots,X_r$.

Число главных компонент k, которое следует выбрать для представления вектора $X=(X_1,\ldots,X_r)^{\mathrm{T}}$, обычно определяют одним из следующих эмпирических способов [17, с. 114]:

1) выбор числа k равным числу собственных значений матрицы K_X, которые принимают значения больше единицы;

2) выбор числа k равным такому числу, при котором значение

$$\mu_k\left(\tilde{f}_1,\ldots,\tilde{f}_k\right)=\frac{\sum\limits_{i=1}^{k}\mathrm{D}\tilde{f}_i}{\sum\limits_{i=1}^{r}\mathrm{D}X_i}=\frac{\sum\limits_{i=1}^{k}\lambda_i}{\sum\limits_{i=1}^{r}\lambda_i}$$

относительной доли дисперсии, вносимой первыми k компонентами в суммарную дисперсию исходных показателей, было бы, по мнению исследователя, достаточно близким к единице.

Построенные главные компоненты позволяют представить вектор X в виде

$$\overset{\circ}{X} = AF = A\Lambda^{1/2}\Lambda^{-1/2}F = L\Lambda^{-1/2}F = LF^{*}, \tag{2}$$

где $L = A\Lambda^{1/2}$ — детерминированная матрица с элементами l_{ij}, $1 \le i, j \le r$, а $F^{*} = \left(f_1^{*}, \ldots, f_r^{*}\right)^{\mathrm{T}}$ — нормированный вектор главных компонент с $f_i^{*} = \tilde{f}_i \big/ \sqrt{\lambda_i}$, $i = 1, \ldots, r$. Соотношение (2) называют линейной моделью главных компонент.

Отметим, что ковариационная матрица K_X вектора X полностью воспроизводится матрицей L, так как

$$\operatorname{cov}\left(X_i, X_j\right) = \mathrm{E}\,\overset{\circ}{X}_i\,\overset{\circ}{X}_j = \mathrm{E}\left(\sum_{m=1}^{r} l_{im} f_m^{*}\right)\left(\sum_{k=1}^{r} l_{jk} f_k^{*}\right) = \sum_{m=1}^{r} l_{im} l_{jm}.$$

Поскольку $\mathrm{D}\,\overset{\circ}{X}_i = \sum\limits_{m=1}^{r} l_{im}^2$, величины l_{im}^2, $m = 1, \ldots, r$, являются вкладами главных компонент f_m^{*} в дисперсию показателя X_i. Кроме того,

$$\operatorname{cov}\left(\overset{\circ}{X}_i,\, f_j^{*}\right) = E\left(\sum_{m=1}^{r} l_{im} f_m^{*}\right) f_j^{*} = l_{ij}.$$

Отметим (см. [16, с. 344–345], [18, с. 354]), что в случае, когда исходные показатели $X_1, \ldots, X_r$ имеют различные единицы измерения, следует провести нормировку и перейти к безразмерным величинам. Тогда ковариационной матрицей для нормированных случайных величин $X_1^{*}, \ldots, X_r^{*}$ будет являться корреляционная матрица R_X исходных величин $X_1, \ldots, X_r$, матрицей A в представлении (1) будет матрица, состоящая из соответствующих собственных векторов матрицы R_X, а элементы l_{ij} матрицы L будут коэффициентами корреляции $\rho_{ij} = \rho(X_i^{*}, f_j^{*})$ случайных величин X_i^{*} и f_j^{*}. В связи с этим элементы l_{ij} называют нагрузками i-го показателя на j-ю главную компоненту.

Еще одним важным аргументом относительно использования в МГК именно корреляционных матриц является наглядная интерпретируемость результатов сжатия данных. Таким образом, для построения главных компонент системы показателей $X_1, \ldots, X_r$ и получения линейного представле-

ния (2) требуется знать ковариационную или корреляционную матрицу вектора X. Как правило, на практике эти матрицы неизвестны. Поэтому при нахождении главных компонент матрицу K_X или R_X заменяют их оценками $\hat{K}_X$ или $\hat{R}_X$ соответственно. Традиционно в качестве таких оценок принято использовать выборочные ковариационные или корреляционные матрицы.

Свойства оценок главных компонент, построенных на базе выборочных оценок $\hat{K}_X$, в условиях, когда наблюдаемые показатели имеют нормальное распределение, подробно исследованы в [16, §13.4]. Однако распределение реальных данных может отличаться от нормального, данные могут содержать единичные выбросы или кластеры выбросов. Классические выборочные оценки чувствительно реагируют на такие отклонения. Для преодоления этого недостатка в [19] предложена робастная версия МГК, в которой вместо функционала $\mu_k(f_1,\ldots,f_k)$ использован робастный функционал Хьюбера. В настоящей работе предложено использовать такой важнейший (согласно, например, [6, 7]) подход к построению робастных главных компонент, который состоит в замене выборочных оценок ковариационных и корреляционных матриц их робастными аналогами. Известные робастные оценки корреляционных матриц представлены далее.

Оценивание ковариационных матриц. Робастными оценками будем полагать те оценки, которые обладают ненулевой пороговой точкой. Пороговой точкой (breakdown point) ε^* оценки называют наименьшую долю выбросов в выборке, которая может привести к тому, что оценка будет принимать произвольные значения. Формальное определение пороговой точки для оценки ковариационной матрицы приведено в [11, с. 309]. Пороговая точка, равная 0,5, — лучшее значение, которое можно ожидать от оценки, поскольку при большей доле загрязнения становится невозможным провести различие между «хорошей» и «плохой» частями выборки.

Обозначим X_{ij} — результат измерения i-й компоненты вектора $X=(X_1,\ldots,X_r)^{\mathrm{T}}$ для j-го наблюдаемого объекта $j=1,\ldots,n$. Опишем методы оценивания ковариационной и корреляционной матриц вектора X и укажем пороговые точки рассмотренных оценок.

1. Выборочные оценки Пирсона. Выборочной ковариационной матрицей $\hat{K}_X$ называют матрицу с элементами $\hat{k}_{ij}$, $1\le i,j\le r$, где

$$\hat{k}_{ij}=\frac{1}{n}\sum_{m=1}^{n}(X_{im}-\overline{X}_{i\cdot})(X_{jm}-\overline{X}_{j\cdot}),\quad \overline{X}_{j\cdot}=\frac{1}{n}\sum_{m=1}^{n}X_{jm}.$$

Выборочной корреляционной матрицей $\hat{R}_X$ называют матрицу с элементами $\hat{\rho}_{ij}$, $1 \le i, j \le r$, где $\hat{\rho}_{ij} = \hat{k}_{ij} / (s_i s_j)$, а $s_j = \sqrt{\frac{1}{n} \sum\limits_{m=1}^{n} (X_{jm} - \overline{X}_{j\cdot})^2}$ — выборочное среднеквадратическое отклонение случайной величины X_j.

Пороговая точка выборочной оценки ковариационной матрицы равна $1/n$ и стремится к нулю при большом объеме выборки n [9, с. 271]. Таким образом, классическая выборочная оценка $\hat{K}_X$ не является робастной.

2. Ранговая оценка Спирмена. Одним из способов робастного оценивания корреляционных матриц является построение парных ранговых оценок коэффициентов корреляции для каждой пары наблюдаемого вектора. Коэффициентом ранговой корреляции Спирмена ρ_{YZ} (например, [20, с. 120]) случайных величин Y и Z, построенным по наблюдениям $(Y_1, Z_1), \ldots, (Y_n, Z_n)$, называется статистика

$$\rho_{YZ} = \frac{\sum\limits_{m=1}^{n} (R_m - \overline{R})(S_m - \overline{S})}{\sqrt{\sum\limits_{m=1}^{n} (R_m - \overline{R})^2 \sum\limits_{m=1}^{n} (S_m - \overline{S})^2}}.$$

Здесь R_m — ранг элемента Y_m в выборке $Y_1, \ldots, Y_n$; S_m — ранг элемента Z_m в выборке $Z_1, \ldots, Z_n$; $\overline{R} = \frac{1}{n} \sum\limits_{m=1}^{n} R_m = \frac{n+1}{2}$; $\overline{S} = \frac{1}{n} \sum\limits_{m=1}^{n} S_m = \frac{n+1}{2}$ — средние арифметические рангов.

С использованием численного моделирования в [21] показано, что пороговая точка рангового коэффициента корреляции составляет примерно 0,2. В связи с этим оценку Спирмена можно полагать робастной.

К числу преимуществ ранговой оценки следует отнести то, что она позволяет оперировать с данными, измеренными не только в количественной, но и в порядковой шкале измерений, а также способность выявлять любые (не только линейные) монотонные зависимости между показателями.

3. MCD-оценка. Эта оценка — Minimum Covariance Determinant — предложена Руссо и Лероем в [9]. Идея построения оценки состоит в нахождении таких h (среди имеющихся n) наблюдений, которые имеют выборочную ковариационную матрицу с наименьшим определителем. MCD-оценка $\hat{K}_X$ ковариационной матрицы определяется как выборочная ковариационная матрица этих h наблюдений. Соответственно оценкой среднего будет служить выборочное среднее, построенное по h наблюде-

ниям. Значением h может быть любое целое число из промежутка $\left[\frac{n+1+r}{2}, n\right]$, но, как правило, h выбирают равным $\left[\frac{n+1+r}{2}\right]$. Выбор обусловлен тем, что пороговая точка MCD-оценки (см. [9, с. 271]), равная $\frac{n+1-h}{n}$, достигает наибольшего значения при указанном h.

Отметим, что при выборе $h=n$ MCD-оценка совпадает с выборочной оценкой Пирсона. Состоятельность MCD-оценки доказана в [22].

Алгоритм FAST-MCD быстрого вычисления MCD-оценки описан в [23] и в настоящее время реализован в пакете *MATLAB*.

4. Ортогонализованные оценки типа Гнанадесикана — Кетенринга, основанные на MAD-оценках. Подход к оцениванию матриц K_X и R_X, предложенный в [10], опирается на тождество

$$\operatorname{cov}(\xi,\eta)=\frac{\mathrm{D}(a\xi+b\eta)-\mathrm{D}(a\xi-b\eta)}{4ab}, \tag{3}$$

справедливое для любых случайных величин ξ и η с конечными вторыми моментами. Постоянные a и b в (3) можно выбирать произвольным образом. В частности, если принять $a=1/\sigma_\xi$ и $b=1/\sigma_\eta$, то

$$\rho(\xi,\eta)=\frac{\mathrm{D}(a\xi+b\eta)-\mathrm{D}(a\xi-b\eta)}{4}. \tag{4}$$

Теперь для построения робастных оценок элементов ковариационных или корреляционных матриц предлагается заменить в тождествах (3) и (4) дисперсии и среднеквадратические отклонения их робастными оценками. В частности, робастной оценкой среднеквадратического отклонения является MAD-оценка (Median Absolute Deviation about the Median), определяемая (см. [24]) для выборки $Y_1,\ldots,Y_n$ следующим образом:

$$\mathrm{MAD}(Y)=\mathrm{MAD}(Y_1,\ldots,Y_n)=\operatorname{med}\left(\left|Y-\operatorname{med}(Y)\right|\right),$$

где

$$\operatorname{med}(Y)=\begin{cases} Y_{(k+1)} & \text{при нечетном } n=2k+1,\\ \dfrac{Y_{(k)}+Y_{(k+1)}}{2} & \text{при четном } n=2k, \end{cases}$$

— выборочная медиана выборки $Y_1,\ldots,Y_n$; $\operatorname{med}\left(\left|Y-\operatorname{med}(Y)\right|\right)$ — выборочная медиана выборки $\left|Y_i-\operatorname{med}(Y)\right|$, $i=1,\ldots,n$.

Элементами оценки Гнанадесикана — Кетенринга корреляционной матрицы являются оценки соответствующих парных коэффициентов кор-

реляции. Хьюбером (см. [25, с. 202]) было отмечено, что построенные указанным образом оценки ковариационных и корреляционных матриц могут не обладать свойством положительной определенности. Отсутствие положительной определенности может привести к появлению отрицательных собственных значений. Для устранения этого изъяна в [11, с. 308] предложена процедура ортогонализации, а построенные новые оценки ковариационных матриц названы ортогонализованными оценками Гнанадесикана — Кетенринга. В настоящей работе при построении оценок использована коррекционная процедура, описанная в [11]. Указано также, что оценки Гнанадесикана — Кетенринга (OGK-оценки) [11, с. 309] сохраняют пороговую точку оценок дисперсии и среднего, использованных при оценивании коэффициентов корреляции. В [24] показано, что пороговая точка MAD-оценки равна 0,5.

5. Оценки Олива — Хокинса (Olive — Hawkins). Еще один метод робастного оценивания ковариационной матрицы предложен в [26], а затем развит в [12]. Согласно этому методу, предлагается построить два аттрактора ковариационной матрицы, первый из которых будет обладать свойством $\sqrt{n}$-состоятельности, а второй — высокой пороговой точкой. Затем из двух аттракторов будет выбран наилучший, согласно указанному ниже способу.

Итерационная процедура построения первого аттрактора следующая. В качестве стартовой точки выбирается выборочное среднее и выборочная ковариационная матрица и вычисляются расстояния Махалонобиса D_i, $i=1,\ldots,n,$ для всех наблюдений. Затем выбирается примерно $n/2$ наблюдений, имеющих наименьшие расстояния Махалонобиса, и по этим наблюдениям вычисляется выборочное среднее и выборочная ковариационная матрица, затем пересчитываются расстояния Махалонобиса D_i, $i=1,\ldots,n,$ до нового центра множества. Итерационную процедуру рекомендуется повторить 5 раз. Показано в [26], что такой аттрактор сходится к DGK-оценке, предложенной в [18]. В [27] доказана $\sqrt{n}$-состоятельность DGK-оценки, а в [18] с использованием численного моделирования показано, что эта оценка имеет примерно 10%-ную пороговую точку.

Для построения второго аттрактора, называемого Median Ball (MB), вычисляется покоординатная выборочная медиана MED(X) и выбирается $n/2$ наблюдений, ближайших к MED(X) в смысле евклидова расстояния. MB-оценки для среднего и ковариационной матрицы определяются как выборочное среднее и выборочная ковариационная матрица, вычисленные по этой выбранной половине наблюдений. Назовем «медианным шаром»

гиперсферу, содержащую половину наблюдений, ближайших к MED(X) в евклидовой метрике.

Если оценка среднего, полученная первым аттрактором, лежит вне медианного шара, то для оценивания ковариационной матрицы выбирают второй аттрактор. В другом случае из двух аттракторов выбирают тот, который дает оценку матрицы с наименьшим детерминантом. Итоговая оценка ковариационной матрицы определяется следующим образом:

$$\hat{K}_X = \frac{\operatorname{med}\left(D_1^2(m_A, K_A), \ldots, D_n^2(m_A, K_A)\right)}{\chi^2_{r,0,5}} K_A,$$

где m_A, K_A — оценки среднего и ковариационной матрицы, построенные выбранным аттрактором; $\chi^2_{r,0,5}$ — квантиль уровня 0,5 распределения хи-квадрат с r степенями свободы.

Методы сравнения качества сжатия многомерных показателей. Для того чтобы проводить сравнение различных методов, предназначенных для решения задачи сжатия многомерного вектора, требуется ввести количественный показатель, характеризующий эффективность МГК. В задаче с реальными данными с априорно неизвестной структурой связей такой показатель вряд ли можно определить. Однако моделированные данные имеют определенную заданную структуру зависимостей, и это позволяет ввести метрики качества. Так, в [14] предложено выбирать в качестве метрики усредненные квадраты отклонений собственных значений корреляционной матрицы от соответствующих скорректированных оценок собственных значений. В качестве метрики в [13] рассмотрены среднеквадратические отклонения фишеровских преобразований оценок коэффициентов корреляции от их истинных значений. Более продуктивной представляется мера, предложенная в [6]. Для каждого метода в [6] определена мера относительной ошибки прогноза как

$$e_{pred} = \frac{\lambda_1 + \ldots + \lambda_k}{\hat{\lambda}_1 + \ldots + \hat{\lambda}_k} - 1,$$

где k — число оптимальным образом выбранных главных компонент; λ_i, $i = 1, \ldots, k$ — истинные собственные числа ковариационной и корреляционной матриц; $\hat{\lambda}_i$, $i = 1, \ldots, k$ — оценки соответствующих собственных чисел.

Бесспорно то, что с использованием указанных выше метрик можно сравнивать различные модификации МГК. Однако ключевым моментом компонентного анализа является интерпретируемость полученных резуль-

татов: каждую из первых k главных компонент $f_1^*, \ldots, f_k^*$ следует трактовать как обобщенный показатель, представляющий некоторую группу исходных коррелированных показателей. Провести такую интерпретацию позволяет построенная на базе оцененной корреляционной матрицы матрица нагрузок L с элементами $l_{ij} = \rho(X_i^*, f_j^*)$. Понятно, что структура матрицы нагрузок определяется структурой корреляционной матрицы. Следовательно, если структура корреляционной матрицы R вектора X известна, матрица нагрузок также должна иметь определенную известную структуру. Так, если r-мерный вектор X состоит из m подвекторов и при этом компоненты, принадлежащие одному подвектору, коррелированы между собой и некоррелированы с компонентами других подвекторов вектора X, то идеальная матрица нагрузок для вектора X будет иметь m столбцов, состоящих из нулей и единиц, и $r–m$ столбцов, состоящих только из нулей. Каждый из m первых столбцов будет соответствовать главной компоненте, объединяющей компоненты одного подвектора. Единичные значения в этих столбцах будут располагаться в строках, соответствующих коррелированным компонентам данного подвектора. Остальные элементы столбца будут нулевыми, так как компоненты вектора X, принадлежащие разным подвекторам, независимы. Удалим из матрицы нагрузок $r–m$ нулевых столбцов. Назовем такую «идеальную» матрицу L_e размером $r \times m$ эталонной. Будем полагать, что из двух рассматриваемых методов главных компонент более эффективен метод, доставляющий нагрузочную матрицу, первые m столбцов которой наиболее близки (в смысле среднеквадратического отклонения) к столбцам эталонной матрицы.

Приведем формальное определение предлагаемого показателя качества сжатия векторов, обладающих структурой корреляционных связей, описанных выше. Пусть одним из указанных выше способов выбрано оптимальное число k главных компонент. Удалим из оцененной матрицы нагрузок столбцы с номерами $k+1, \ldots, r$ и обозначим эту матрицу $\tilde{L}$. Через d_{ij} обозначим евклидово расстояние между абсолютными значениями i-го столбца $\tilde{L}_i$ оцененной матрицы $\tilde{L}$ и j-м столбцом L_{ej} эталонной матрицы. Теперь переставим столбцы матрицы $\tilde{L}$ так, чтобы для каждого столбца с номером $i = 1, \ldots, k$ выполнялось равенство $d_{ii} = \min\limits_{1 \le j \le m} d_{ij}$. Показателем, измеряющим эффективность метода определения главных компонент, назовем величину $\gamma = \sum\limits_{i=1}^{m} d_{ii}$.

В случае полного совпадения оцененной матрицы $\tilde{L}$ и эталонной матрицы введенная величина γ будет равна нулю. Таким образом, при сравнении нескольких модификаций МГК будем полагать лучшей ту, которая дает наименьшее значение показателя γ. Идея такой меры предложена в [3] в задаче факторного анализа.

Далее для проведения сравнительного анализа будут смоделированы девятимерные векторы, состоящие из трех трехмерных подвекторов так, чтобы компоненты разных подвекторов были не коррелированы, а компоненты внутри каждого подвектора — сильно коррелированы между собой с парными коэффициентами корреляции 0,75–0,9. Для демонстрационных данных такой структуры эталонная матрица будет иметь следующий вид:

$$L_e = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}^{\mathrm{T}}.$$

Численный сравнительный анализ. Для проведения сравнительного анализа качества сжатия данных с помощью традиционного и робастных МГК смоделируем выборки объемом $n = 100$ девятимерных векторов с указанной выше структурой зависимостей. Векторы $X = (X_1, \ldots, X_9)^{\mathrm{T}}$ будут иметь одно из следующих распределений.

1. Многомерное нормальное распределение $N(0, K_X)$ с нулевым средним и ковариационной матрицей K_X размера 9×9 вида $\begin{pmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{pmatrix}$, где A, B и C — симметричные матрицы, имеющие вид

$$A = \begin{pmatrix} 1 & -0,9 & 0,8 \\ -0,9 & 1 & -0,75 \\ 0,8 & -0,75 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -0,9 & -0,8 \\ -0,9 & 1 & 0,75 \\ -0,8 & 0,75 & 1 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0,9 & -0,8 \\ 0,9 & 1 & 0,75 \\ -0,8 & 0,75 & 1 \end{pmatrix}.$$

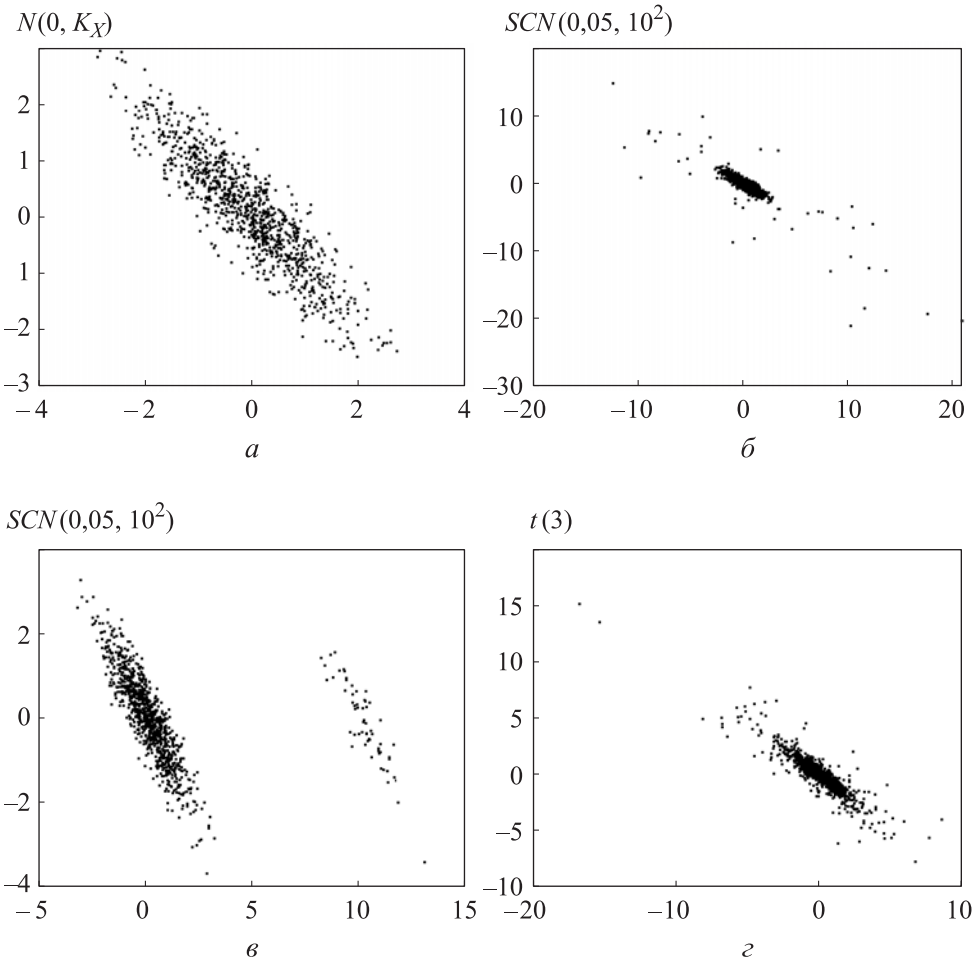
2. Многомерное распределение Тьюки $SCN(\delta, c^2)$ вида $(1-\delta)N(0, K_X) + \delta N(0, c^2 K_X)$ с долей засорения $0 \le \delta \le 1$ и параметром засорения c^2.

3. Распределение $ACN(\delta, \mu)$ вида $(1-\delta)N(0, K_X) + \delta N(\mu, K_X)$ с долей засорения $0 \le \delta \le 1$ и вектором средних μ.

4. Многомерное распределение Стьюдента $t(3)$ с тремя степенями свободы.

Распределение Тьюки $SCN(\delta, c^2)$ имитирует симметричное загрязнение нормального распределения, при котором с вероятностью δ в выборке появляются наблюдения, дисперсия которых в c^2 раз превышает дисперсию основной части наблюдений; распределение $ACN(\delta, \mu)$ имитирует асимметричное засорение кластером выбросов с математическим ожиданием μ; распределение Стьюдента имеет тяжелые хвосты.

Диаграммы рассеяния первых двух компонент вектора X для распределений $N(0, K_X)$, $SCN(0,05, 10^2)$, $ACN(0,05, \mu)$ с $\mu = (10, 0, \ldots, 0)$ и $t(3)$ приведены на рисунке.



Диаграммы рассеяния первых двух компонент вектора X:
$а$ — $N(0, K_X)$; $б$ — $SCN(0, 05, 10^2)$; $в$ — $ACN(0, 05, \mu)$ с $\mu = (10, 0, \ldots, 0)$; $г$ — $t(3)$

Для оценивания качества сжатия демонстрационных данных проведем серию из 1000 моделирований девятимерных векторов указанной структуры объемом $n=100$ для каждого распределения 1–4. Классический МГК (Pearson), использующий оценку Пирсона для корреляционной матрицы, и рассматриваемые робастные модификации МГК, использующие оценки Спирмена, MCD-оценки, ортогонализованные оценки типа Гнанадесикана — Кетенринга (OGK), основанные на MAD, и оценки Олива — Хокинса, будут сравниваться с помощью усредненного по 1000 повторов показателя

$$\overline{\gamma}=\frac{1}{1000}\sum_{i=1}^{1000}\gamma^{(i)},$$

где $\gamma^{(i)}$ — значение величины γ в i-м моделировании.

Усредненные значения $\overline{\gamma}$ отклонения оцененной матрицы $\tilde{L}$ от эталонной матрицы для выборок объемом $n=100$ при различных вероятностных распределениях наблюдаемых случайных векторов приведены в табл. 1.

Таблица 1

Усредненные значения $\overline{\gamma}$ отклонения оцененной матрицы $\tilde{L}$ от эталонной матрицы для выборок объемом $n=100$ при различных вероятностных распределениях наблюдаемых случайных векторов

Метод	Распределение				
	$N(0,K_X)$	$SCN(0,1,3^2)$	$SCN(0,05,10^2)$	$ACN(0,05,\mu)$, $\mu=(10,0,\ldots,0)$	$t(3)$
Пирсона	0,358	0,442	0,451	0,542	0,554
Спирмена	0,416	0,438	0,437	0,483	0,487
MCD	0,427	0,420	0,420	0,461	0,455
OGK	0,394	0,396	0,398	0,430	0,431
Олива — Хокинса	0,398	0,396	0,396	0,445	0,444

На основании результатов, представленных в табл. 1, можно сделать вывод о том, что в случаях, когда наблюдения имеют нормальное распределение, наилучшим является классический МГК. Таким образом, численный эксперимент подтверждает известный аналитический результат (см. [16])

о том, что оценки собственных значений ковариационной матрицы, получаемые классическим методом, являются асимптотически эффективными при нормальном распределении. Среди робастных модификаций лучше других в нормальном случае выглядит модификация OGK, использующая медианные оценки среднеквадратических отклонений. Это обстоятельство может быть объяснено тем, что MCD-оценки и оценки Олива — Хокинса фактически отсекают «худшую» половину наблюдений и строят оценки по выборке объемом $n/2$. Поэтому для выборок не слишком большого объема $n=100$ оценки Гнанадесикана — Кетенринга, использующие при оценивании выборочные медианы, построенные по всей выборке объемом n, имеют некоторое незначительное преимущество перед MCD-оценками и оценками Олива — Хокинса. Ниже будет показано, что при увеличении объема выборки на порядок оценка Гнанадесикана — Кетенринга потеряет это незначительное преимущество. В случае когда в нормальной выборке присутствует 5 % засорений с дисперсиями в 9 или 100 раз, превышающими дисперсию основной части выборки, качество классического МГК снижается. Лучшее качество здесь показывает МГК с модификацией оценки Олива — Хокинса, а модификация оценки Гнанадесикана — Кетенринга уступает ей незначительно. При 5%-ном засорении данных кластером выбросов, где среднее значение первой компоненты вектора сдвигается на 10, и для данных, имеющих распределение Стьюдента с тремя степенями свободы, наблюдается схожая картина — качество сжатия классического МГК и модификации МГК, основанной на ранговом коэффициенте Спирмена, ухудшается, лучшей оказывается OGK-модификация, модификации MCD и Олива — Хокинса незначительно уступают OGK. Отметим, что для наблюдений, имеющих распределение с тяжелыми хвостами $t(3)$, и при засорении выборки кластером выбросов качество сжатия ухудшается у всех рассматриваемых методов.

Рассмотрим еще один тип засорения данных, который имитирует ошибку постановки десятичной запятой при вводе данных. Для этого выберем случайным образом 5 % наблюдений из указанного выше распределения $N(0, K_X)$ и умножим их на $10\,m$, $m=1,2$.

Усредненные значения $\overline{\gamma}$ отклонения оцененной матрицы $\tilde{L}$ от эталонной матрицы для нормальных величин с неверно поставленной десятичной запятой приведены в табл. 2.

Результаты, приведенные в табл. 2, также свидетельствуют о чувствительности классического МГК и модификации, основанной на ранговом коэффициенте Спирмена, к такому типу аномальных данных. Методы, ос-

нованные на MCD-оценках, оценках Гнанадесикана — Кетенринга и Олива — Хокинса, абсолютно нечувствительны к такому типу засорения, поскольку при построении MCD-оценок и оценок Олива — Хокинса аномальные наблюдения отбраковываются, а при построении оценок Гнанадесикана — Кетенринга 5 % аномальных наблюдений, попадающих на хвосты, не изменяют выборочные медианы.

Таблица 2

Усредненные значения $\bar{\gamma}$ отклонения оцененной матрицы $\tilde{L}$ от эталонной матрицы для нормальных величин с неверно поставленной десятичной запятой

Метод	Распределение $N(0, K_X)$ с 5 % данных, умноженных	
	на 10	на 100
Пирсона	0,849	1,306
Спирмена	0,490	0,531
MCD	0,418	0,418
OGK	0,397	0,397
Олива — Хокинса	0,398	0,397

Рассмотрим вопрос о тенденциях изменения качества сжатия данных при увеличении объема выборки. Смоделируем $n = 1000$ наблюдений для девятимерных векторов, имеющих те же распределения, которые указаны в табл. 1, и повторим численный эксперимент по оцениванию матрицы нагрузок 1000 раз.

Усредненные значения $\bar{\gamma}$ функционала качества для выборок объемом $n = 1000$ при различных вероятностных распределениях наблюдаемых случайных векторов приведены в табл. 3.

Таблица 3

Усредненные значения $\bar{\gamma}$ функционала качества для выборок объемом n = 1000 при различных вероятностных распределениях наблюдаемых случайных векторов

Метод	Распределение				
	$N(0, K_X)$	$SCN(0,1,3^2)$	$SCN(0,05,10^2)$	$ACN(0,05,\mu)$, $\mu = (10,0,\ldots,0)$	$t(3)$
Пирсона	0,240	0,272	0,392	0,940	0,353
Спирмена	0,283	0,319	0,338	0,335	0,345
MCD	0,245	0,246	0,245	0,243	0,264
OGK	0,249	0,249	0,249	0,248	0,261
Олива — Хокинса	0,244	0,245	0,244	0,242	0,262

Все методы показывают уменьшение значения усредненного функционала качества $\overline{\gamma}$ при увеличении объема выборки. Преимущество классического МГК перед робастными методами MCD, OGK и Олива — Хокинса в нормальном случае уже не такое явное, как в случаях с выборками объемом $n=100$. Робастные модификации MCD, OGK и Олива — Хокинса показывают более высокое качество по сравнению с классическим МГК как при симметричных засорениях, описываемых распределениями Тьюки $SCN(0,05,3^2)$ и $SCN(0,05,10^2)$, так и при асимметричных засорениях. Усредненное значение функционала качества $\overline{\gamma}$ для MCD, OGK и Олива — Хокинса практически не меняется при рассмотренных засорениях данных. Однако можно отметить, что при увеличении объема данных незначительное преимущество на распределениях $N(0,K_X)$ и $ACN(0,05,\mu)$ с $\mu=(10,0,\ldots,0)$ перешло от модификации OGK к модификации Олива — Хокинса.

Заключение. Рассмотрены робастные модификации метода главных компонент, основанные на следующих робастных оценках корреляционных матриц — ранговых коэффициентах корреляции Спирмена, MCD-оценках, ортогонализованных оценках типа Гнанадесикана — Кетенринга (OGK-оценках) и оценках Олива — Хокинса. Для численного сравнения классического МГК и его робастных модификаций введено понятие эталонной матрицы и определен функционал, измеряющий эффективность метода выявления главных компонент. С помощью компьютерного моделирования вычислены значения указанного функционала, определяющего качество сжатия девятимерных векторов с коррелированными компонентами для рассмотренных модификаций МГК. Результаты численного моделирования подтвердили положение о том, что при нормальном распределении наблюдений наилучшим способом сжатия является классический МГК. Если данные имеют распределение Стьюдента с тремя степенями свободы, а также при наличии в данных симметричных засорений, описываемых распределением Тьюки, кластера выбросов или отдельных аномальных наблюдений, качество сжатия классического МГК и ранговой модификации Спирмена снижается. Робастные модификации довольно устойчивы к таким типам выбросов, лучшее качество (в смысле наименьшего значения указанного функционала) показывают модификации OGK и Олива — Хокинса. Причем OGK имеет некоторое незначительное преимущество на выборках умеренного объема ($n=100$), а Олива — Хокинса — на выборках большого объема.

ЛИТЕРАТУРА

[1] Hubert M., Engelen S. Robust PCA and classification in biosciences. *Bioinformatics*, 2004, vol. 20, iss. 11, pp. 1728–1736. DOI: https://doi.org/10.1093/bioinformatics/bth158

[2] Hubert M., Rousseeuw P.J., Branden K.V. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 2005, vol. 47, iss. 1, pp. 64–79.
DOI: https://doi.org/10.1198/004017004000000563

[3] Горяинова Е.Р., Шалимова Ю.А. Снижение размерности многомерных показателей с нелинейно зависимыми компонентами. *Бизнес-информатика*, 2015, № 3, с. 24–33.

[4] Wright J., Peng Y., Ma Y., et al. Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. *22nd NIPS*. ACM, 2009, pp. 2080–2088.

[5] Wilcox R.R. Robust principal components: a generalized variance perspective. *Behav. Res.*, 2008, vol. 40, no. 1, pp. 102–108.
DOI: https://doi.org/10.3758/BRM.40.1.102

[6] Maronna R. Principal components and orthogonal regression based on robust scales. *Technometrics*, 2005, vol. 47, no. 3, pp. 264–273.

[7] Croux C., Haesbroeck G. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 2000, vol. 87, iss. 3, pp. 603–618. DOI: https://doi.org/10.1093/biomet/87.3.603

[8] Spearman C. The proof and measurement of association between two things. *Am. J. Psych.*, 1904, vol. 15, no. 1, pp. 72–101. DOI: https://doi.org/10.2307/1412159

[9] Rousseeuw P.J., Leroy A.M. Robust regression and outlier detection. Wiley, 1987.

[10] Gnanadesikan R., Kettenring J.R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 1972, vol. 28, no. 1, Special Multivariate Issue, pp. 81–124. DOI: https://doi.org/10.2307/2528963

[11] Maronna R., Zamar R.H. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 2002, vol. 44, iss. 4, pp. 307–317.
DOI: https://doi.org/10.1198/004017002188618509

[12] Olive D.J. Robust multivariate analysis. Cham, Springer, 2017.
DOI: https://doi.org/10.1007/978-3-319-68253-2

[13] Zhang J., Olive D.J., Ye P. Robust covariance matrix estimation with canonical correlation analysis. *Int. J. Stat. Probab.*, 2012, vol. 1, no. 2, pp. 119–136.
DOI: https://doi.org/10.5539/ijsp.v1n2p119

[14] Croux C., Garcia-Escudero L.A., Gordaliza A., et al. Robust principal component analysis based on trimming around affine subspaces. *Stat. Sin.*, 2017, vol. 27, no. 3, pp. 1437–1459.

[15] Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. М., ЛКИ, 2010.

[16] Айвазян С.А., ред. Прикладная статистика. Классификация и снижение размерности. М., Финансы и статистика, 1989.

[17] Jolliffe I.T. Principal component analysis. *Springer Series in Statistics*. New York, Springer-Verlag, 2002. DOI: https://doi.org/10.1007/b98835

[18] Delvin S.J., Gnanadesikan R., Kettenring J.R. Robust estimation of dispersion matrices and principal components. *J. Am. Stat. Assoc.*, 1981, vol. 76, no. 374, pp. 354–362.

[19] Поляк Б.Т., Хлебников М.В. Метод главных компонент: робастные версии. *Автомат. и телемех.*, 2017, № 3, с. 130–148.

[20] Горяинова Е.Р., Панков А.Р., Платонов Е.Н. Прикладные методы анализа статистических данных. М., НИУ ВШЭ, 2012.

[21] Abdullah M.B. On a robust correlation coefficient. *J. R. Stat. Soc. Ser. D*, 1990, vol. 39, no. 4, pp. 455–460. DOI: https://doi.org/10.2307/2349088

[22] Cator E.A., Lopuhaä H.P. Asymptotic expansion of the minimum covariance determinant estimators. *J. Multivar. Anal.*, 2010, vol. 101, iss. 10, pp. 2372–2388. DOI: https://doi.org/10.1016/j.jmva.2010.06.009

[23] Rousseeuw P.J., van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 1999, vol. 41, iss. 3, pp. 212–223. DOI: https://doi.org/10.2307/1270566

[24] Maronna R.A., Martin D., Yohai V. Robust statistics theory and methods. Wiley, 2006.

[25] Хьюбер П.Дж. Робастность в статистике. М., Мир, 1984.

[26] Olive D.J. A resistant estimator of multivariate location and dispersion. *Comput. Stat. Data Anal.*, 2004, vol. 46, iss. 1, pp. 93–102. DOI: https://doi.org/10.1016/S0167-9473(03)00119-1

[27] Lopuhaä H.P. Asymptotics of reweighted estimators of multivariate location and scatter. *Ann. Stat.*, 1999, vol. 27, iss. 5, pp. 1638–1665. DOI: https://doi.org/10.1214/aos/1017939145

Горяинов Владимир Борисович — д-р физ.-мат. наук, доцент, профессор кафедры «Математическое моделирование» МГТУ им. Н.Э. Баумана (Российская Федерация, 105005, Москва, 2-я Бауманская ул., д. 5, корп. 1).

Горяинова Елена Рудольфовна — канд. физ.-мат. наук, доцент департамента математики факультета экономических наук НИУ ВШЭ (Российская Федерация, 101000, Москва, Мясницкая ул., 20).

Просьба ссылаться на эту статью следующим образом:

Горяинов В.Б., Горяинова Е.Р. Сравнительный анализ качества робастных модификаций метода главных компонент при сжатии коррелированных данных. *Вестник МГТУ им. Н.Э. Баумана. Сер. Естественные науки*, 2021, № 3 (96), с. 23–45. DOI: https://doi.org/10.18698/1812-3368-2021-3-23-45

COMPARATIVE ANALYSIS OF ROBUST MODIFICATION QUALITY FOR PRINCIPAL COMPONENT ANALYSIS TO PERFORM CORRELATED DATA COMPRESSION

V.B. Goryainov[1] vb-goryainov@bmstu.ru
E.R. Goryainova[2] el-goryainova@mail.ru

[1] **Bauman Moscow State Technical University, Moscow, Russian Federation**
[2] **National Research University Higher School of Economics, Moscow, Russian Federation**

Abstract

Principal component analysis is one of the methods traditionally used to solve the problem of reducing the dimensionality of a multidimensional vector with correlated components. We constructed the principal components using a special representation of the covariance or correlation matrix of the indicators observed. The classical principal component analysis uses Pearson sample correlation coefficients as estimates of the correlation matrix elements. These estimates are extremely sensitive to sample contamination and anomalous observations. To robustify the principal component analysis, we propose to replace the sample estimates of correlation matrices with well-known robust analogues, which include Spearman's rank correlation coefficient, Minimum Covariance Determinant estimates, orthogonalized Gnanadesikan — Kettenring estimates, and Olive — Hawkins estimates. The study aims to carry out a comparative numerical analysis of the classical principal component analysis and its robust modifications. For this purpose, we simulated nine-dimensional vectors with known correlation matrix structures and introduced a special metric that allows us to evaluate the quality of data compression. Our extensive numerical experiment has shown that the classical principal component analysis boasts the best compression quality for a Gaussian distribution of observations. When observations are characterised by a Student's *t*-distribution with three degrees of freedom, as well as when a cluster of outliers, individual anomalous observations, or symmetric contaminations described by the Tukey distribution are present in the data, it is the Gnanadesikan — Kettenring and Olive — Hawkins estimates modifying the principal component analysis that show the best compression quality. The quality of the classical principal component analysis and Spearman's rank modification decreases in these cases

Keywords

Robust principal component analysis, MCD estimate, Gnanadesikan — Kettenring estimate, Olive — Hawkins estimate

Received 16.07.2020
Accepted 29.01.2021

REFERENCES

[1] Hubert M., Engelen S. Robust PCA and classification in biosciences. *Bioinformatics*, 2004, vol. 20, iss. 11, pp. 1728–1736. DOI: https://doi.org/10.1093/bioinformatics/bth158

[2] Hubert M., Rousseeuw P.J., Branden K.V. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 2005, vol. 47, iss. 1, pp. 64–79. DOI: https://doi.org/10.1198/004017004000000563

[3] Goryainova E.R., Shalimova Yu.A. Reducing the dimensionality of multivariate indicators containing non-linearly dependent components. *Business Informatics*, 2015, no. 3, pp. 24–33 (in Russ.).

[4] Wright J., Peng Y., Ma Y., et al. Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. *22nd NIPS*. ACM, 2009, pp. 2080–2088.

[5] Wilcox R.R. Robust principal components: a generalized variance perspective. *Behav. Res*., 2008, vol. 40, no. 1, pp. 102–108. DOI: https://doi.org/10.3758/BRM.40.1.102

[6] Maronna R. Principal components and orthogonal regression based on robust scales. *Technometrics*, 2005, vol. 47, no. 3, pp. 264–273.

[7] Croux C., Haesbroeck G. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 2000, vol. 87, iss. 3, pp. 603–618. DOI: https://doi.org/10.1093/biomet/87.3.603

[8] Spearman C. The proof and measurement of association between two things. *Am. J. Psych*., 1904, vol. 15, no. 1, pp. 72–101. DOI: https://doi.org/10.2307/1412159

[9] Rousseeuw P.J., Leroy A.M. Robust regression and outlier detection. Wiley, 1987.

[10] Gnanadesikan R., Kettenring J.R. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 1972, vol. 28, no. 1, Special Multivariate Issue, pp. 81–124. DOI: https://doi.org/10.2307/2528963

[11] Maronna R., Zamar R.H. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 2002, vol. 44, iss. 4, pp. 307–317. DOI: https://doi.org/10.1198/004017002188618509

[12] Olive D.J. Robust multivariate analysis. Cham, Springer, 2017. DOI: https://doi.org/10.1007/978-3-319-68253-2

[13] Zhang J., Olive D.J., Ye P. Robust covariance matrix estimation with canonical correlation analysis. *Int. J. Stat. Probab*., 2012, vol. 1, no. 2, pp. 119–136. DOI: https://doi.org/10.5539/ijsp.v1n2p119

[14] Croux C., Garcia-Escudero L.A., Gordaliza A., et al. Robust principal component analysis based on trimming around affine subspaces. *Stat. Sin*., 2017, vol. 27, no. 3, pp. 1437–1459.

[15] Ivchenko G.I., Medvedev Yu.I. Vvedenie v matematicheskuyu statistiku [Introduction to mathematical statistics]. Moscow, LKI Publ., 2010.

[16] Aivazyan S.A., ed. Prikladnaya statistika. Klassifikatsia i snizheniye razmernosti [Applied statistics. Classification and dimension reduction]. Moscow, Finansy i statistika Publ., 1989.

[17] Jolliffe I.T. Principal component analysis. *Springer Series in Statistics*. New York, Springer-Verlag, 2002. DOI: https://doi.org/10.1007/b98835

[18] Delvin S.J., Gnanadesikan R., Kettenring J.R. Robust estimation of dispersion matrices and principal components. *J. Am. Stat. Assoc*., 1981, vol. 76, no. 374, pp. 354–362.

[19] Polyak B.T., Khlebnikov M.V. Principle component analysis: robust versions. *Autom. Remote Control*, 2017, vol. 78, no. 3, pp. 490–506.
DOI: https://doi.org/10.1134/S0005117917030092

[20] Goryainova E.R., Pankov A.P., Platonov E.N. Prikladnye metody analiza statisticheskikh dannykh [Applied methods of statistical data analysis]. Moscow, HSE Univ. Publ., 2012.

[21] Abdullah M.B. On a robust correlation coefficient. *J. R. Stat. Soc. Ser. D*, 1990, vol. 39, no. 4, pp. 455–460. DOI: https://doi.org/10.2307/2349088

[22] Cator E.A., Lopuhaä H.P. Asymptotic expansion of the minimum covariance determinant estimators. *J. Multivar. Anal*., 2010, vol. 101, iss. 10, pp. 2372–2388. DOI: https://doi.org/10.1016/j.jmva.2010.06.009

[23] Rousseeuw P.J., van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 1999, vol. 41, iss. 3, pp. 212–223.
DOI: https://doi.org/10.2307/1270566

[24] Maronna R.A., Martin D., Yohai V. Robust statistics theory and methods. Wiley, 2006.

[25] Huber P.J. Robust statistics. Wiley, 1981.

[26] Olive D.J. A resistant estimator of multivariate location and dispersion. *Comput. Stat. Data Anal*., 2004, vol. 46, iss. 1, pp. 93–102.
DOI: https://doi.org/10.1016/S0167-9473(03)00119-1

[27] Lopuhaä H.P. Asymptotics of reweighted estimators of multivariate location and scatter. *Ann. Stat*., 1999, vol. 27, iss. 5, pp. 1638–1665.
DOI: https://doi.org/10.1214/aos/1017939145

Goryainov V.B. — Dr. Sc. (Phys.-Math.), Assoc. Professor, Professor, Department of Mathematical Simulation, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5/1, Moscow, 105005 Russian Federation).

Goryainova E.R. — Cand. Sc. (Phys.-Math.), Assoc. Professor, Department of Mathematics, Faculty of Economic Sciences, National Research University Higher School of Economics (Myasnitskaya ul. 20, Moscow, 101000 Russian Federation).

Please cite this article in English as:

Goryainov V.B., Goryainova E.R. Comparative analysis of robust modification quality for principal component analysis to perform correlated data compression. *Herald of the Bauman Moscow State Technical University, Series Natural Sciences*, 2021, no. 3 (96), pp. 23–45 (in Russ.). DOI: https://doi.org/10.18698/1812-3368-2021-3-23-45