DOI: 10.25728/mlsd.2023.1377

СРАВНЕНИЕ РОБАСТНЫХ ВЕРСИЙ МЕТОДА ГЛАВНЫХ КОМПОНЕНТ, ОСНОВАННЫХ НА ПРОЕКЦИОННОМ ПРЕСЛЕДОВАНИИ И ОЦЕНИВАНИИ КОРРЕЛЯЦИОННЫХ МАТРИЦ

Горяинов В.Б.
МГТУ им. Н.Э. Баумана, Москва, Россия
vb-goryainov@mail.ru
Горяинова Е.Р.
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
el-goryainova@mail.ru

Аннотация. При помощи компьютерного моделирования сравнивается качество классической и робастных версий метода главных компонент при сжатии данных, имеющих вероятностные распределения с тяжелыми хвостами. На примере редукции многомерного вектора социально-экономических показателей продемонстрирована согласованность полученных результатов с результатами моделирования.

Ключевые слова: Метод главных компонент, проекционное преследование, MCD-оценка, оценка Гнанадесикана — Кетенринга, оценка Олива — Хокинса, распределение Тьюки, бимодальное распределение.

Введение

Метод главных компонент (МГК) — статистический метод компактного описания случайного вектора с коррелированными координатами при помощи его линейного преобразования в вектор существенно меньшей размерности, называемый вектором главных компонент. При этом вектор главных компонент имеет некоррелированные координаты и содержит бо́льшую часть информации о корреляционной структуре исходного вектора. [1, 2].

МГК основан на собственных векторах и собственных значениях ковариационной матрицы наблюдений, которая на практике обычно неизвестна и, как правило, оценивается выборочной ковариационной матрицей. Однако, как показано, например, в [3] и [4], выборочная ковариационная матрица чувствительна к выбросам, что приводит МГК к грубым ошибкам в итоговых результатах. Для исправления этого недостатка были предложены различные варианты робастного МГК.

Методы робастного МГК можно разделить на две группы. К первой группе относятся методы, основанные на замене используемой в МГК выборочной корреляционной матрицы ее робастными аналогами [5].

Второй подход к робастному анализу главных компонент заключается в вычислении при помощи проекционного преследования (projection pursuit) [6] непосредственных робастных оценок собственных значений и собственных векторов, минуя оценку ковариационной матрицы. Как и классический МГК, этот метод ищет направления с максимальной дисперсией проецируемых на эти направления данных, но вместо дисперсии использует робастную оценку масштаба.

В данной работе проводится сравнительный анализ качества этих методов в зависимости от вероятностного распределения исходных данных. В качестве распределений выбраны многомерные нормальное распределение, многомерное распределение Стьдента и распределение Тьюки. Нормальное распределение описывает идеальную ситуацию классического МГК К. Пирсона и Г. Хотеллинга. Распределение Тьюки моделирует наиболее часто встречающуюся на практике ситуацию, когда распределение исходных данных в той или иной степени отклоняется от нормального, в частности за счет грубых ошибок в наблюдениях. Сравнение всех рассмотренных версий МГК иллюстрируется примером построения главных компонент 13-мерного вектора коррелированных социально-экономических показателей.

1. Построение главных компонент

Рассмотрим случайный вектор $X = (X_1, \dots, X_p)^T$ с математическим ожиданием $\mathrm{E}X = 0$ и корреляционной матрицей Σ. Обозначим через $\lambda_1, \dots, \lambda_p$ и $e_1, \dots, e_p$ упорядоченные в порядке убывания

собственные значения и соответствующие им нормированные собственные векторы матрицы Σ соответственно.

Классический МГК ищет направления с максимальной дисперсией проецируемых на эти направления данных, и этими направлениями оказываются $e_1, \dots, e_p$. А именно, в классическом МГК первая главная компонента e_1 определяется как

$$e_1 = arg \ \ max_{|a|=1} \mathrm{D}(a^T X),$$

где $X = (X_1, \dots, X_p)^T$ — случайный вектор наблюдений (исходных данных), а $\mathrm{D}(a^T X)$ — дисперсия случайной величины $a^T X$. Другими словами вектор единичной длины e_1 максимизирует дисперсию спроецированных на него данных X. При этом собственное число λ_1 вычисляется как $\lambda_1 = \mathrm{D}(e_1^T X)$. Аналогично находятся остальные главные компоненты: в предположении, что первые $k-1$ собственных векторов уже найдены ($k > 1$) k-й собственный вектор будет

$$e_k = arg \ \ max_{|a|=1, a \perp e_1, \dots, a \perp e_{k-1}} \mathrm{D}(a^T X), \quad (1)$$

а k-е собственное значение есть $\lambda_k = \mathrm{D}(e_k^T X)$.

В выборочной версии классического МГК дисперсия заменяется выборочной дисперсией $S^2(u_1, \dots, u_n) = \frac{1}{n}\sum_{i=1}^n (u_i - \overline{u})^2$, где $\overline{u} = \frac{1}{n}\sum_{i=1}^n u_i$. А именно, для последовательности наблюдений $x_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$, вектора X оценки собственных векторов и собственных чисел определяются соответственно как

$$e_k = arg \ \ max_{|a|=1, a \perp e_1, \dots, a \perp e_{(k-1)}} S^2(a^T x_1, \dots, a^T x_n), \quad (2)$$

$$\lambda_k = S^2(a^T x_1, \dots, a^T x_n).$$

В работе рассматриваются робастные версии МГК двух типов. В первом случае вместо выборочной ковариационной матрицы рассматриваются ее робастные оценки такие, как оценки MCD-оценка, оценка Гнанадесикана — Кетенринга и оценка Олива — Хокинса. Во втором случае дисперсия $\mathrm{D}(a^T X)$ в (1) заменяется робастной мерой параметра масштаба — квадратом среднего абсолютного отклонения $\mathrm{MAD}(a^T X) = \mathrm{med}|a^T X - \mathrm{med}(a^T X)|$, где med — медиана случайной величины. В выборочной версии медиана заменяется выборочной медианой

$$\mathrm{MAD}(a^T x_1, \dots, a^T x_n), \quad (3)$$

где MAD-оценка среднеквадратического отклонения $\sqrt{\mathrm{D}\xi}$ произвольной случайной величины ξ по ее наблюдениям $u_1, \dots, u_k$ определяется как

$$MAD(u_1, \dots, u_n) = \mathrm{med}(|u_1 - \mathrm{med}(u_1, \dots, u_n)|, \dots, |u_n - \mathrm{med}(u_1, \dots, u_n)|),$$

а $\mathrm{med}(u_1, \dots, u_n)$ — медиана выборки $u_1, \dots, u_n$. Максимизация (1) или (2) является частным случаем проекционного преследования. В терминах проекционного преследования и дисперсия $D(a^T X)$, и робастная мера параметра масштаба (3) называются проекционными индексами.

К сожалению, (3) как функция от a не является выпуклой, что значительно затрудняет ее максимизацию. Вместо этого в данной работе используется приближенный метод нахождения собственных чисел и собственных векторов, предложенный в [7]. Идея этого метода заключается в замене континуального максимума по единичной сфере $|a| = 1$ максимумом на конечном числе точек, а именно, на наблюдениях $x_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$. Соответствующий алгоритм приведен в [7] и заключается в следующем.

Обозначим через μ L_1-медиану, определяемую как $\hat{\mu} = arg \ \ min_{\mu \in \mathbb{R}^p} |x_i - \mu|$, а через $x_i^1 = x_i - \hat{\mu}$, $i = 1, \dots, n$ — центрированные данные.

Первый собственный вектор определяется как $e_1 = arg \ \ max_{a \in A_1} S(a^T x_1^1, \dots, a^T x_n^1)$, где $A_1 = \{x_i^1 / |x_i^1|, i = 1, \dots, n\}$, затем вычисляются $y_i^1 = e_1^T x_i^1$, $i = 1, \dots, n$.

Далее для $k = 2, \dots, q$ последовательно находятся $x_i^k = x_i^{k-1} - y_i^{k-1} e_{k-1}$, $A_k = \{x_i^k / |x_i^k|, i = 1, \dots, n\}$, $e_k = arg \ \ max_{a \in A_k} S(a^T x_1^k, \dots, a^t x_n^k)$, $y_i^k = e_k^T x_i^k$.

В работе версия, основанная на методе проекционного преследования, сравнивается с модификациями МГК, базирующимися на следующих оценках корреляционной матрицы Σ вектора X — оценке Пирсона, Гнанадесикана—Кетенринга, Олива-Хокинса и MCD-оценке. Последние три оценки являются робастными, т.е. имеют ненулевую пороговую точку и, как следствие, являются слабо чувствительными к засорению выборки аномальными наблюдениями (выбросами).

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

MCD-оценка корреляционной матрицы определяется как выборочная корреляционная матрица $\left[\frac{n+1+p}{2}\right]$ наблюдений, которые имеют выборочную корреляционную матрицу с наименьшим определителем среди имеющихся n наблюдений [4].

Оценка Гнанадесикана — Кетенринга [8] элементов матрицы Σ основана на равенстве $\mathrm{cov}(X_i, X_j) = \frac{1}{4}(\mathrm{D}(X_i + X_j) - \mathrm{D}(X_i - X_j))$, в котором $\mathrm{D}(X_i + X_j)$ и $\mathrm{D}(X_i - X_j)$ заменяются квадратами MAD-оценок [9, с.5].

Оценка Олива — Хокинса [10, 11] пропорциональна выборочной корреляционной матрице, построенной из половины $n/2$ наблюдений. Эти $n/2$ наблюдений либо являются ближайшими к медиане наблюдений $\mathrm{M}ed(x) = (\mathrm{m}ed(x_{11}, \dots, x_{n1}), \dots, \mathrm{m}ed(x_{1p}, \dots, x_{np}))$ в смысле евклидова расстояния, либо имеют наименьшее расстояние Махаланобиса от выборочного среднего значения наблюдений, рассчитанного с использованием выборочной корреляционной матрицы $\hat{\Sigma}$. Точное определение оценки и алгоритм ее вычисления приведены в [10, 11].

2. Результаты моделирования

Основной задачей МГК является интерпретируемость построенных главных компонент, определяемая структурой матрицы нагрузок, столбцами которой являются собственные векторы корреляционной матрицы наблюдений. С этой целью в [12] была предложена мера качества МГК равная евклидовому расстоянию между редуцированной оцененной матрицей нагрузок и редуцированной эталонной матрицей нагрузок, соответствующей смоделированным зависимостям. Но вычисление этой метрики оказывается достаточно трудоёмким.

В данной работе метрика основана на мере относительной ошибки прогнозирования из [13] и определяется величиной

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

где q —количество оптимальным образом выбранных главных компонент, $\hat{\lambda}_1, \dots, \hat{\lambda}_q$ — оценки собственных чисел $\lambda_1, \dots, \lambda_q$ корреляционной матрицы Σ . Наилучшей считалась модификация МГК с наименьшим значением d.

При сравнительном анализе корреляционная матрица Σ имела вид

$$\Sigma = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}, \qquad A = \begin{pmatrix} 1 & -0.9 & 0.8 \\ -0.9 & 1 & -0.7 \\ 0.8 & -0.7 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & -0.8 & -0.7 \\ -0.8 & 1 & 0.6 \\ -0.7 & 0.6 & 1 \end{pmatrix}. \qquad (4)$$

Это позволило моделировать шестимерные ($p = 6$) векторы, состоящие из двух двумерных подвекторов так, чтобы координаты разных подвекторов были некоррелированны, а координаты каждого подвектора были сильно коррелированны между собой с парными коэффициентами корреляции 0.6–0.9. Вследствие этого все версии МГК выделяли две ($q = 2$) главные компоненты.

В работе классический МГК, основанный на выборочной корреляционной матрице Пирсона, сравнивался с робастными модификациями МГК, использующими MCD-оценку, ортогонализованные оценки Гнанадесикана — Кетенринга и оценки Олива — Хокинса, а также с методом проекционного преследования.

Для всех указанных версий МГК вычислялся показатель эффективности d в ситуациях, когда вектор наблюдений X имел одно из следующих распределений: нормальное распределение, распределение Тьюки, распределение Стьюдента и бимодальное (двугорбое) нормальное распределение. Для этого $N =$

10^4 раз моделировалась выборка вектора X объёма $n = 100$, имеющего одно из вышеперечисленных распределений с корреляционной матрицей (4). Качество каждой из пяти рассмотренных версий МГК оценивалось величиной $\overline{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$, где d_i — значение d в i-м численном эксперименте, $i = 1, \dots, N$. Ниже дано описание указанных вероятностных распределений X и представлены результаты моделирования для каждого из них при объёмах выборок.

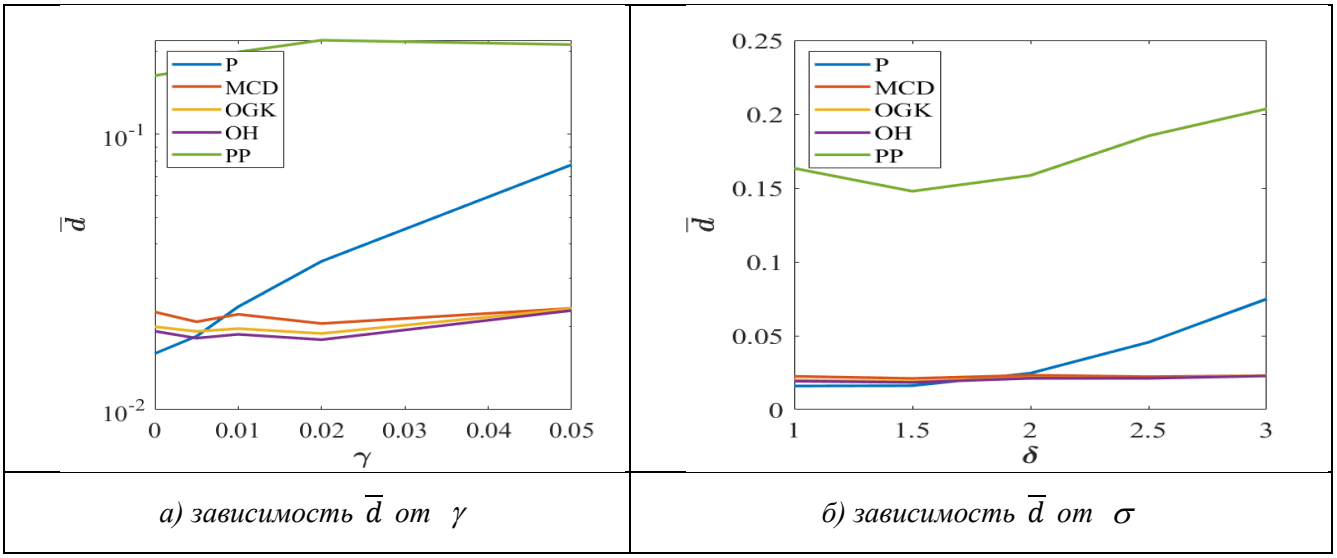
Распределение Тьюки или загрязненное (засоренное) нормальное распределение с долей загрязнения γ и величиной загрязнения δ описывается плотностью

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

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.

Распределение Тьюки имитирует засорение нормально распределенных наблюдений небольшой долей γ наблюдений, дисперсия которых в δ^2 раз превышает дисперсию основной части наблюдений. При $\delta = 1$ или $\gamma = 0$ оно совпадает с нормальным распределением. Считается [3], 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 при помощи компьютерного моделирования для различных версий МГК исследовалась зависимость от γ и δ оценки $\overline{d}$ показателя эффективности d.



а) зависимость $\overline{d}$ от γ

б) зависимость $\overline{d}$ от σ

Рис. 1. Зависимость метрики качества $\overline{d}$ пяти версий МГК от доли загрязнения γ и величины загрязнения σ распределения Тьюки. На каждом графике рисунка 2 линия P соответствует оценке Пирсона, линия MCD — MCD-оценке, линия OGK — оценке Гнанадесикана — Кетенринга, линия OH — оценке Олива — Хокинса, линия PP — методу проекционного преследования.

На рис. 1а) для различных версий МГК приведены графики зависимости $\overline{d}$ от γ при фиксированном $\delta = 3$. Видно, что МГК, основанный на выборочной корреляционной матрице, является лучшим только при отсутствии засорений ($\gamma = 0$) и становится наихудшим при $\delta = 3$ уже для $\gamma > 0.01$. Наилучшими следует признать методы МГК, использующие MCD-оценку, оценку Гнанадесикана — Кетенринга и оценку Олива — Хокинса. В отсутствие засорений они практически не уступают методам, основанным на выборочной корреляционной матрице, а при наличии даже небольших засорений заметно их превосходят.

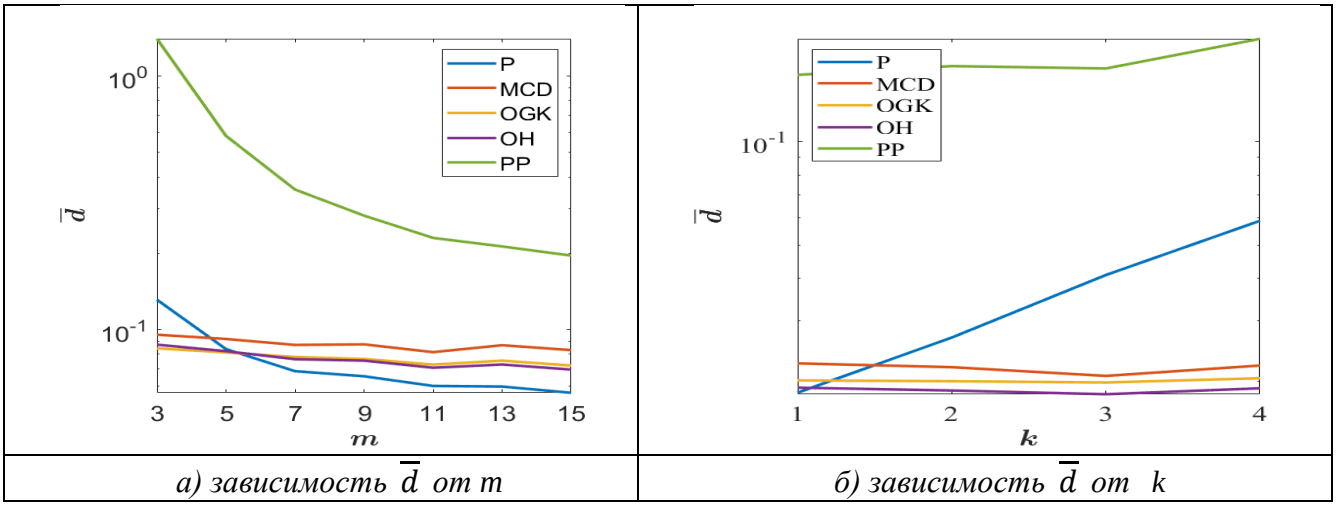
На рис. 1б) для для этих же версий МГК приведены графики зависимости $\overline{d}$ от δ при фиксированном $\gamma = 0.01$. Видно, что МГК, основанный на выборочной корреляционной матрице, является худшим уже при $\delta > 2$. И опять наилучшими следует признать методы МГК, использующие MCD-оценку, оценку Гнанадесикана — Кетенринга и оценку Олива — Хокинса.

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

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

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

В работе при помощи компьютерного моделирования исследовалась зависимость показателя $\overline{d}$ эффективности различных версий МГК от m в предположении, что матрица Σ имеет вид (4). На рис. 2а) для пяти изучаемых версий МГК и выборок объёма $n=100$ приведены графики зависимости $\overline{d}$ от m. Видно, что МГК, основанный на выборочной корреляционной матрице, является конкурентоспособным при $m=3$ и 4, и становится лучшим при $m>5$. Методы МГК, использующие оценки MCD, оценки Гнанадесикана — Кетенринга и оценки Олива — Хокинса являются наилучшими при $m<5$ и конкурентоспособными при $m>5$. Таким образом, если случайный вектор наблюдений X имеет многомерное распределение Стьюдента, то разумно всегда использовать методы МГК, основанные на оценках MCD, Гнанадесикана — Кетенринга и Олива — Хокинса.



а) зависимость $\overline{d}$ от m | *б) зависимость $\overline{d}$ от k*

Рис. 2. Зависимость метрики качества $\overline{d}$ пяти версий МГК от степени свободы m распределения Стьюдента и расстояния k между горбами двугорбого нормального распределения. На каждом графике рисунка 2 линия P соответствует оценке Пирсона, линия MCD — MCD-оценке, линия OGK — оценке Гнанадесикана — Кетенринга, линия OH — оценке Олива — Хокинса, линия PP — методу проекционного преследования.

Бимодальное нормальное распределение. Рассмотрим ещё 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 $f(x,\mu,\Sigma)$, где μ — расстояние между центрами кластеров, γ — доля засорения. Канонической иллюстрацией к бимодальному распределению является распределение антропометрических характеристик (рост, окружность головы и т.д.) индивидов, выбранных из двух

кластеров (например, мужчин и женщин). Показатели в обоих кластерах имеют схожую структуру связей, но разные средние значения, соответствующие горбам бимодального распределения. В работе предполагалось, что $\mu = (k, 0,0,0,0,0)$. Для различных версий МГК на рис. 2б) приведены графики зависимости $\overline{d}$ от k при фиксированном $\gamma = 0.01$. Видно, что МГК, основанный на выборочной корреляционной матрице, является лучшим только при $k < 1$ и становится наихудшим при $k > 2$. При $k > 2$ наилучшими являются методы МГК, использующие MCD-оценку, оценку Гнанадесикана — Кетенринга и оценку Олива — Хокинса.

Метод проекционного преследования оказался наихудшим на всех рассмотренных распределениях. Причина этого состоит в наличии множества локальных максимумов функции (3). Поскольку поиск максимума функции (3) осуществляется с помощью предложенной в [7] приближённой процедуры, то получаемые оценки собственных значений оказываются достаточно грубыми.

3. Пример с реальными социально-экономическими данными

С сайта https://www.cia.gov/the-world-factbook/countries/ были взяты данные о 13 социально-экономических показателях по 85 странам. Исследуемыми показателями являются: 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 страны 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}. Отметим, что каждый из показателей $X_1, \dots, X_{13}$ имеет высокие коэффициенты корреляции с большинством других показателей. Согласно критерию Шапиро — Уилка [14] гипотеза о гауссовости всех показателей, за исключением X_{10}, была отвергнута на уровне значимости не более 0.008.

Проведём поиск резко выделяющихся наблюдений. Будем считать наблюдение y_i аномальным наблюдением среди $y_1, \dots, y_n$, $n = 85$, если $y_i \notin (\hat{\mu} - 3\hat{\sigma}, \hat{\mu} + 3\hat{\sigma})$, где $\hat{\mu} = \mathrm{med}(y_1, \dots, y_n)$, $\hat{\sigma} = \mathrm{MAD}(y_1, \dots, y_n)/\Phi^{-1}(0.75)$, а $\Phi^{-1}(x)$ — функция, обратная к функции распределения вероятностей стандартной нормальной случайной величины. У 29 стран выбросы были обнаружены по крайней мере в одном из 13 показателей. А такие страны, как Афганистан, Джибути, Мали, Мозамбик, ЦАР, имели аномальные значения не менее, чем по трем показателям. Теперь дадим компактное описание наблюдаемых коррелированных признаков с помощью меньшего числа некоррелированных показателей (главных компонент). Определим оптимальное количество главных компонент, необходимых для описания наблюдаемого вектора показателей, как количество собственных значений больших единицы. Все рассмотренные модификации выделили одинаковое оптимальное число главных компонент равное трём.

Рассмотрим теперь вопрос о качестве сжатия данных. В отличие от рассмотренных выше моделированных данных истинная структура корреляционной матрицы наблюдаемых показателей неизвестна. По этой причине метрика качества d не подходит для сравнения различных версий сжатия реальных данных. Естественным показателем качества редукции реальных данных представляется доля дисперсии исходных признаков, которая описывается первыми (в данном примере тремя первыми) главными компонентами. График зависимости суммарной объяснённой дисперсии от количества главных компонент для пяти рассматриваемых методов представлен на рис. 3. Наилучшее качество в данном примере показывает МГК, основанный на оценке Олива-Хокинса (77,8% дисперсии исходных показателей объясняется первыми тремя главными компонентами). Незначительно уступают ему МГК на основе оценки MCD (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 $X_1, \dots, X_9, X_{11}, \dots, X_{13}$ не являются гауссовскими. Метод проекционного преследования показал наихудший результат (53%). Итак, результаты полученные в этом примере с реальными данными хорошо согласуются с результатами численного моделирования.

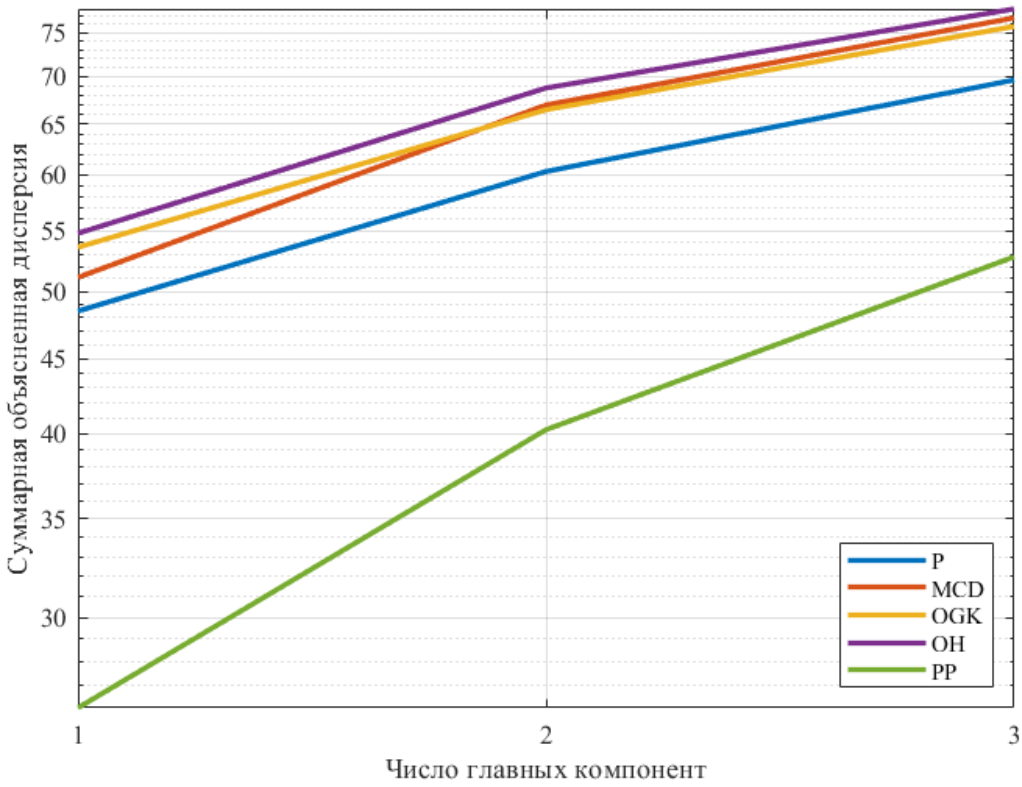


Рис. 3. Линия P соответствует оценке Пирсона, линия MCD — MCD-оценке, линия OGK — оценке Гнанадесикана — Кетенринга, линия OH — оценке Олива — Хокинса, линия PP — PP-оценке

4. Заключение

В работе была рассмотрена традиционная и робастные версии МГК, предназначенные для редукции многомерных показателей. С помощью введённой метрики был проведён сравнительный анализ качества сжатия полученных методом численного моделирования многомерных коррелированных показателей для всех рассмотренных модификаций МГК. Результаты статистического моделирования, представленные на рис. 1 и 2, показывают, что в случае гауссовского распределения показателей, наилучшим является традиционный МГК, использующий выборочную оценку корреляционной матрицы. Однако для бимодального распределения, распределений Тьюки и Стьюдента, имитирующих различные виды засорения данных, традиционный метод уступает первенство робастным версиям МГК, основанным на робастных оценках Олива-Хокинса, Гнанадесикана-Кеттенринга и MCD. Среди трёх указанных робастных модификаций некоторое преимущество на всех распределениях имела версия Олива-Хокинса. Важно отметить, что в гауссовском случае робастные версии МГК незначительно уступают традиционной. Неконкурентоспособность метода проекционного преследования в данной задаче обусловлена тем, что приближённый метод нахождения максимума невыпуклой функции не даёт хороших приближений на датасетах умеренного объёма. Выводы, полученные методом численного моделирования, достаточно хорошо согласуются с результатами сжатия многомерного датасета коррелированных социально-экономических показателей, содержащих аномальные наблюдения.

Литература

1. *Дронов С.В.* Многомерный статистический анализ. – Барнаул: Изд-во Алт. гос. ун-та, 2003. - 213 с.
2. *Jolliffe I.T.* Principal component analysis. - Second edition. New York: Springer-Verlag, 2002. – 487 p.
3. *Huber P.J., Ronchetti E.M.* Robust statistics. – Hoboken: Wiley, 2009. 360 p.
4. *Rousseeuw P.J., Leroy A.M.* Robust Regression and Outlier Detection. – Chichester: Wiley, 1987. – 347 p.

5. *Devlin S.J., Gnanadesikan R., Kettenring J.R.* Robust estimation of dispersion matrices and principal components // J. Amer. Statist. Assoc. – 1981. – Vol. 76. – P. 354–362.
6. *Li G., Chen Z.* Projection-Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo // Journal of the American Statistical Association. – 1985. – Vol. 80. – P.759–766.
7. *Croux C., Ruiz-Gazen A.* High breakdown estimators for principal components: The projection-pursuit approach revisited // J. Multivariate. Anal. – 2005. – Vol. 95. – P. 206 – 226.
8. *Gnanadesikan R., Kettenring J.R.* Robust estimates, residuals, and outlier detection with multiresponse data // Biometrics. – 1972. – Vol. 28. N 1. – P. 81–124.
9. *Maronna R.A., Martin R.D., Yohai V.J., Salibián-Barrera M.* Robust Statistics: theory and Methods (with R). – Chichester: Wiley, 2019 – 453 p.
10. *Olive D.J.* A resistant estimator of multivariate location and dispersion // Comput. Statist. Data Anal. – 2004. – Vol. 46, N 1. – P. 93–102.
11. *Olive D.J.* Robust multivariate analysis. – Cham: Springer, 2017. – 501 p.
12. *Горяинов В.Б., Горяинова Е.Р.* Сравнительный анализ качества робастных модификаций метода главных компонент при сжатии коррелированных данных // Вестник Московского государственного технического университета им. Н.Э. Баумана. Серия: Естественные науки. – 2021. N 3 (96). – С. 23–45.
13. *Maronna R.* Principal components and orthogonal regression based on robust scales // Technometrics. – 2005. – Vol. 47, N 3. – P. 264–273.
14. *Razali N.M., Wah Y.B.* Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling Tests // Journal of Statistical Modeling and Analytics. Kuala Lumpur: Institut Statistik Malaysia. – 2011. – V 2, N 1. – P. 21–33.