Использование метода главных компонент при построении интегральных индикаторов*

М. М. Медведникова
medvmasha@rambler.ru
Московский физико-технический институт, ФУПМ, каф. «Интеллектуальные системы»

В данной работе рассматривается использование метода главных компонент при построении интегральных индикаторов. Полученные результаты сравниваются с результатами, даваемыми методом расслоения Парето. Строится интегральный индикатор для российских вузов. Для этого используются биографии 30 богатейших бизнесменов России по версии журнала «Forbes» за 2011 год.

Ключевые слова:*интегральный индикатор, экспертные оценки, веса параметров, метод главных компонент, метод расслоения Парето.*

Principal component analysis for building integral indicators*

M. M. Medvednikova
Moscow Institute of Physics and Technology

The main goal of this work is to present principal component analysis for integral indicators construction. Derived results are compared with Pareto slicing method's results. The integral indicator for Russian universities is built by using biographies of 30 the richest Russian businessmen according to magazine "Forbes"in 2011.

Keywords: *integral indicator, expert estimations, feature weights, principal component analysis, Pareto slicing.*

Введение

Современная востребованность рейтингов высших учебных заведений обусловлена существованием большого числа вузов. В частности, в России на данный момент их насчитывается более тысячи. Существует достаточно много рейтингов, построенных с использованием различных критериев [1, 2]. В данной работе предлагается в качестве критерия для оценки вуза использовать успешность карьеры выпускников. Предполагается, что чем выше качество образования, тем выше человек продвигается по карьерной лестнице, не зависимо от того, работает он по полученной специальности, или нет.

Для построения рейтингов используются интегральные индикаторы. Построение интегрального индикатора — введение отношения порядка на множестве сравнимых объектов. Предполагается, что каждый объект описан вектором, компоненты которого являются результатами измерений соответствующих показателей. Множество рассматриваемых объектов называется *выборкой*. Выборка полностью описывается матрицей, строками которой являются векторы, сопоставляемые объектам. Все измерения выполнены в линейных шкалах. *Линейная шкала* — это шкала, на которой равным отрезкам соответствует равные абсолютные приращения показателя. *Интегральный индикатор* — скаляр, поставленный в соответствие объекту. Интегральный индикатор для набора объектов — вектор, компоненты которого поставлены в соответствие сравниваемым объектам.

Научный руководитель В. В. Стрижов

Распространенным алгоритмом [4, 5, 6, 7, 8, 9] построения интегральных индикаторов для объектов, описанных в линейных шкалах, является линейная комбинация значений показателей. Основная задача заключается в определении весов показателей.

Существуют две основные разновидности рассматриваемой задачи. Первая — построение интегрального индикатора методом «с учителем». В этом случае имеются экспертные оценки качества объектов и важности показателей, необходимо согласовать значения интегрального индикатора и весов показателей. Для этого разработаны различные алгоритмы: использующие экспертные оценки качества объектов [3], использующие оценки качества обектов и весов признаков и уточняющие эти оценки [4, 5]. Вторая разновидность — построение индикатора методом «без учителя». Веса вычисляются исходя из некоторого заданного критерия информативности описаний. В этом случае используется метод расслоения Парето [6], вычисления расстояний [7], метод главных компонент [8, 9, 10, 11].

В настоящей работе для построения интегрального индикатора будут использованы метод главных компонент и метод расслоения Парето. Метод главных компонент заключается в том, что к множеству описаний обектов применяется преобразование вращения, которое соответствует критерию наибольшей информативности С. Р. Рао [10]. Согласно этому критерию, наибольшая информативность есть минимальное значение суммы квадратов расстояния от описания обектов до их проекций на первую главную компоненту. Приводится подробное изложение теоретического обоснования метода главных компонент в методических целях. Метод расслоения Парето состоит в разделении выборки на слои несравнимых объектов. В статье приведены описание алгоритма и его теоретическое обоснование, базирующееся на [10], для метода главных компонент и описание алгоритма для метода расслоения Парето, также представлены результаты вычислительных экспериментов для рассматриваемых методов и проведено их сопоставление.

Постановка задачи в общем виде

Дана матрица «объекты-признаки» A. Каждая строка $\mathbf{a}_i^T$, $i = 1, ..., p$ этой матрицы — это вектор, описывающий обект. В данной работе предполагается, что в матрице A данные представлены полностью, без пропусков.

Требуется найти отображение

$$F : A \to \mathbf{q},$$

сопостовляющее каждой строке $\mathbf{a}_i^T$ матрицы A интегральный индикатор q_i.

Предложенное ниже обоснование метода главных компонент, используемого при решении данной задачи, является авторской версией изложения [10].

Базис Грама — Шмидта. Пусть $\mathbf{u}^T = (u_1, ..., u_p)$ — p-мерная случайная величина, с нулевым математическим ожиданием и ковариационной матрицей Σ ранга $m \leqslant p$.

$$M(\mathbf{u}) = \mathbf{0}.$$

$$\operatorname{rang}(\Sigma) = m \leqslant p.$$

Введем обозначение:

$$\mathfrak{M}(\mathbf{u}) = \mathfrak{M}(u_1, ..., u_p) = \{c_1 u_1 + ... + c_p u_p \mid c_i \in \mathbb{R}\}.$$

По определению $\mathfrak{M}(\mathbf{u})$ является линейным пространством. Определим скалярное произведение двух элементов y_1, y_2 из $\mathfrak{M}(\mathbf{u})$ следующим образом:

$$\langle y_1, y_2 \rangle = \operatorname{cov}(y_1, y_2) = \mathrm{M}(y_1 y_2).$$

Тогда нормой элемента y_1 является квадратный корень из его дисперсии:

$$\|y_1\| = \sqrt{\mathrm{D}(y_1)} = \sqrt{\mathrm{M}(y_1^2)}.$$

Из линейной алгебры известно, что $\mathfrak{M}(\mathbf{u})$ имеет ортонормированный базис $g_1, ..., g_m$, где $g_1, ..., g_m$ — попарно не коррелированные случайные величины с единичной дисперсией. Тогда каждый элемент пространства $\mathfrak{M}(\mathbf{u})$ может быть представлен в виде:

$$u_i = a_{i1}g_1 + ... + a_{im}g_m, \quad i = 1, ..., p.$$

Или в матричной форме:

$$\mathbf{u} = A\mathbf{g},$$

где $\mathbf{g}^T = (g_1, ..., g_m)$ и матрица $A = \|a_{ij}\|$. При этом элементы ковариационой матрицы Σ выражаются следующим образом:

$$\Sigma_{ij} = \langle u_i, u_j \rangle = \mathrm{cov}(u_i, u_j) = \sum_{k=1}^{m} a_{ik}a_{jk},$$

$$\Sigma = AA^T.$$

Обратно, если $\mathfrak{M}(\mathbf{u}) = \mathfrak{M}(\mathbf{g})$, то вектор $\mathbf{g}$ может быть выражен через вектор $\mathbf{u}$:

$$\mathbf{g} = B\mathbf{u},$$

$$I = B\Sigma B^T,$$

где I — единичная матрица.

Также из линейной алгебры известно, что размерность пространства $\mathfrak{M}(\mathbf{u})$ равна рангу матрицы скалярных произведений элементов $u_1, ..., u_p$, которой в данном случае является матрица Σ. Следовательно, число случайных величин в ортонормированном базисе равно $m = \mathrm{rang}(\Sigma)$. Ортонормированный базис не единственен. Однако некоторые специальные базисы представляют статистический интерес, их мы и будем рассматривать.

Линейный предиктор как проекция. Пусть $P(u_i)$ — проекция u_i на $\mathfrak{M}(u_1, ..., u_{i-1})$. По определению это линейная функция переменных $u_1, ..., u_{i-1}$, которая определяется как

$$P(u_i) = b_1^* u_1 + ... + b_{i-1}^* u_{i-1} = \underset{b_1,...,b_{i-1}}{\arg\min} \|u_i - \sum_{r=1}^{i-1} b_r u_r\|^2 = \underset{b_1,...,b_{i-1}}{\arg\min} \mathrm{M}(u_i - \sum_{r=1}^{i-1} b_r u_r)^2.$$

Следовательно, $P(u_i)$ является линейным предиктором случайной величины u_i, основанным на величинах $u_1, ..., u_{i-1}$, с минимальной среднеквадратичной ошибкой. Более того,

$$u_i - P(u_i) \perp \mathfrak{M}(u_1, ..., u_{i-1}).$$

Поэтому коэффициенты $b_1^*, ..., b_{i-1}^*$ наилучшей линейной функции определяются из условия равенства нулю скалярных произведений разности $u_i - \sum_{r=1}^{i-1} b_r u_r$ и некоторых $u_s, s = 1, ...i-1$. Поскольку скалярное произведение определено как ковариация, то условие имеет вид:

$$\mathrm{cov}(u_s, u_i - \sum_{r=1}^{i-1} b_r u_r) = 0, \quad s = 1, ..., i-1,$$

откуда следует, что

$$\operatorname{cov}[u_s, u_i - P(u_i)] = 0, \quad s = 1, ..., i-1,$$

что равносильно следующему:

$$\operatorname{cov}(u_s, u_i) = \operatorname{cov}[u_s, P(u_i)], \quad s = 1, ..., i-1.$$

В таком случае имеем:

$$\operatorname{cov}[u_i - P(u_i), u_j - P(u_j)] = \operatorname{cov}(u_i, u_j) - \operatorname{cov}[u_i, P(u_j)] + \operatorname{cov}[P(u_i), P(u_j)] - \operatorname{cov}[P(u_i), u_j] = 0, \quad i \neq j.$$

Если обозначить остаток $u_i - P(u_i)$ через $u_{i,12...i-1}$, то предыдущая формула эквивалентна утверждению, что остатки

$$u_1,\ u_{2,1},\ u_{3,12},\ \ldots,\ u_{p,12...p-1}$$

попарно не коррелированы.

Ортонормированный базис Грама — Шмидта. Пусть $t_{ii} = \|u_i - P(u_i)\|$ — норма (квадратный корень из дисперсии) остатка $u_{i,12...i-1}$. Рассмотрим

$t_{11}g_1 = u_1 = u_1,$
$t_{22}g_2 = u_2 - P(u_2) = u_2 - t_{21}g_1,$
.....
$t_{pp}g_p = u_p - P(u_p) = u_p - t_{p1}g_1 - ... - t_{p,p-1}g_{p-1}.$

Отметим, что проекция случайной величины u_i на $\mathfrak{M}(u_1, ..., u_{i-1})$ может быть выражена через $g_1, ..., g_{i-1}$, а коэффициенты последовательно определены с помощью процесса ортогонализации Грама — Шмидта. Если $t_{ii} = 0$, то u_i полностью определяется предыдущими случайными величинами. В противном случае, если $t_{ii} \neq 0$, то

$$g_i = \frac{1}{t_{ii}}[u_i - P(u_i)] = \frac{1}{t_{ii}}u_{i,12...i-1},$$

так что $\|g_i\| = 1$. Определенные таким образом величины g_i, соответствующие ненулевым t_{ii}, составляют ортонормированный базис. Обратные соотношения имеют вид:

$u_1 = t_{11}g_1,$
$u_2 = t_{21}g_1 + t_{22}g_2,$
.....
$u_p = t_{p1}g_1 + t_{p2}g_2 + ... + t_{pp}g_p$

и могут быть переписаны в виде:

$$\mathbf{u} = T\mathbf{g},$$

где T — нижняя треугольная матрица. Тогда можно выразить ковариационную матрицу Σ через матрицу T следующим образом:

$$\Sigma = TT^T.$$

Анализ главных компонент

Рассмотрим p случайных величин $\mathbf{u} = (u_1, ..., u_p)$ с ковариационной матрицей Σ. Пусть $\lambda_1 \geqslant ... \geqslant \lambda_p$ — собственные числа, а $\mathbf{p}_1, ..., \mathbf{p}_p$ — соответствующие им собственные векторы матрицы Σ. Тогда, как известно из линейной алгебры,

$$\mathbf{p}_i^T \Sigma \mathbf{p}_i = \lambda_i; \quad \mathbf{p}_i^T \Sigma \mathbf{p}_j = 0, \quad i \neq j.$$

Рассмотрим случайные величины, получающиеся в результате преобразования

$$y_i = \mathbf{p}_i^T \mathbf{u}, \quad i = 1, ..., p.$$

Обозначим через $\mathbf{y}$ вектор новых случайных величин и через P ортогональную матрицу со столбцами из собственных векторов матрицы Σ:

$$\mathbf{y}^T = (y_1, ..., y_p), \quad P = (\mathbf{p}_1, ..., \mathbf{p}_p).$$

Тогда вектор $\mathbf{y}$ можно получить из вектора $\mathbf{u}$ с помощью ортогонального преобразования:

$$\mathbf{y} = P\mathbf{u}.$$

Случайная величина y_i называется i-той главной компонентой случайной величины $\mathbf{u}$.

Свойства главных компонент

1. *Главные компоненты не коррелированы. Дисперсия i-той главной компоненты равна* λ_i.
   Это следует из соотношений:

$$\mathrm{D}(\mathbf{p}_i^T \mathbf{u}) = \langle \mathbf{p}_i^T \mathbf{u}, \mathbf{p}_i^T \mathbf{u} \rangle = \langle p_{i1}u_1 + ... + p_{ip}u_p,\ p_{i1}u_1 + ... + p_{ip}u_p \rangle = \mathbf{p}_i^T \Sigma \mathbf{p}_i = \lambda_i;$$

$$\mathrm{cov}(\mathbf{p}_i^T \mathbf{u}, \mathbf{p}_j^T \mathbf{u}) = \mathbf{p}_i^T \Sigma \mathbf{p}_j = 0, \quad i \neq j.$$

   Таким образом, линейное преобразование $\mathbf{y} = P\mathbf{u}$ переводит коррелированное множество случайных величин в некоррлерованное.
2. *Пусть* $g_i = \frac{1}{\sqrt{\lambda_i}} \boldsymbol{p}_i^T \boldsymbol{u}$ *для* $\lambda_i \neq 0$ *и* $rang(\Sigma) = r$, *так что отличными от нуля оказываются первые* r *собственных чисел матрицы* Σ. *Тогда* $g_1, ..., g_r$ — *ортонормированный базис случайной величины* $\boldsymbol{u}$.
3. *Пусть* $\boldsymbol{b}$ — *произвольный вектор, такой что* $\|\boldsymbol{b}\| = 1$. *Тогда дисперсия* $D(\boldsymbol{b}^T \boldsymbol{u})$ *достигает максимума при* $\boldsymbol{b} = \boldsymbol{p}_1$ *и этот максимум равен* λ_1:

$$\boldsymbol{p}_1 = \arg\max_{\|\boldsymbol{b}\|=1} D(\boldsymbol{b}^T \boldsymbol{u});$$

$$\lambda_1 = \max_{\|\boldsymbol{b}\|=1} D(\boldsymbol{b}^T \boldsymbol{u}).$$

4. Следствия из пунктов 1-3:

   —

$$\min_{\|\mathbf{b}\|=1} \mathrm{D}(\mathbf{b}^T \mathbf{u}) = \lambda_p = \mathrm{D}(\mathbf{p}_p^T \mathbf{u})$$

   —

$$\max_{\|\mathbf{b}\|=1, \mathbf{b} \perp \mathbf{p}_1, ..., \mathbf{p}_{l-1}} \mathrm{D}(\mathbf{b}^T \mathbf{u}) = \lambda_l = \mathrm{D}(\mathbf{p}_l^T \mathbf{u})$$

   — Пусть $\mathbf{b}_1, ..., \mathbf{b}_k$ — множество ортогональных вектров с единичной нормой. Тогда

$$\lambda_1 + ... + \lambda_k = \max_{\mathbf{b}_1, ..., \mathbf{b}_k} [\mathrm{D}(\mathbf{b}_1^T \mathbf{u}) + ... + \mathrm{D}(\mathbf{b}_k^T \mathbf{u})] = \mathrm{D}(\mathbf{p}_1^T \mathbf{u}) + ... + \mathrm{D}(\mathbf{p}_k^T \mathbf{u}).$$

5. **Утверждение**
Пусть $\boldsymbol{b}_1^T\boldsymbol{u}, ..., \boldsymbol{b}_k^T\boldsymbol{u}$ *— k линейных функций случайной величины* $\boldsymbol{u}$ *и* σ_i^2 *— остаточная дисперсия в предсказании* u_i *с помощью наилучшего линейного предиктора, основанного на* $\boldsymbol{b}_1^T\boldsymbol{u}, ..., \boldsymbol{b}_k^T\boldsymbol{u}$*. Тогда*

$$\min_{\boldsymbol{b}_1,...,\boldsymbol{b}_k} \sum_{i=1}^{p} \sigma_i^2$$

достигаетсяв случае, если множество $\boldsymbol{b}_1^T\boldsymbol{u}, ..., \boldsymbol{b}_k^T\boldsymbol{u}$ *эквивалентно множеству* $\boldsymbol{p}_1^T\boldsymbol{u}, ..., \boldsymbol{p}_k^T\boldsymbol{u}$*, то есть каждая из величин* $\boldsymbol{b}_i^T\boldsymbol{u}$ *есть линейная комбинация первых k главных компонент.*

Доказательство
По определению

$$\sigma_i^2 = \|u_i - \sum_{j=1}^{k} \langle \mathbf{b}_j^T\mathbf{u}, u_i\rangle \mathbf{b}_j^T\mathbf{u}\|^2 = \mathrm{D}(\langle \mathbf{b}_j^T\mathbf{u}, u_i\rangle \mathbf{b}_j^T\mathbf{u}) = \mathrm{D}(u_i - P(u_i)).$$

Без ограничения общности можно считать $\mathbf{b}_1^T\mathbf{u}, ..., \mathbf{b}_k^T\mathbf{u}$ некоррелированными функциями с единичной дисперсией. Для оптимального решения эти функции должны быть линейно независимы.
Проведем преобразование выражения для σ_i^2.

$$\sigma_i^2 = \mathrm{D}(u_i - P(u_i)) = \mathrm{D}(u_i) + \mathrm{D}[P(u_i)] - 2\mathrm{cov}[u_i, P(u_i)] =$$

$$= \mathrm{D}(u_i) + \mathrm{D}[\sum_{j=1}^{k} \langle \mathbf{b}_j^T\mathbf{u}, u_i\rangle \mathbf{b}_j^T\mathbf{u}] - 2\sum_{j=1}^{k} \langle u_i, \langle \mathbf{b}_j^T\mathbf{u}, u_i\rangle \mathbf{b}_j^T\mathbf{u}\rangle$$

Т.к. все $\mathbf{b}_i^T\mathbf{u}$ были заменены на некоррелированные величины с единичной дисперсией, то можно продолжить следующим образом:

$$\sigma_i^2 = \mathrm{D}(u_i) + \sum_{j=1}^{k} \langle \mathbf{b}_j^T\mathbf{u}, u_i\rangle^2 - 2\sum_{j=1}^{k} \langle \mathbf{b}_j^T\mathbf{u}, u_i\rangle^2 =$$

$$= \mathrm{D}(u_i) - \sum_{j=1}^{k} \langle \mathbf{b}_j^T\mathbf{u}, u_i\rangle^2$$

Обозначим $\mathrm{D}(u_i)$ как σ_{ii}, а i-тый столбец матрицы Σ как Σ_i. Тогда получаем, что

$$\sigma_i^2 = \sigma_{ii} - \sum_{j=1}^{k} [\mathrm{cov}(u_i, \mathbf{b}_j^T\mathbf{u})]^2 =$$

$$= \sigma_{ii} - \sum_{j=1}^{k} [b_{j1}\langle u_i, u_1\rangle + ... + b_{jp}\langle u_i, u_p\rangle]^2 =$$

$$= \sigma_{ii} - \sum_{j=1}^{k} [\mathbf{b}_j^T\Sigma_i]^2 = \sigma_{ii} - \sum_{j=1}^{k} [\mathbf{b}_j^T\Sigma_i\Sigma_i^T\mathbf{b}_j].$$

Теперь просуммируем все остаточные дисперсии.

$$\sum_{i=1}^{p}\sigma_i^2 = \sum_{i=1}^{p}\sigma_{ii} - \mathbf{b}_1^T(\sum_{i=1}^{p}\Sigma_i\Sigma_i^T)\mathbf{b}_1 - ... - \mathbf{b}_k^T(\sum_{i=1}^{p}\Sigma_i\Sigma_i^T)\mathbf{b}_k =$$

$$= \mathrm{Tr}(\Sigma) - \mathbf{b}_1^T(\Sigma\Sigma^T)\mathbf{b}_1 - ... - \mathbf{b}_k^T(\Sigma\Sigma^T)\mathbf{b}_k.$$

Для того чтобы минимизировать $\sum\sigma_i^2$ нужно найти максимальное значение суммы

$$\mathbf{b}_1^T(\Sigma\Sigma^T)\mathbf{b}_1 + ... + \mathbf{b}_k^T(\Sigma\Sigma^T)\mathbf{b}_k$$

при условиях

$$\mathbf{b}_i^T\Sigma\mathbf{b}_i = 1; \quad \mathbf{b}_i^T\Sigma\mathbf{b}_j = 0, \quad i \neq j,$$

которые обеспечивают, что случайные величины $\mathbf{b}_1^T\mathbf{u}, ..., \mathbf{b}_k^T\mathbf{u}$ не коррелированы и имеют единичную дисперсию. В таком случае при оптимальном выборе векторов $\mathbf{b}_i$ они являются собственными векторами для характеристического уравнения

$$\det(\Sigma\Sigma - \lambda\Sigma) = 0.$$

Но собственные числа и векторы такого уравнения совпадают с собственными числами и векторами уравнения

$$\det(\Sigma - \lambda I) = 0.$$

■

Интерпретация главных компонент

Как показано в [10], полученный результат дает возможность интерпретировать главные компоненты следующим образом. Предположим, что мы хотим заменить p-мерную случайную величину на $k < p$ линейных функций, теряя не слишком много информации. Эффективность выбора этих функций зависит от того, в какой степени они дают возможность реконструировать p первоначальных случайных величин. Один из методов реконструкции случайной величины u_i состоит построении ее наилучшего линейного предиктора на основе k линейных функций. В этом случае эффективность предиктора может быть измерена с помощью остаточной дисперсии σ_i^2. Полная мера эффективности предиктора равна $\sum\sigma_i^2$. *Наилучшим выбором линейных функций, для которых $\sum\sigma_i^2$ минимальна, является выбор первых k главных компонент случайной величины* $\boldsymbol{u}$.

Уточнение постановки задачи

Введем следующие обозначения:

$\hat{\Sigma} = A^TA$ — оценка ковариационной матрицы объектов;

$\mathbf{y}$ — вектор главных компонент;

W — весовая матрица (матрица преобразования вращения). Она является ортогональной, то есть $I = W^TW$;

$\mathbf{q}$ — интегральный индикатор.

Обозначим $Z = WA$ матрицу, состоящую из столбцов $(\mathbf{z}_1, ..., \mathbf{z}_p)$. Для нахождения первой главной компоненты необходимо найти такие линейные комбинации строк матрицы A, что векторы-столбцы матрицы Z обладали бы наибольшей дисперсией.

Требуется найти:

$$W_* = \arg\min_{W^T W = I} \sum_{i=1}^{p} \|\mathbf{a}_i - (\mathbf{a}_i, \mathbf{w})\mathbf{w}\|^2,$$

где $\mathbf{a}_i$ — векторы-строки матрицы «объекты-признаки» A, $\mathbf{w}$ — один из столбцов матрицы W, причем $\|\mathbf{w}\| = 1$.

Такая постановка задачи в силу доказанного С. Р. Рао [10] утверждения эквивалентна следующей:

$$W_* = \arg\max_{W^T W = I} \sum_{j=1}^{p} D\mathbf{z}_j.$$

При этом интегральный индикатор строится в виде:

$$\mathbf{q} = A\mathbf{w},$$

где $\mathbf{w}$ —один из столбцов матрицы W_*.

Описание алгоритма

Как было показано, для нахождения первой главной компоненты, используемой при построении интегрального индикатора, в качестве матрицы преобразования вращения нужно рассматривать матрицу, составленную из собственных векторов ковариационной матрицы вектора признаков:

$$W_* = P = \{\mathbf{p}_1, \mathbf{p}_2, ...\mathbf{p}_n\},$$

где $\mathbf{p}_1$, $\mathbf{p}_2$, ... $\mathbf{p}_n$ — собственные векторы ковариационной матрицы $\hat{\Sigma}$.

Направление первой главной компоненты определяет собственный вектор, соответсвующий максимальному собственному числу. Если $\lambda_1, \lambda_2, ...\lambda_n$ — собственные числа матрицы $\hat{\Sigma}$ и для них выполнено:

$$\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_n,$$

то искомый вектор $\mathbf{w}$ определяется как

$$\mathbf{w} = \mathbf{p}_1.$$

Метод расслоения Парето

В предположении, что признаки могут быть измерены в ранговых шкалах, используем для построения интегрального индикатора метод расслоения Парето.

Уточнение постановки задачи. Введем отношение доминирования на наборе объектов $\{\mathbf{a}_i\}_{i=1}^m$. Объект $\mathbf{a}_i$ доминирует объект $\mathbf{a}_k$ ($\mathbf{a}_i \succ \mathbf{a}_k$, $\mathbf{a}_i \neq \mathbf{a}_k$), если все компоненты вектора $\mathbf{a}_i$ больше или равны соответсвующим компонентам вектора $\mathbf{a}_k$:

$$a_{ij} \geqslant a_{kj},\ j = 1, ...n.$$

Определим Парето-оптимальный фронт как набор недоминируемых объектов.

Требуется разделить имеющийся набор объектов на Парето-слои из недоминируемых объектов.

Описание алгоритма. Рассмотрим строки матрицы «объекты-признаки» как набор сравниваемых объектов. Будем отсекать Парето-слои, начиная с нижнего. Обозначим P_l l-тый Парето-слой.

$$P_l = \{\mathbf{a} \in A \mid \neg\exists \mathbf{x} \in A : x_i \leqslant a_i, i = 1, ...n\}$$

Исключим из матрицы полученный слой перейдем к получению следующего. Процесс остановится, когда матрица A станет пустой.

Алгоритм Парето-расслоения описан в [12].

Вычислительный эксперимент

Метод главных компонент на модельных данных

В ходе эксперимента использовались синтетические данные (рис. 1), которые подбирались вручную так, чтобы точки, соответствующие описаниям объектов, находились на плоскости вблизи одной прямой.

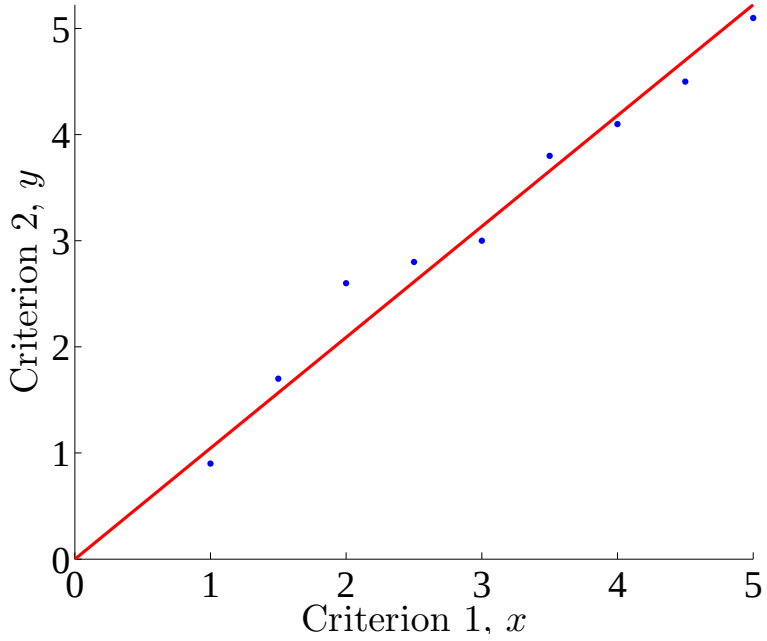

Рис. 1. Построение первой главной компоненты для синтетических данных. Синие точки соответствуют описаниям объектов, красная прямая показывает направление главной компоненты.

Метод главных компонент на реальных данных

В настоящей работе для построения интегрального индикатора для вузов предлагается разбивать биографии выпускников на группы, сооветвующие следующим сферам деятельности:

1. Образование, наука, инновационные разработки;
2. Бизнес, экономика;
3. Политика, государственное управление, деятельность в общественных организациях;
4. Культура и искусство;
5. Спортивная карьера.

В каждой группе будет построен интегральный индикатор методом главных компонент. Интегральный индикатор вуза будет определяться как среднее арифметическое из индикаторов его выпускников.

Эксперимент проводился для выборки в сфере деятельности «Бизнес, экономика». В выборку вошли описания биографий 30 богатейших бизнесменов России по версии жур-

нала «Forbes» за 2011 год. Для визуализации результатов использовались попарно 3 признака, оказавшиеся наиболее близкими к главной компоненте. На рисунке 2 представлены в виде точек входные данные, красные линии — направления первых главных компонент для пары признаков.

Так же при помощи вычисления коэффициента ранговой корреляции Спирмена между полученным интегральным индикатором и доходом бизнесменов была проверена гипотеза о том, что определяющую роль в индикаторе играет доход. Гипотеза не была подтверждена: значение коэффициента Спирмена равно 0.4794, что соответствует слабой прямой связи между величинами.

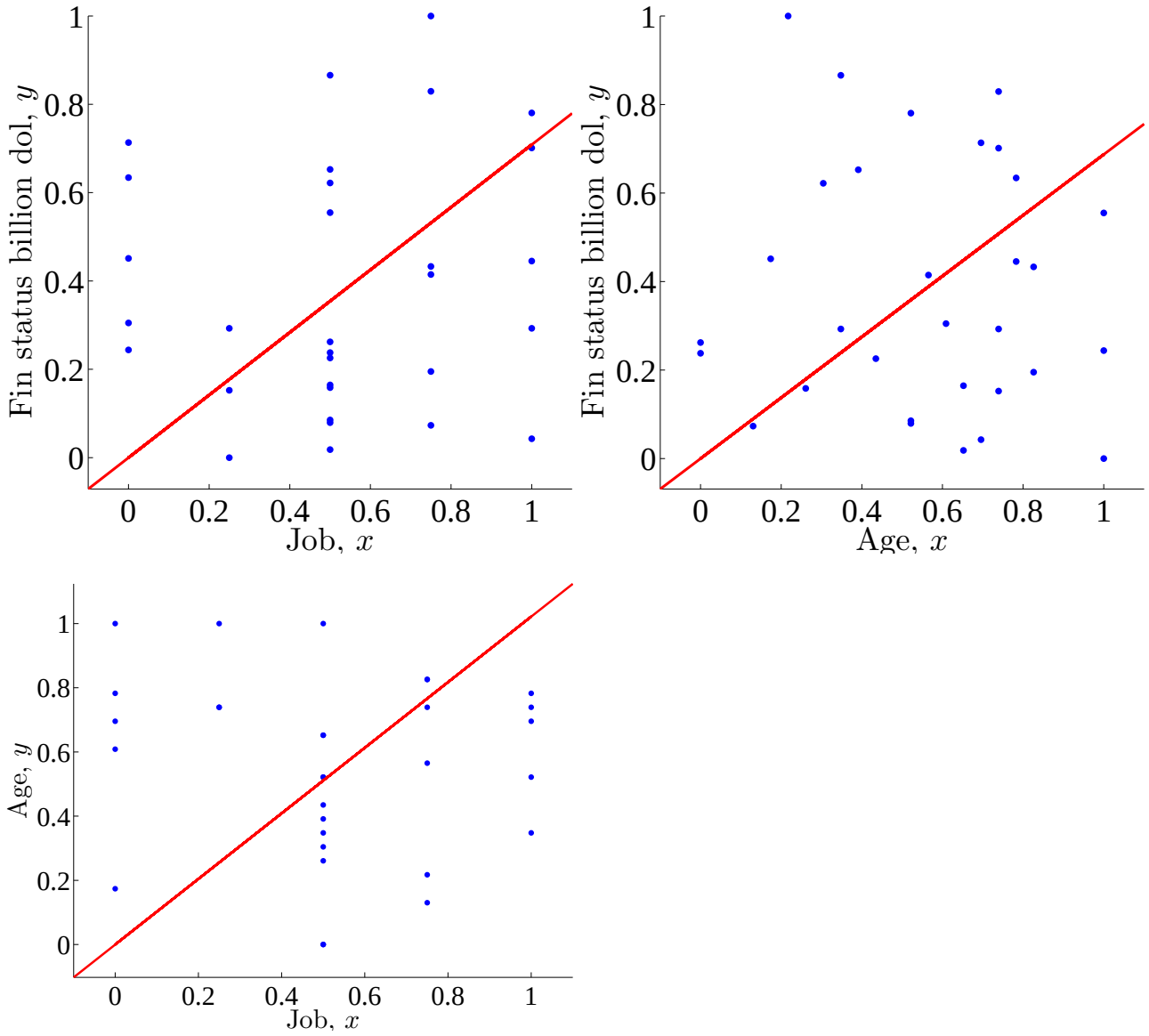

Рис. 2. Построение первой главной компоненты. Синие точки соответствуют описаниям объектов, красная прямая показывает направление главной компоненты. Оси нормированы.

Сравнение результатов для разных способов построения интегральных индикаторов

Для возможности визуализации интегральные индикаторы строились для трех признаков. Результаты, полученные с помощью метода главных компонент, представлены в том же формате, что и в предыдущем случае (рис. 3). Результаты, полученные при помощи расслоения Парето, представлены на рисунке 4, где различными цветами обозначены разные слои.

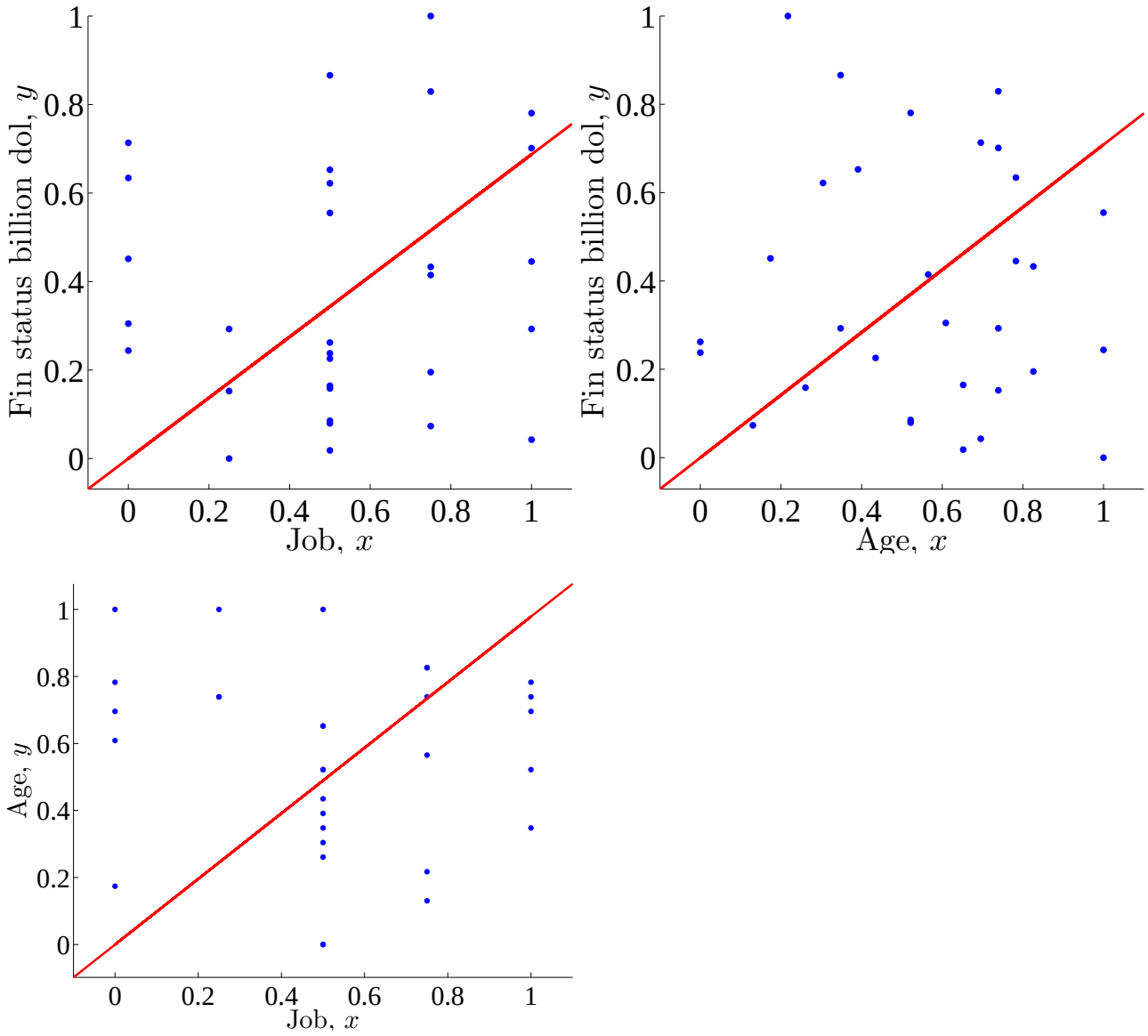

Рис. 3. Построение первой главной компоненты. Синие точки соответствуют описаниям объектов, красная прямая показывает направление главной компоненты. Оси нормированы.

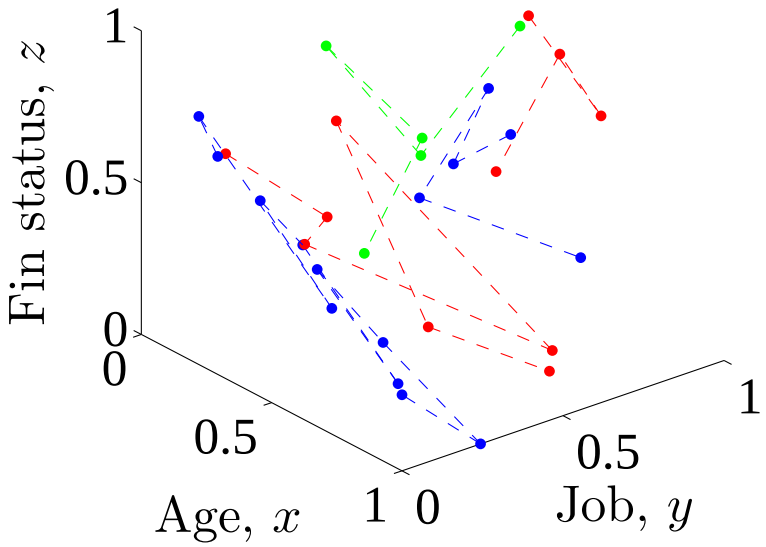

Рис. 4. Парето-слои. Оси нормированы.

Сравнение полученных результатов представлено на рисунке 5, где по горизонтальной оси отложены интегральные индикаторы объектов, полученные методом расслоения Парето, а по вертикальной оси — индикаторы, полученные методом главных компонент. Как можно видеть, результаты, даваемые разными методами, отличаются, но между ними прослеживается линейная зависимость.

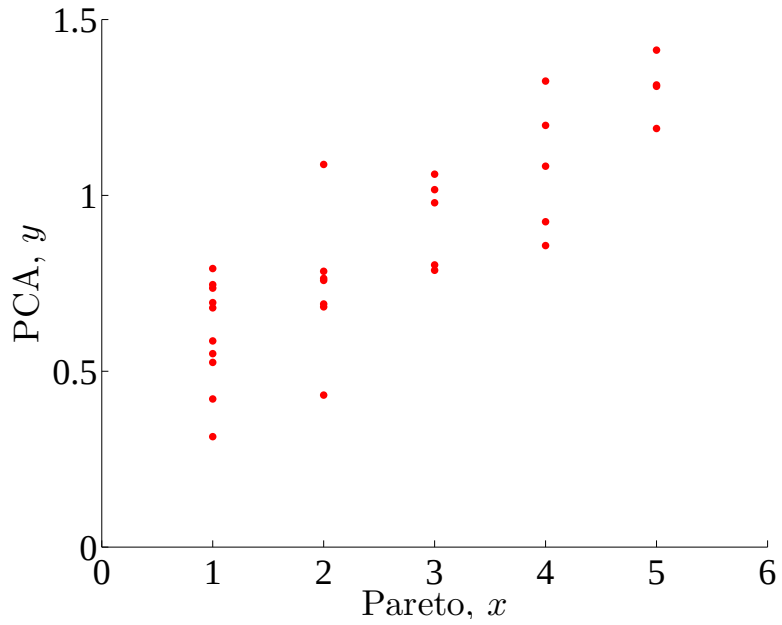

Рис. 5. Сравнение полученных результатов.

Заключение

В работе были приведены описание и теоретическое обоснование метода главных компонент, а также описание алгоритма расслоения Парето для построения интегральных индикаторов. Проведен вычислительный эксперимент и сравнение полученных результатов. Сделан вывод, что результаты, даваемые рассмотренными двумя методами связаны между собой линейно.

Литература

[1] О. М. Карпенко, М. Д. Бершадская. *Международный рейтинг университетов Webometrics: основные идеи, индикаторы, результаты*, Педагогические Измерения, 2010, №2.

[2] С.Ċ. Донецкая. *Российский подход к ранжированию ведущих университетов мира*, ЭКО, 2009, №9: 137-150

[3] С. А. Айвазян, В. М. Бухштабер, И. С. Енюков, Л. Д. Мешалкин. *Прикладная статистика. Классификация и снижение размерностей*, Финансы и статистика, 1989, с.334, 421-424.

[4] М. П. Кузнецов, В. В. Стрижов. *Уточнение ранговых экспертных оценок с использованием монотонной интерполяции*. Всероссийская конференция «Математические методы распознавания образов», сборник докладов. МАКС-Пресс, 2011.

[5] В. В. Стрижов. *Уточнение экспертных оценок с помощью измеряемых данных*, Заводская лаборатория. Диагностика материалов, 2006, с.59-64.

[6] Strijov, V. *Expert estimations concordance for biosystems under extreme conditions. Notes on applied mathematics*, Moscow, Coumpiting Center of RAS, 2002.

[7] Vadim Strijov and Goran Granic and Jeljko Juric and Branka Jelavic and Sandra Antecevic Maricic. *Integral indicator of ecological impact of the Croatian thermal power plants*, Energy, 2011, №7: 4144-4149.

[8] Strijov, V. and Shakin, V. *Index construction: the expert-statistical method*, Proc. Conference on Sustainability Indicators and Intelligent Decisions, 2003, 56-57.

[9] В. В. Стрижов, Т. В. Казакова. *Устойчивые интегральные индикаторы с выбором опорного множества описаний*, Заводская лаборатория. Диагностика материалов. 2007, 72-74.

[10] С. Р. Рао. *Линейные и статистические методы и их применения*, Наука, 1968, 530-533.

[11] I. T. Jolliffe *Principal Conmponent Analysis*, Springer, 2002.

[12] М. М. Медведникова. *Алгоритм Парето-расслоения*,
https://mlalgorithms.svn.sourceforge.net/svnroot/mlalgorithms/Medvednikova2012PCA/code/Pareto, 2012.