УДК 519.234.3 DOI: 10.18698/1812-3368-2023-2-17-34

ИССЛЕДОВАНИЕ УСТОЙЧИВОСТИ К АНОМАЛЬНЫМ НАБЛЮДЕНИЯМ МОДИФИКАЦИЙ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ

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

Рассмотрена задача редукции многомерных коррелированных показателей. Один из подходов к решению этой задачи основан на методе главных компонент, который позволяет компактно описать вектор с коррелированными координатами (компонентами) с помощью вектора главных компонент с некоррелированными координатами существенно меньшей размерности, сохраняя при этом большую часть информации о корреляционной структуре исходного вектора. На моделированных и реальных данных проведено сравнение несколько модификаций метода главных компонент, отличающихся способом оценивания корреляционной матрицы вектора наблюдений. Цель работы — демонстрация преимущества робастных модификаций метода главных компонент в тех случаях, когда данные содержат аномальные значения. Для сравнения рассматриваемых модификаций на модельных данных введена метрика, измеряющая различие оцененных и истинных собственных значений корреляционной матрицы исходных данных. Методом компьютерного моделирования исследовано поведение этой меры в зависимости от вероятностного распределения наблюдений. В качестве распределений выбраны имитирующие засоренную выборку многомерные распределения с недиагональными корреляционными матрицами. Далее изучена выборка 13 коррелированных социально-экономических показателей по 85 странам, в которой выявлено 46 аномальных значений. Рассмотренные модификации метода главных компонент выбрали одина-

Ключевые слова

Метод главных компонент, робастные оценки корреляционной матрицы, MCD-оценка, оценка Гнанадесикана — Кетенринга, оценка Олива — Хокинса, распределение Тьюки, бимодальное распределение

ковое оптимальное число главных компонент, равное трем. Однако качество сжатия реальных данных, которое определяется как доля суммарной дисперсии исходных показателей, описываемая первыми тремя главными компонентами, оказалась существенно выше у робастных модификаций метода главных компонент. Полученные на реальных данных результаты хорошо согласуются с выводами компьютерного моделирования

Поступила 06.08.2022
Принята 14.10.2022

Введение. Метод главных компонент (МГК) — статистический метод, позволяющий по наблюдениям случайного вектора $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ с коррелированными компонентами получить его компактное описание, линейно преобразуя X в вектор главных компонент $Z = (Z_1, \ldots, Z_p)^{\mathrm{T}}$ с некоррелированными координатами. При этом большую часть информации о корреляционной структуре исходного вектора X можно описать подвектором $(Z_1, \ldots, Z_q)^{\mathrm{T}}$ вектора Z существенно меньшей размерности q, $q < p$ [1, 2]. Метод главных компонент с успехом используется во многих областях науки и техники. Большое количество примеров из физики, химии, биологии, генетики, сельского хозяйства, геологии, экологии, климатологии, метеорологии, океанографии, демографии, экономики, психологии и других наук приведено, например, в [2]. Нередко МГК предшествует другим многомерным методам, например, регрессионному, дискриминантному и кластерному анализу.

Построение главных компонент основано на специальном представлении корреляционной матрицы исходных данных. Как правило, на практике корреляционная матрица неизвестна и ее необходимо оценивать. В классическом МГК в качестве оценки используется выборочная корреляционная матрица Пирсона, которая является эффективной оценкой при нормальном распределении наблюдаемого вектора X [1]. Однако эта оценка чувствительна к отклонениям распределения исходных данных от нормального, в частности, к выбросам (большим ошибкам в наблюдениях) [3], что приводит МГК к грубым ошибкам в итоговых результатах.

В настоящее время существуют методы оценивания корреляционных матриц [4], которые достаточно хорошо справляются с отклонениями распределения исходных данных от нормального, в том числе с выбросами в наблюдениях. Кроме того, эти методы доступны для рядового пользователя, поскольку имеются в различных компьютерных средах, например, в MATLAB. Поэтому логично использовать подход к нахождению главных

компонент, который состоит в замене выборочной корреляционной матрицы Пирсона ее робастными аналогами. Однако для этого необходимо осознавать преимущества и недостатки для МГК каждого метода оценивания корреляционной матрицы исходных данных.

Цель работы — демонстрация преимущества робастных модификаций МГК перед классическим МГК в случаях, когда данные имеют распределение, отличающееся от гауссова. Для этого вводится специальная метрика, характеризующая качество сжатия данных, а затем с помощью методов компьютерного моделирования проводится сравнительный анализ различных модификаций МГК на данных, имеющих многомерное нормальное распределение, распределение Тьюки, распределение Стьюдента, бимодальное (двугорбое) нормальное распределение.

Приведен пример сжатия 13-мерного вектора социально-экономических показателей по 85 странам. Полученные на этих реальных данных результаты хорошо согласуются с выводами компьютерного моделирования.

Метод главных компонент. Рассмотрим случайный вектор наблюдений (исходных данных) $X = (X_1, \ldots, X_p)^{\mathrm{T}}$ с математическим ожиданием $\mathrm{E}X = 0$ и корреляционной матрицей Σ. Случайные величины $X_1, \ldots, X_p$ также называются показателями, или признаками. Обозначим через $\lambda_1, \ldots, \lambda_p$ и $e_1, \ldots, e_p$ упорядоченные в порядке убывания собственные значения и соответствующие им нормированные собственные векторы матрицы Σ.

Рассмотрим вектор $Y = (Y_1, \ldots, Y_q)^{\mathrm{T}}$, $q < p$, являющийся ортогональным линейным преобразованием $Y = B^{\mathrm{T}}X$ исходных данных X, где B — произвольная ортогональная матрица размером $p \times q$, т. е. $B^{\mathrm{T}}B = I$, I — единичная матрица.

Обозначим через $\mathcal{L}_x$ и $\mathcal{L}_y$ подпространства в $\mathbb{R}^p$, порожденные линейными комбинациями случайных величин $X_1, \ldots, X_p$ и $Y_1, \ldots, Y_q$ соответственно. Ортогональная проекция $\hat{X}$ вектора X на подпространство $\mathcal{L}_y$ имеет вид $\hat{X} = BB^{\mathrm{T}}X$. Метод главных компонент утверждает, что математическое ожидание евклидовой нормы $\mathrm{E}\,|X - \hat{X}|^2$ минимально, когда матрица B совпадает с матрицей A_q, столбцами которой являются координаты векторов $e_1, \ldots, e_q$. Другими словами, произвольные q случайных величин из $\mathcal{L}_x$ содержат наибольшую информацию о корреляционной структуре X, если они совпадают с первыми q координатами $Z_1, \ldots, Z_q$ случайного вектора $Z = A_p^{\mathrm{T}}X$.

Координаты $Z_1, \ldots, Z_p$ вектора Z называют главными компонентами исходного вектора X, их дисперсии $\mathrm{D}Z_j$, $j = 1, \ldots, p$, удовлетворяют условию $\mathrm{D}Z_1 \geq \mathrm{D}Z_2 \geq \ldots \geq \mathrm{D}Z_p$, а ковариационная матрица вектора Z, равная $A_p^{\mathrm{T}} \Sigma A_p$, является диагональной с элементами $\lambda_1, \ldots, \lambda_p$ на диагонали.

Оценивание корреляционных матриц. В приложениях при отыскании главных компонент неизвестная корреляционная матрица Σ вектора X заменяется ее оценкой, построенной по n наблюдениям $(x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, вектора X. В работе будут сравниваться модификации МГК, основанные на следующих пяти оценках корреляционной матрицы Σ вектора X, подробное описание которых имеется, например, в [4, 5].

1. Оценка Пирсона. Выборочная корреляционная матрица $\hat{\Sigma}$ Пирсона, ij-й элемент которой $\hat{\sigma}_{ij}$ является выборочным коэффициентом корреляции между $(x_{1i}, \ldots, x_{ni})$ и $(x_{1j}, \ldots, x_{nj})$.

2. Оценка Спирмена. Обозначим через r_{ij} ранг x_{ij} в последовательности $x_{1j}, \ldots, x_{nj}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$. Ранговой оценкой Спирмена (см., например, [6]) корреляционной матрицы Σ называют матрицу $\hat{\Sigma}_r$, ij-й элемент которой $\hat{\sigma}_{ij}^{(r)}$ является выборочным коэффициентом корреляции между $(r_{1i}, \ldots, r_{ni})$ и $(r_{1j}, \ldots, r_{nj})$.

3. MCD-оценка. Определяется (см. [7, 8]) как выборочная корреляционная матрица $\left[\dfrac{n+1+p}{2} \right]$ наблюдений, которые имеют выборочную корреляционную матрицу с наименьшим определителем среди имеющихся n наблюдений.

4. Ортогонализованная оценка Гнанадесикана — Кетенринга (OGK). Эта оценка матрицы Σ опирается на тождества

$$\operatorname{cov}(X_i, X_j) = \frac{\mathrm{D}(X_i + X_j) - \mathrm{D}(X_i - X_j)}{4}, \qquad \rho(X_i, X_j) = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{\mathrm{D}X_i \mathrm{D}X_j}},$$

в которых среднеквадратические отклонения $\sqrt{\mathrm{D}(X_i + X_j)}$, $\sqrt{\mathrm{D}(X_i - X_j)}$, $\sqrt{\mathrm{D}X_i}$ и $\sqrt{\mathrm{D}X_j}$ оцениваются MAD-оценками [9]. При этом MAD-оценка среднеквадратического отклонения $\sqrt{\mathrm{D}\xi}$ произвольной случайной величины ξ по ее наблюдениям $u_1, \ldots, u_k$ определяется как

$$\mathrm{MAD}(u_1, \ldots, u_k) = \operatorname{med}\left(\left| u_1 - \operatorname{med}(u_1, \ldots, u_k) \right|, \ldots, \left| u_k - \operatorname{med}(u_1, \ldots, u_k) \right| \right).$$

Здесь $\mathrm{med}(u_1,\ldots,u_k)$ — выборочная медиана выборки $u_1,\ldots,u_k$. Если получившаяся оценка не является положительно определенной матрицей, то она корректируется с использованием процесса ортогонализации [10].

5. Оценка Олива — Хокинса (OH). Эта оценка пропорциональна выборочной корреляционной матрице, построенной по $n/2$ наблюдениям. Эти $n/2$ наблюдений являются либо ближайшими к медиане наблюдений $\mathrm{Med}(x)=(\mathrm{med}(x_{11},\ldots,x_{n1}),\ldots,\mathrm{med}(x_{1p},\ldots,x_{np}))$ в смысле евклидова расстояния, либо имеют от выборочного среднего наблюдений наименьшие расстояния Махалонобиса, вычисленные с использованием выборочной корреляционной матрицы $\hat{\Sigma}$. Точное определение оценки и алгоритм ее вычисления приведены в [4, 11].

Сравнительный анализ различных модификаций метода главных компонент. Для того чтобы проводить сравнение различных методов, предназначенных для решения задачи сжатия многомерного вектора, требуется ввести количественный показатель, характеризующий эффективность МГК. Это возможно для моделированных данных, поскольку они имеют определенную заданную структуру зависимостей. Например, в [12] в качестве такой метрики рассмотрены среднеквадратические отклонения фишеровских преобразований оценок коэффициентов корреляции от их истинных значений. Однако ключевой задачей МГК является не оценивание корреляционных матриц, а интерпретируемость построенных главных компонент. Интерпретируемость главных компонент определяется структурой матрицы нагрузок, столбцами которой являются нормированные собственные векторы корреляционной матрицы наблюдений. Для этого в [13] предложена мера качества МГК, равная евклидовому расстоянию между редуцированной оцененной матрицей нагрузок и редуцированной эталонной матрицей нагрузок, которая соответствует смоделированным зависимостям. Однако в [5] было отмечено, что вычисление введенной в [13] метрики оказывается очень трудоемким. В настоящей работе в качестве метрики будет рассматриваться величина вида

$$d=\max(u,1/u)-1,\qquad u=\sum_{j=1}^{q}\lambda_j\Big/\sum_{j=1}^{q}\hat{\lambda}_j,$$

где q — число оптимальным образом выбранных главных компонент; $\lambda_1,\ldots,\lambda_q$ — собственные числа корреляционной матрицы, $\hat{\lambda}_1,\ldots,\hat{\lambda}_q$ — оценки соответствующих собственных чисел.

При сравнении нескольких модификаций МГК будем полагать лучшей ту, которая дает наименьшее значение показателя d. Идея использования

меры d основана на предложенной в [14] мере относительной ошибки прогнозирования.

Для проведения сравнительного анализа моделировались шестимерные ($p = 6$) векторы, состоящие из двух трехмерных подвекторов так, чтобы координаты разных подвекторов были некоррелированны, а координаты каждого подвектора — сильно коррелированны между собой с парными коэффициентами корреляции 0,6–0,9. Для демонстрационных данных такой структуры выбрана эталонная корреляционная матрица Σ размером 6×6 вида

$$\Sigma = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \quad A = \begin{pmatrix} 1 & -0,9 & 0,8 \\ -0,9 & 1 & -0,7 \\ 0,8 & -0,7 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -0,8 & -0,7 \\ -0,8 & 1 & 0,6 \\ -0,7 & 0,6 & 1 \end{pmatrix}. \quad (1)$$

Вследствие этого все модификации МГК выделяли две ($q = 2$) главные компоненты из шести.

Между собой сравнивались классический МГК, использующий оценку Пирсона корреляционной матрицы, и робастные модификации МГК, основанные на MCD-оценке и оценке Спирмена, Гнанадесикана — Кетенринга и Олива — Хокинса. Предполагалось, что вектор наблюдений X имеет нормальное распределение, распределение Тьюки, распределение Стьюдента и бимодальное (двугорбое) нормальное распределение.

Распределение Тьюки. Обозначим через $f(x, \mu, \Sigma)$, $x \in \mathbb{R}^p$, плотность p-мерного нормального распределения с математическим ожиданием $\mu \in \mathbb{R}^p$ и корреляционной матрицей Σ. Назовем p-мерным распределением Тьюки или p-мерным загрязненным (засоренным) нормальным распределением с долей загрязнения γ и величиной загрязнения δ распределение с плотностью

$$f_T(x, \Sigma, \gamma, \delta) = (1 - \gamma) f(x, 0, \Sigma) + \gamma f(x, 0, \delta\Sigma).$$

Распределение Тьюки описывает засорение нормального распределения, при котором с вероятностью γ в выборке появляются наблюдения, среднеквадратическое отклонение которых в δ раз превышает среднеквадратическое отклонение основной части наблюдений. В частности, при $\delta = 10$ распределение Тьюки имитирует достаточно распространенную ошибку, когда при вводе данных с клавиатуры десятичная запятая ошибочно ставится на одну позицию правее, чем следовало бы. При $\delta = 1$ или $\gamma = 0$ распределение Тьюки совпадает с нормальным.

Мнения о том, какие уровни загрязнения являются типичными на практике, расходятся [3]. На основании дискуссии, ведущейся на эту тему в научной литературе, можно сделать вывод о том, что в приложениях наиболее распространенные уровни засорения описываются распределением Тьюки с $0 < \gamma < 0,15$ и $1 < \delta < 3$.

В предположении, что случайный вектор наблюдений X имеет p-мерное распределение Тьюки, в работе с использованием компьютерного моделирования для различных модификаций МГК исследована зависимость показателя эффективности d от γ и δ. Для этого $N = 10^4$ раз моделировалась выборка вектора X объемом $n = 100$, имеющего p-мерное $(p = 6)$ распределение Тьюки с корреляционной матрицей (1) и различными γ и δ, в частности с $\gamma = 0$ и $\delta = 1$, что соответствует нормальному распределению. Качество каждой из пяти перечисленных выше модификаций МГК оценивалось величиной $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$, d_i — значение d в i-м численном эксперименте, $i = 1, \ldots, N$.

Для пяти различных модификаций МГК приведены зависимости $\bar{d}(\gamma)$ при фиксированном значении $\delta = 2$ (рис. 1, *а*). Метод главных компонент, основанный на выборочной корреляционной матрице, является лучшим только при отсутствии засорений $(\gamma = 0)$ и становится наихудшим при $\delta = 2$ уже для $\gamma > 0,01$. Наилучшими следует признать методы МГК, использующие MCD-оценку, оценки Гнанадесикана — Кетенринга и Олива — Хокинса. При отсутствии засорений они практически не уступают методам, основанным на выборочной и ранговой корреляционных матрицах, а при наличии небольших засорений заметно их превосходят.

Для этих же модификаций МГК приведены зависимости $\bar{d}(\delta)$ при фиксированном значении $\gamma = 0,01$ (рис. 1, *б*). Метод главных компонент, основанный на выборочной корреляционной матрице, является худшим уже при $\delta > 2,1$. Наилучшими следует признать методы МГК, исполь-зующие MCD-оценку, оценки Гнанадесикана — Кетенринга и Олива — Хокинса.

Распределение Стьюдента. Предположим, что случайный вектор наблюдений X имеет p-мерное распределение Стьюдента [15] с m степенями свободы и матричным параметром Σ, т. е. его плотность распределения вероятностей имеет вид:

$$f(x, \Sigma, m) = \frac{1}{|\Sigma|^{1/2}} \frac{1}{\sqrt{(m\pi)^p}} \frac{\Gamma((m+p)/2)}{\Gamma(m/2)} \left(1 + \frac{x^{\mathrm{T}}\Sigma^{-1}x}{m}\right)^{-(m+p)/2}, \quad x \in \mathbb{R}^p.$$

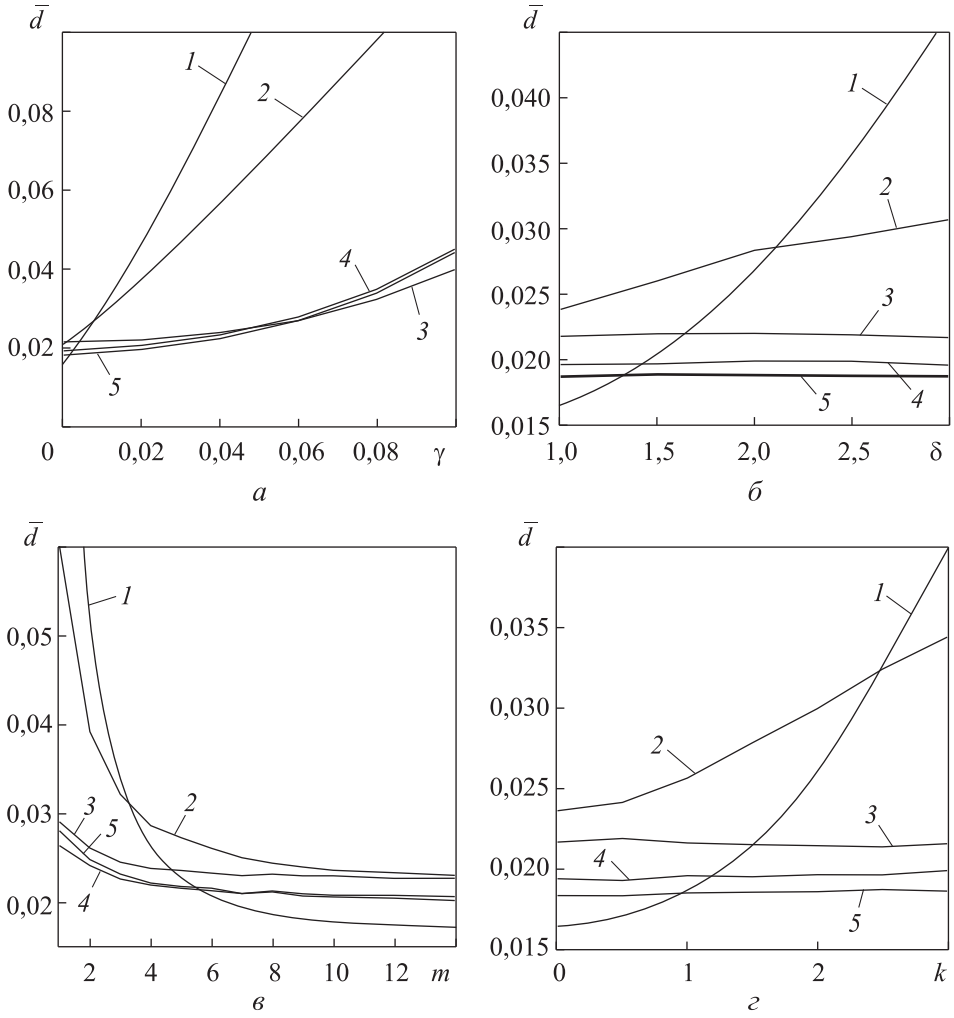

Рис. 1. Зависимости $\bar{d}(\gamma)$ (*а*), $\bar{d}(\delta)$ (*б*), $\bar{d}(m)$ (*в*) и $\bar{d}(k)$ (*г*):

1 — оценка Пирсона; *2* — оценка Спирмена; *3* — MCD-оценка; *4* — оценка Гнанадесикана — Кетенринга; *5* — оценка Олива — Хокинса

Распределение Стьюдента при $m \to \infty$ стремится к нормальному. Поэтому оно является хорошей моделью распределений, отклоняющихся от нормального, причем степень отклонения можно регулировать, изменяя параметр m от 1 до ∞. Отметим, что корреляционная матрица этого распределения существует при $m > 2$ и равна $\dfrac{m}{m-2}\Sigma$.

В работе с использованием компьютерного моделирования с числом повторов $N = 10^4$ исследована зависимость показателя d эффективности различных модификаций МГК от m в предположении, что матрица Σ имеет вид (1). Для пяти модификаций МГК и выборок объемом $n = 100$ приве-

дены зависимости $\overline{d}(m)$ (рис. 1, *в*). Метод главных компонент, основанный на выборочной корреляционной матрице, является конкурентоспособным при $m > 4$ и становится лучшим при $m > 5$. Методы МГК, использующие MCD-оценки, оценки Гнанадесикана — Кетенринга и Олива — Хокинса, являются наилучшими при $m < 5$ и конкурентоспособными при $m > 5$. Метод МГК, основанный на ранговой корреляционной матрице, наихудший при $m > 3$ и далеко не лучший при $m < 3$. Таким образом, если случайный вектор наблюдений X имеет многомерное распределение Стьюдента, то разумно всегда использовать методы МГК, основанные на MCD-оценке, оценках Гнанадесикана — Кетенринга и Олива — Хокинса.

Бимодальное нормальное распределение. Рассмотрим еще один тип засорения данных, предполагая, что случайный вектор наблюдений X имеет бимодальное (двугорбое) нормальное распределение с плотностью

$$f_b(x, \gamma, \mu, \Sigma) = (1-\gamma) f(x, 0, \Sigma) + \gamma f(x, \mu, \Sigma),$$

где $f(x, \mu, \Sigma)$, $x \in \mathbb{R}^p$ — плотность p-мерного нормального распределения с математическим ожиданием $\mu \in \mathbb{R}^p$ и корреляционной матрицей Σ. Бимодальное распределение описывает засорение нормального распределения с плотностью $f(x, 0, \Sigma)$ кластером выбросов с плотностью $f(x, \mu, \Sigma)$, μ — расстояние между центрами основного и загрязняющего распределений; γ — доля засорения. Канонической иллюстрацией к бимодальному распределению является распределение антропометрических характеристик (рост, окружность головы и т. д.) индивидов, выбранных из двух кластеров (мужчин и женщин). Показатели в обоих кластерах имеют схожую структуру связей, но различные средние значения, соответствующие горбам бимодального распределения.

В работе предположено, что $\mu = (k, 0, 0, 0, 0, 0)$ и с помощью компьютерного моделирования с числом повторов $N = 10^4$ исследована зависимость d эффективности различных модификаций МГК от k в предположении, что матрица Σ имеет вид (1).

Для пяти различных модификаций МГК и выборок объемом $n = 100$ приведены зависимости $\overline{d}(k)$ при фиксированном значении $\gamma = 0{,}01$ (рис. 1, *г*). Метод главных компонент, основанный на выборочной корреляционной матрице, является лучшим только при $k < 1$ и становится наихудшим при $k > 2{,}5$. При $k > 1{,}5$ наилучшими являются методы МГК, использующие MCD-оценку, оценки Гнанадесикана — Кетенринга и Олива — Хокинса. Метод рангового МГК не является наилучшим ни при каком значении k.

Отдельно следует отметить MCD-оценку, которая, по существу, является классической выборочной оценкой корреляционной матрицы, построенной по «лучшей» (незагрязненной) половине имеющихся наблюдений. В связи с этим может показаться, что MCD-оценка всегда должна иметь преимущество в качестве перед классической оценкой. Согласно результатам моделирования (см. рис. 1, *б–г*), в случае, когда наблюдения имеют близкое к нормальному распределение (например, распределение Стьюдента с числом степеней свободы $m > 5$, распределение Тьюки с засоряющей дисперсией $\sigma^2 < 1{,}5^2$), МГК, основанный на MCD-оценке, проигрывает классическому. Это связано с тем, что MCD-оценка использует лишь $n/2$ наблюдений и, соответственно, имеет большую дисперсию, что приводит к более медленной скорости сходимости оцениваемых параметров к истинным значениям по сравнению с классическим методом.

Пример с социально-экономическими данными. Взяты данные о 13 социально-экономических показателях по 85 странам*. Исследуемые показатели: ВВП на душу населения X_1; коэффициент младенческой смертности X_2; ожидаемая продолжительность жизни при рождении X_3; плотность врачей X_4; плотность больничных коек X_5; процент ожирения среди взрослого населения X_6; процент населения, находящегося за чертой бедности, X_7; профицит/дефицит бюджета X_8; доля сельского хозяйства в структуре ВВП X_9; доля сферы услуг в структуре ВВП X_{10}; уровень безработицы X_{11}; коэффициент индекса Джини распределения семейного дохода X_{12}; коэффициент чистой миграции X_{13}.

Оценивая корреляционную матрицу вектора $X = (X_1, \ldots, X_{13})^{\mathrm{T}}$ любым из пяти рассмотренных выше методов, можно убедиться, что каждый показатель имеет высокие коэффициенты корреляции с большинством остальных.

Согласно критерию Шапиро — Уилка [16], гипотеза о нормальности всех показателей, за исключением X_{10}, отвергнута на уровне значимости не более 0,008.

Далее выполнен поиск аномальных наблюдений. Наблюдение y_i признавалось выбросом (аномальным наблюдением) среди $y_1, \ldots, y_n$, $n = 85$, если $y_i \notin (\hat{\mu} - 3\hat{\sigma},\ \hat{\mu} + 3\hat{\sigma})$, где $\hat{\mu} = \mathrm{med}(y_1, \ldots, y_n)$, $\hat{\sigma} = \mathrm{MAD}(y_1, \ldots, y_n)/\Phi^{-1}(0{,}75)$, $\Phi^{-1}(x)$ — функция, обратная функции распределения вероятностей стандартной нормальной случайной величины. Аномальные значения хотя бы по одному из 13 показателей обнаружены

* https://www.cia.gov/the-world-factbook/countries

в 29 странах. Такие страны, как Афганистан, Джибути, Мали, Мозамбик, ЦАР имели выбросы не менее чем в трех показателях.

Сформулируем задачу сжатия наблюдаемых показателей и попытаемся дать компактное описание наблюдаемых признаков с использованием меньшего числа некоррелированных показателей (главных компонент). Для этого найдем собственные числа корреляционных матриц, вычисленных пятью методами оценивания. Определим оптимальное число главных компонент, необходимых для описания наблюдаемого вектора показателей, как число собственных значений, больших единицы. Указанные собственные числа приведены в табл. 1.

Таблица 1

Собственные числа корреляционных матриц для пяти методов оценивания

Оценка	1	2	3	4	5	6	7	8	9	10	11	12	13
Пирсона	6,30	1,54	1,20	0,82	0,69	0,65	0,49	0,36	0,26	0,24	0,21	0,14	0,05
Спирмена	7,04	1,41	1,06	0,72	0,59	0,55	0,50	0,37	0,25	0,19	0,15	0,08	0,04
MCD	6,73	1,90	1,39	0,68	0,55	0,47	0,38	0,27	0,24	0,13	0,11	0,05	0,05
Гнанадесикана — Кетенринга	6,96	1,67	1,20	0,77	0,55	0,42	0,37	0,29	0,25	0,16	0,13	0,10	0,06
Олива — Хокинса	7,12	1,82	1,17	0,72	0,52	0,40	0,38	0,30	0,19	0,12	0,11	0,06	0,04

Все методы определяют одинаковое оптимальное число главных компонент, равное трем. Для того чтобы дать качественную интерпретацию новых показателей Z_1, Z_2, Z_3, вычислим корреляционную матрицу A между векторами $X_1, \ldots, X_{13}$ и Z_1, Z_2, Z_3. Элементы a_{ij}, $i = 1, \ldots, 13$, $j = 1, 2, 3$, этой матрицы равны коэффициентам корреляции между i-м исходным показателем X_i и j-й главной компонентой Z_j. Для каждого из пяти методов оценивания матрица A приведена в табл. 2, высокие коэффициенты корреляции выделены полужирным.

На основании матрицы A во всех пяти случаях можно утверждать, что первая главная компонента представляет собой первые десять показателей, вторая — уровень безработицы и индекс Джини, третья — коэффициент чистой миграции. Таким образом, противоречий в интерпретации данных при применении различных модификаций МГК в рассмотренном примере нет. Три новых показателя условно можно описать следующим образом: первая главная компонента характеризует социально-экономическое благополучие страны; вторая — степень экономического неравенства внутри страны; третья — желание проживать в стране.

Таблица 2

Корреляционные матрицы между исходными показателями и тремя главными компонентами для пяти методов оценивания

Оценка														
Пирсона			Спирмена			MCD			Гнанадесикана — Кетенринга			Олива — Хокинса		
–0,894	–0,002	0,173	**–0,885**	–0,010	0,104	**0,886**	0,031	–0,114	**–0,883**	0,092	–0,135	**0,883**	0,111	–0,080
0,810	0,032	0,398	**0,820**	0,009	0,266	**–0,821**	–0,103	–0,301	**0,819**	–0,221	–0,286	**–0,815**	–0,180	–0,369
0,921	0,005	–0,020	**0,916**	0,007	–0,000	**–0,917**	–0,055	–0,015	**0,914**	–0,146	0,017	**–0,914**	–0,139	–0,064
0,804	–0,063	–0,207	**0,803**	–0,030	–0,323	**–0,802**	–0,090	0,371	**0 ,803**	–0,083	0,324	**–0,806**	–0,119	0,293
0,607	–0,299	–0,207	**0,615**	–0,263	–0,339	**–0,593**	–0,321	0,516	**0,594**	–0,267	0,561	**–0,596**	–0,323	0,511
0,633	0,322	–0,327	**0,619**	0,358	–0,326	**–0,643**	0,318	0,234	**0,643**	0,307	0,090	**–0,646**	0,261	0,162
–0,653	0,132	0,316	**–0,645**	0,187	0,254	**0,631**	0,324	–0,222	**–0,633**	0,313	–0,346	**0,638**	0,334	–0,180
0,660	–0,305	0,212	**0,661**	–0,302	0,096	**–0,657**	–0,372	–0,117	**0,650**	–0,417	–0,055	**–0,649**	–0,417	–0,082
–0,858	–0,339	0,080	**–0,856**	–0,319	0,005	**0,862**	–0,223	–0,028	**–0,866**	–0,159	0,012	**0,866**	–0,161	0,054
0,667	0,533	0,186	**0,671**	0,526	0,212	**–0,685**	0,410	–0,186	**0,693**	0,332	–0,294	**–0,689**	0,381	–0,269
–0,492	**0,515**	0,142	–0,492	**0,513**	0,359	0,485	**0,594**	–0,298	–0,482	**0,579**	–0,394	0,486	**0,582**	–0,283
–0,264	**0,733**	–0,313	–0,260	**0,741**	–0,287	0,248	**0,695**	0,229	–0,221	**0,761**	0,103	0,218	**0,763**	0,125
0,480	0,173	**0,724**	0,502	0,101	**0,640**	–0,499	–0,030	**–0,586**	0,503	–0,172	**–0,561**	–0,497	–0,092	**–0,650**

Теперь следует задать вопрос о качестве сжатия данных. В отличие от моделированных данных истинная структура корреляционной матрицы наблюдаемых показателей неизвестна. В связи с этим определенная выше метрика качества d не подходит для сравнения различных модификаций редукции реальных данных. Естественным показателем качества сжатия реальных данных представляется доля (или процент) дисперсии исходных признаков, которая описывается первыми (в примере — тремя первыми) главными компонентами. Зависимость суммарной объясненной дисперсии от числа главных компонент для пяти рассматриваемых методов приведена на рис. 2. Наилучшее качество в примере показывает МГК, основанный на оценке Олива — Хокинса (77,8 % дисперсии исходных показателей объясняется первыми тремя главными компонентами). Незначительно уступают ему МГК на основе MCD-оценки (77,1 %) и оценки Гнанадесикана — Кетенринга (75,7 %). Классический МГК, основанный на выборочной оценке Пирсона (69,6 %), существенно проигрывает указанным робастным оценкам. Полученный результат обусловлен тем, что распределения наблюдаемых показателей $X_1, \ldots, X_9, X_{11}, \ldots, X_{13}$ не являются гауссовыми.

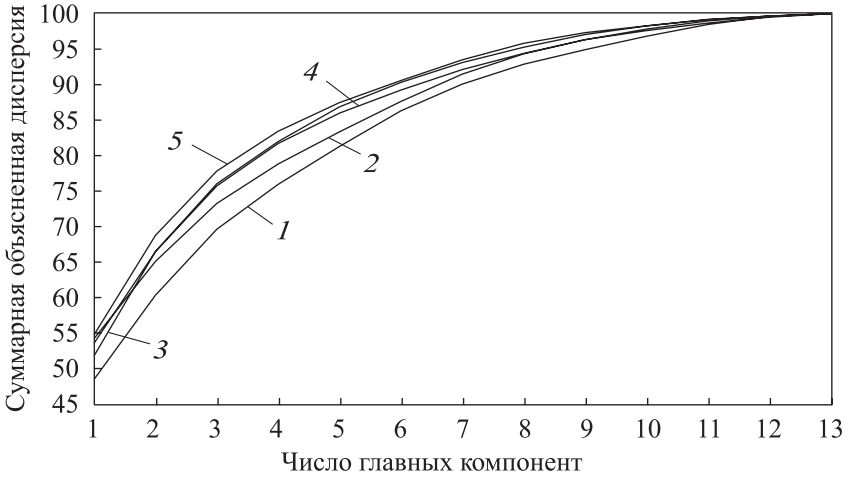

Рис. 2. Зависимость суммарной объясненной дисперсии от числа главных компонент:

1 — оценка Пирсона; *2* — оценка Спирмена; *3* — MCD-оценка; *4* — оценка Гнанадесикана — Кетенринга; *5* — оценка Олива — Хокинса

Заключение. По результатам, представленным на рис. 1, можно сделать вывод о том, что в случаях, когда наблюдения имеют нормальное распределение, наилучшим является классический МГК, основанный на выборочной корреляционной матрице. Однако при самом незначительном

отклонении распределения наблюдений от нормального наилучшими становятся методы, использующие MCD-оценку, оценки Гнанадесикана — Кетенринга и Олива — Хокинса. Следует отметить, что в гауссовом случае преимущество классического МГК по сравнению с указанными робастными модификациями не является существенным. Поэтому при практическом применении МГК необходимо отказаться от классического МГК в пользу МГК, основанного на какой-либо из трех робастных оценках Гнанадесикана — Кетенринга, Олива — Хокинса и MCD-оценке, которые во всех случаях демонстрируют примерно одинаковую эффективность. Метод, основанный на ранговой оценке корреляционной матрицы, хотя и является робастным, но практически всегда уступает этим трем методам. Следовательно, его целесообразно использовать лишь тогда, когда между наблюдениями есть только отношение порядка, поскольку остальные методы в этом случае не работают. Применение робастных модификаций МГК к реальным данным, содержащим аномальные наблюдения, показало хорошую согласованность полученных выводов с результатами моделирования.

ЛИТЕРАТУРА

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

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

[3] Huber P.J., Ronchetti E.M. Robust statistics. Wiley, 2009.

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

[5] 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

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

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

[8] 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

[9] Maronna R.A., Martin R.D., Yohai V.J., et al. Robust statistics. Theory and methods (with *R*). Wiley, 2019.

[10] 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

[11] Olive D.J. A resistant estimator of multivariate location and dispersion. *Comput. Stat. Data Anal.*, 2004, vol. 46, no. 1, pp. 93–102.
DOI: https://doi.org/10.1016/S0167-9473(03)00119-1

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

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

[14] Maronna R. Principal components and orthogonal regression based on robust scales. *Technometrics*, 2005, vol. 47, iss. 3, pp. 264–273.
DOI: https://doi.org/10.1198/004017005000000166

[15] Kotz S., Nadarajah S. Multivariate *T*-distributions and their applications. Cambridge Univ. Press, 2004.

[16] Razali N.M., Wah Y.B. Power comparisons of Shapiro — Wilk, Kolmogorov — Smirnov, Lilliefors, and Anderson — Darling tests. *JOSMA*, 2011, vol. 2, no. 1, pp. 21–33.

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, стр. 1).

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

Просьба ссылаться на эту статью следующим образом:

Горяинов В.Б., Горяинова Е.Р. Исследование устойчивости к аномальным наблюдениям модификаций метода главных компонент. *Вестник МГТУ им. Н.Э. Баумана. Сер. Естественные науки*, 2023, № 2 (107), с. 17–34.
DOI: https://doi.org/10.18698/1812-3368-2023-2-17-34

STUDY OF THE PRINCIPAL COMPONENTS METHOD MODIFICATIONS RESISTANCE TO ABNORMAL OBSERVATIONS

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

The paper considers the problem of reducing multidimensional correlated indicators. One of the approaches to solving this problem is based on the method of principal components, which makes it possible to compactly describe the vector with correlated coordinates (components) using the principal components vector with uncorrelated coordinates of much smaller dimension, while retaining most of the information about correlation structure of the original vector. On simulated and real data, several modifications of the principal components method were compared differing in the method of evaluating correlation matrix of the observation vector. The work objective is to demonstrate advantages of the robust modifications of the principal components method in cases, where data contained the abnormal values. To compare the considered modifications on the model data, metric was introduced that measured the difference between estimated and true eigenvalues of the initial data correlation matrix. This metric behavior depending on the probability distribution of observations was studied by computer simulation. As the distributions, multivariate distributions with the off-diagonal correlation matrices simulating a polluted sample were selected. Next, a sample of 13 correlated socioeconomic indicators for 85 countries was considered, where 46 abnormal values were identified. The considered modifications of the principal components method chose the same optimal number of principal components equal to three. However, the real data compression quality, which was defined as the share of the initial indicators total variance described by the first three principal components, turned out to be significantly higher for the robust modifications of the principal components method. Results obtained on these real data are in good agreement with conclusions of the computer simulation

Keywords

Principal components method, correlation matrix robust evaluation, MCD estimate, Gnanadesikan — Ketenring estimate, Olive — Hawkins estimate, Tukey distribution, bimodal distribution

Received 06.08.2022
Accepted 14.10.2022

REFERENCES

[1] Ayvazyan S.A., ed. Prikladnaya statistika. Klassifikatsiya i snizhenie razmernosti [Applied statistics. Classification and dimension reduction]. Moscow, Finansy i statistika Publ., 1989.

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

[3] Huber P.J., Ronchetti E.M. Robust statistics. Wiley, 2009.

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

[5] 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

[6] Goryainova E.R., Pankov A.P., Platonov E.N. Prikladnye metody analiza statisticheskikh dannykh [Applied metods of statistical data analysis]. Moscow, ID NIU VShE Publ., 2012.

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

[8] 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

[9] Maronna R.A., Martin R.D., Yohai V.J., et al. Robust statistics. Theory and methods (with *R*). Wiley, 2019.

[10] 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

[11] Olive D.J. A resistant estimator of multivariate location and dispersion. *Comput. Stat. Data Anal.*, 2004, vol. 46, no. 1, pp. 93–102.
DOI: https://doi.org/10.1016/S0167-9473(03)00119-1

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

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

[14] Maronna R. Principal components and orthogonal regression based on robust scales. *Technometrics*, 2005, vol. 47, iss. 3, pp. 264–273.
DOI: https://doi.org/10.1198/004017005000000166

[15] Kotz S., Nadarajah S. Multivariate *T*-distributions and their applications. Cambridge Univ. Press, 2004.

[16] Razali N.M., Wah Y.B. Power comparisons of Shapiro — Wilk, Kolmogorov — Smirnov, Lilliefors, and Anderson — Darling tests. *JOSMA*, 2011, vol. 2, no. 1, pp. 21–33.

Goryainov V.B. — Dr. Sc. (Phys.-Math.), Professor, Department of Mathematical Simulation, Bauman Moscow State Technical University (2-ya Baumanskaya ul. 5, str. 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. Study of the principal components method modifications resistance to abnormal observations. *Herald of the Bauman Moscow State Technical University*, *Series Natural Sciences*, 2023, no. 2 (107), pp. 17–34 (in Russ.).
DOI: https://doi.org/10.18698/1812-3368-2023-2-17-34