Снижение размерности пространства. Метод главных компонент

(principal components analysis, PCA)

Примеры практического использования метода главных компонент для снижения размерности пространства признаков

Снижение размерности пространства признаков

Зачем?

-наглядно представить исходные данные

-упростить исследуемую модель

-снизить объемы хранимой информации

Без потери информативности!

Легко снизить пространство при:

Дублировании информации (сильно взаимосвязанные показатели) – **исключаем из рассмотрения**

Наличии неинформативных переменных (переменных, практически не меняющихся при переходе от объекта к объекту) – **исключаем из рассмотрения**

Наличии однотипных переменных - агрегируем (или простое суммирование) однотипные переменные

Два способа снижения размерности

1 способ (удаляем неинформативные из исходного перечня данных)

Без видоизменения пространства исходных переменных

(корреляционный анализ)

2 способ (переходим к новому пространству, каждая переменная в новом пространстве – линейная комбинация исходных переменн

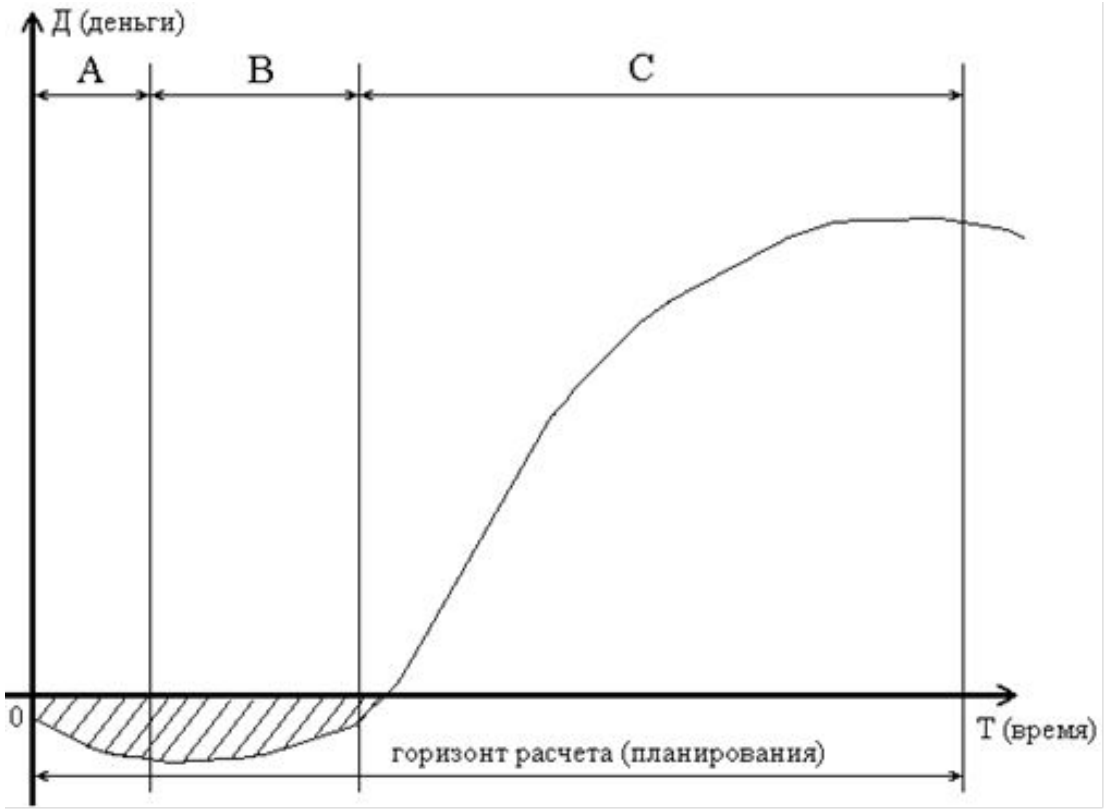
$$F_k = \sum_{j=1}^{m} a_{jk} X_j$$

С преобразованием пространства

(факторный анализ, метод главных компонент (МГК)) – *см. 4 задание в практической работе «Очистка данных»*

Пример снижения размерности

Жизненный цикл инвестиционного проекта (ИП)



Для оценки коммерческой эффективности инвестиционных проектов используются следующие показатели:

1) Доход на капитал
2) Срок окупаемости
3) Будущая стоимость проекта
4) NPV – чистая дисконтированная стоимость
5) IRR – внутренняя норма рентабельности
6) PI – индекс доходности
7) PBP – период возврата вложений
…

Смысловая нагрузка показателей

y1 – Доход на капитал - отношение среднегодовой прибыли от реализации проекта к первоначальным вложениям в проект (max)

y2 – Срок окупаемости проекта - период, требуемый для возврата первоначальных инвестиционных расходов посредством накопленных чистых потоков реальных денег, полученных с помощью проекта (min)

y3 – Будущая стоимость проекта - сумма чистых денежных потоков, связанных с реализацией проекта, за весь период его осуществления (max)

y4 – NPV - сумма текущих чистых денежных потоков за весь расчетный период, приведенная к начальному шагу расчета (max)

y5 – IRR - ставка дисконтирования, при которой NPV (чистая дисконтированная стоимость) за весь срок жизни инвестиционного проекта равна нулю (max)

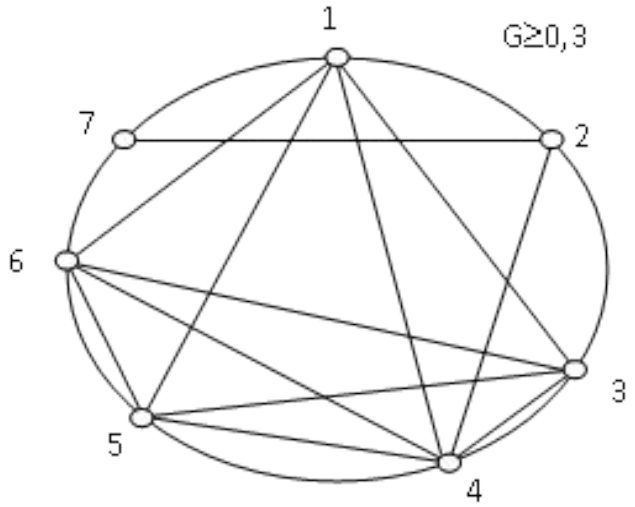
y6 – PI - отношение суммы дисконтированных чистых денежных потоков проекта к дисконтированной величине инвестиций (max)

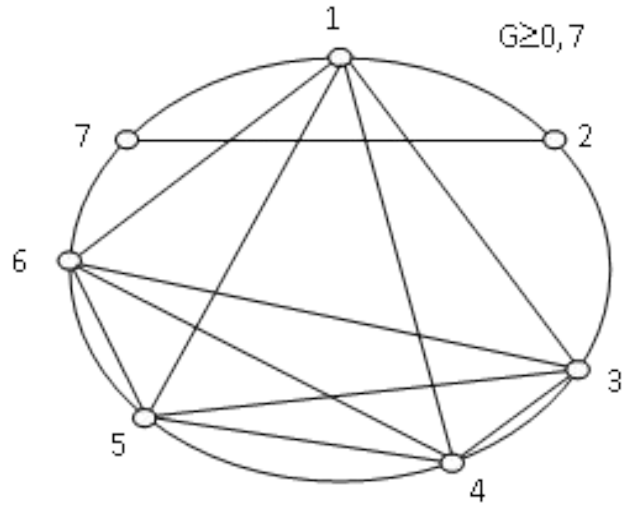
y7 – PBP - количество периодов, в течении которых дисконтированная денежная прибыль возмещает дисконтированную сумму капитальных вложений (min)

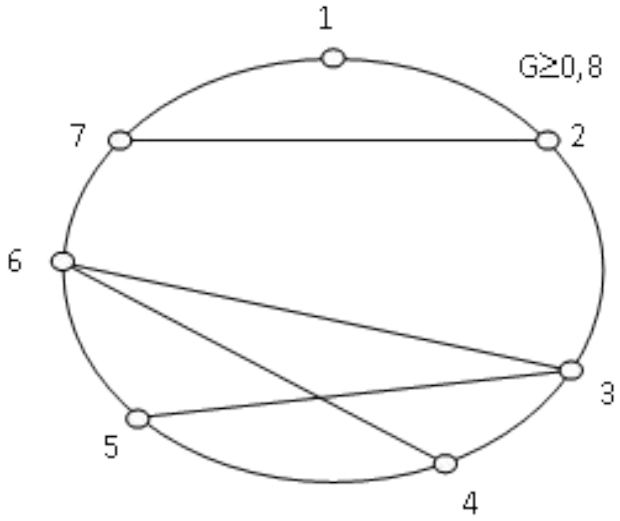
1 способ – корреляционный анализ

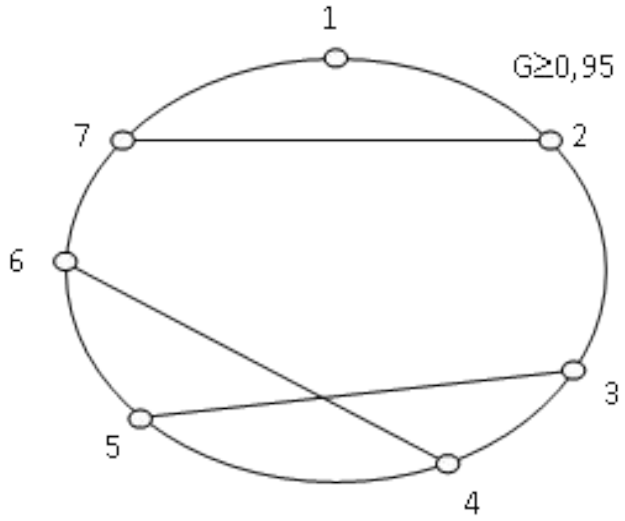
	Y1	y2	y3	y4	y5	y6	y7
y1	1						
y2	0,16	1					
y3	0,76	0	1				
y4	0,79	0,31	0,81	1			
y5	0,73	0,13	**0,95**	**0,82**	1		
y6	0,77	0,12	**0,93**	**0,96**	**0,88**	1	
y7	0,13	0,97	-0,09	0,2	0,09	0	1

1 способ – корреляционный анализ









1 способ – корреляционный анализ

Наиболее сильно коррелируют:

У2 и У7 (срок окупаемости и период возврата капитальных вложений);

У3 и У5 (будущая стоимость проекта и чистая дисконтированная стоимость NPV);

У4 и У6 (внутренняя норма прибыли IRR и индекс прибыльности PI).

Это значит, что можно без ущерба для качества принятия инвестиционного решения исключить из целевой функции три признака. Пусть это будут У2, У3, У6. Таким образом целевая функция вполне может быть описана следующими частными критериями:

1 – доход на капитал – У1;

2 – внутренняя норма прибыли IRR - У4;

3 – чистая дисконтированная стоимость NPV – У5;

4 – период возврата капитальных вложений PBP – У7.

Для углубленного исследования признакового пространства применяется Метод главных компонент – PCA – это 2 способ снижения размерности

НЕКОТОРЫЕ ПОНЯТИЯ ИЗ СТАТИСТИКИ

Средняя величина $\bar{x} = \sqrt[k]{\frac{\sum x_i^k}{n}}$

x_i - i-й вариант усредняемого признака, n - число вариантов, к - порядок средней.

Средняя арифметическая: (математическое ожидание) $\bar{x} = \frac{\sum x_i}{n}$

Среднее квадратическое отклонение признака: $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$

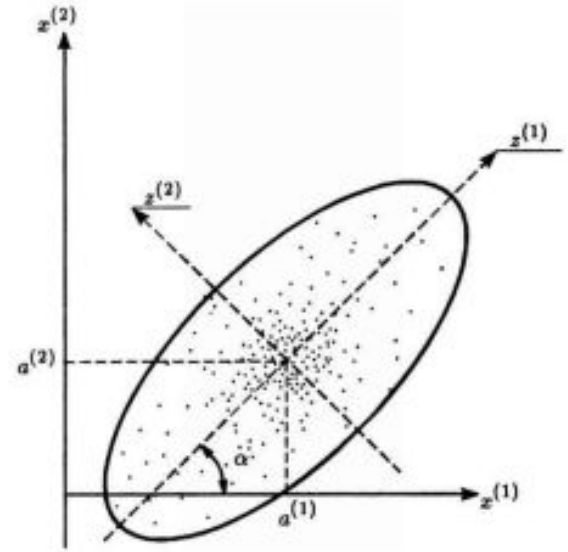
Дисперсия случайной величины : $D = \frac{\sum (x - \bar{x})^2}{n}$

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Метод главных компонент - это один из способов понижения размерности, состоящий в переходе к новому ортогональному базису, оси которого ориентированы по направлениям максимальной дисперсии набора входных данных. Вдоль первой оси нового базиса дисперсия максимальна, вторая ось максимизирует дисперсию при условии ортогональности первой оси, и т.д., последняя ось имеет минимальную дисперсию из всех возможных. Такое преобразование позволяет понижать информацию путем отбрасывания координат, соответствующих направлениям с минимальной дисперсией. Предполагается, что если нам надо отказаться от одного из базисных векторов, то лучше, если это будет тот вектор, вдоль которого набор входных данных меняется менее значительно.

В основе метода главных компонент лежат следующие допущения:

- Допущение о том, что размерность данных может быть понижена за счет линейного преобразования.
- Допущение о том, что больше всего информации несут те направления, в которых дисперсия входных данных максимальна.



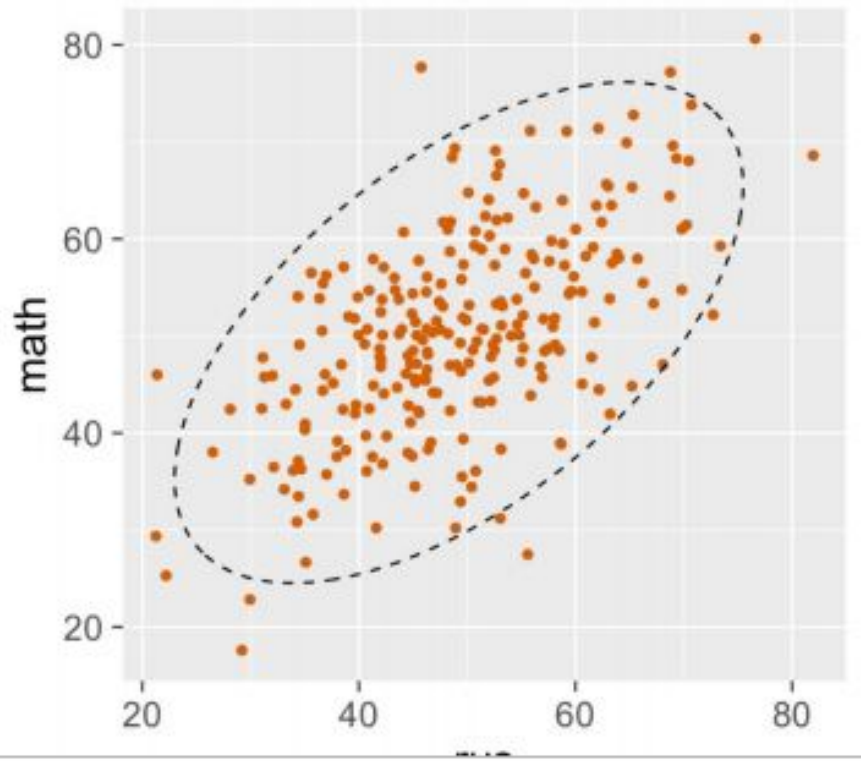
МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Простой пример
Пусть у нас есть список школьников с оценками по русскому языку и математике:

```
##        rus     math
## 1 38.62011 33.67848
## 2 46.22913 54.53733
## 3 46.40963 38.32976
## 4 53.17011 51.07601
## 5 62.86754 65.64322
```

Эти оценки можно визуализировать на плоскости:

Оценки по математике и русскому языку могут коррелировать между собой.



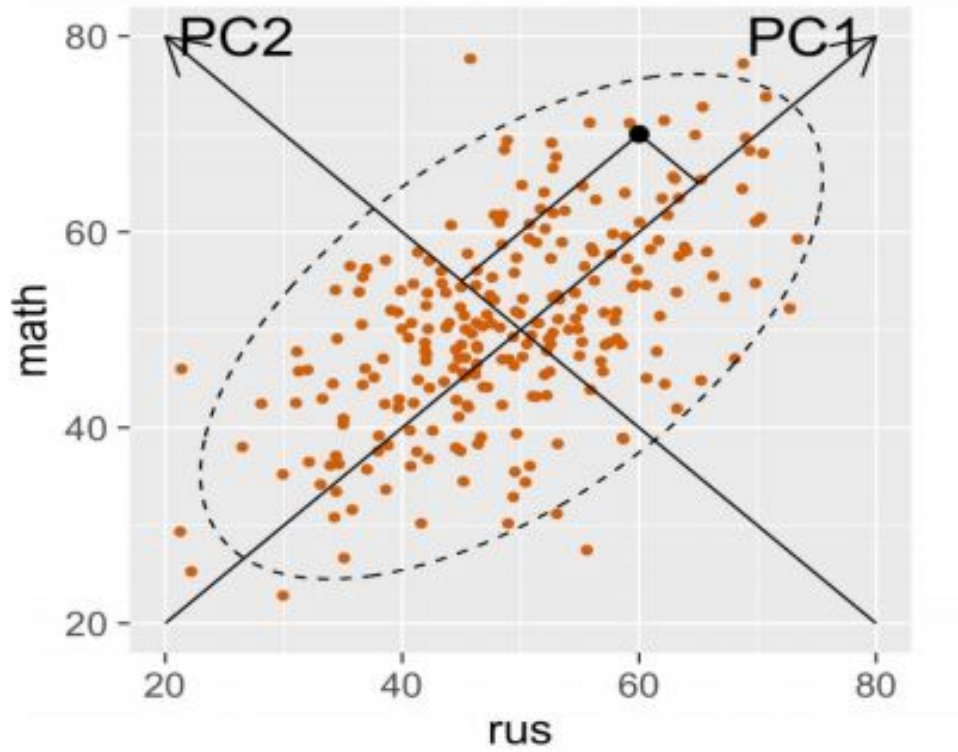
МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Мы можем ввести новые оси на этой же плоскости как линейные комбинации старых осей.

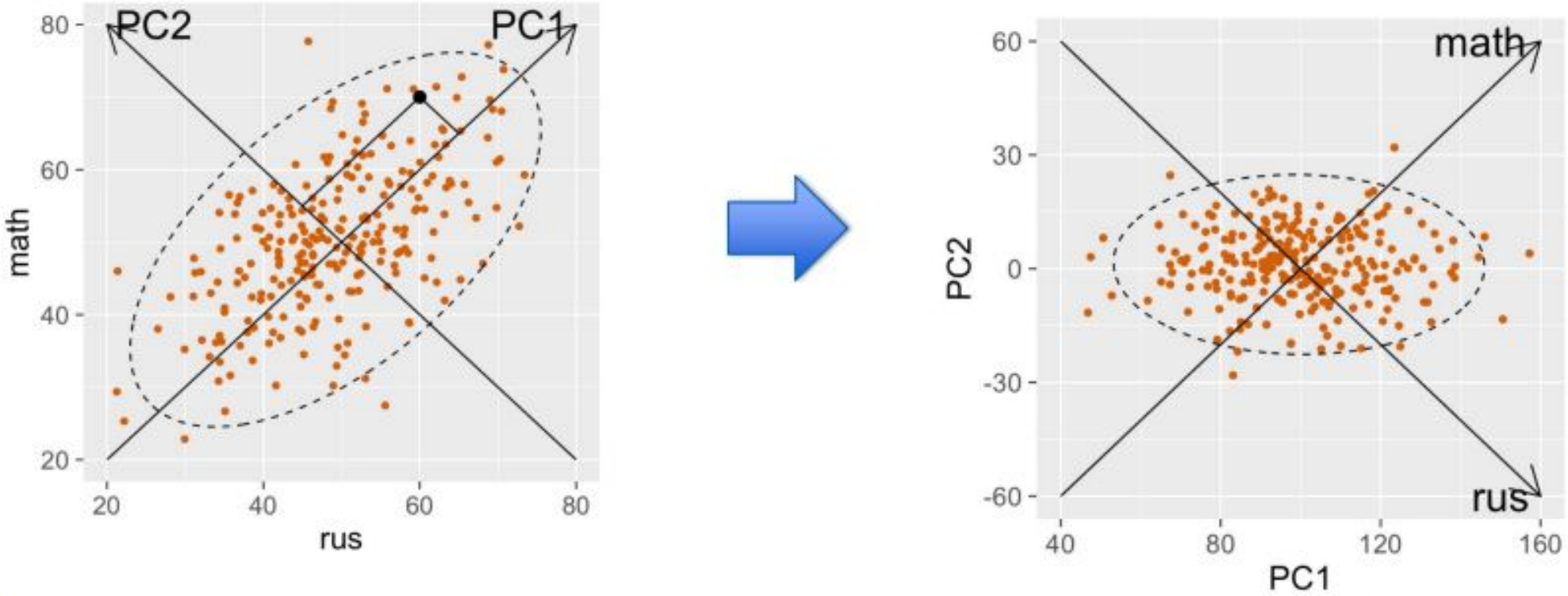
$$PC_1 = rus + math$$

$$PC_2 = math - rus$$

В этом случае наши оси будут выглядеть вот так:
Новые оси, так же перпендикулярны друг другу. Однако первая ось проходит через область максимального разброса оценок, вторая ось проходит через оставшуюся часть максимального разброса оценок.



МЕТОД ГЛАВНЫХ КОМПОНЕНТ



В рамках старых осей, мы могли судить про корреляцию между оценками math и rus, то в новых осях это сделать уже сложно. Координата PC_1 называется ***первой главной компонентой***, а PC_2 — ***второй***. Заметим, что «главных компонент» получилось столько же, сколько изначально было переменных, но, и что PC_1 «главнее» (содержит больше информации), чем PC_2. **Таким образом, идея метода заключается в преобразовании старых компонент в новые, при этом происходит ранжирования новых компонент по степени объяснения дисперсии.**

АЛГОРИТМ МОДЕЛИ

$f_1(x), \ldots, f_n(x)$ — исходные числовые признаки;
$g_1(x), \ldots, g_m(x)$ — новые числовые признаки, $m \leqslant n$;

Требование: старые признаки должны линейно восстанавливаться по новым:

$$\hat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, \quad j = 1, \ldots, n, \quad \forall x \in X,$$

как можно точнее на обучающей выборке $x_1, \ldots, x_\ell$:

$$\sum_{i=1}^{\ell} \sum_{j=1}^{n} \left(\hat{f}_j(x_i) - f_j(x_i)\right)^2 \to \min_{\{g_s(x_i)\}, \{u_{js}\}}$$

АЛГОРИТМ МОДЕЛИ

Матрицы «объекты–признаки», старая и новая:

$$\underset{\ell\times n}{F} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}; \quad \underset{\ell\times m}{G} = \begin{pmatrix} g_1(x_1) & \dots & g_m(x_1) \\ \dots & \dots & \dots \\ g_1(x_\ell) & \dots & g_m(x_\ell) \end{pmatrix}.$$

Матрица линейного преобразования новых признаков в старые:

$$\underset{n\times m}{U} = \begin{pmatrix} u_{11} & \dots & u_{1m} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nm} \end{pmatrix}; \qquad \hat{F} = GU^{\mathsf{T}} \overset{\text{хотим}}{\approx} F.$$

Найти: и новые признаки G, и преобразование U:

$$\sum_{i=1}^{\ell}\sum_{j=1}^{n}\left(\hat{f}_j(x_i) - f_j(x_i)\right)^2 = \left\|GU^{\mathsf{T}} - F\right\|^2 \to \min_{G,U},$$

АЛГОРИТМ МОДЕЛИ

$$\sum_{i=1}^{\ell}\sum_{j=1}^{n}\left(\hat{f}_j(x_i) - f_j(x_i)\right)^2 = \left\|GU^{\intercal} - F\right\|^2 \to \min_{G,U},$$

Можно показать, что если вычислить собственные вектора матрицы F, потом составить матрицу U из этих векторов у которых максимальные собственные значения. Далее рассчитать матрицу G=FU, то тогда мы получим минимум. И таким образом получим решение задачи.

Подробности можно найти в видеолекции Воронцова К.

https://www.youtube.com/watch?v=wcJ0nSUr7ws

или вот тут
http://www.machinelearning.ru/wiki/images/archive/a/a2/20150509140209%21Voron-ML-regression-slides.pdf

МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Проблема интерпретации новых осей.
В рассмотренном примере старые оси имели смысл оценок по двум разным предметам. Новые оси это разности и суммы оценок.

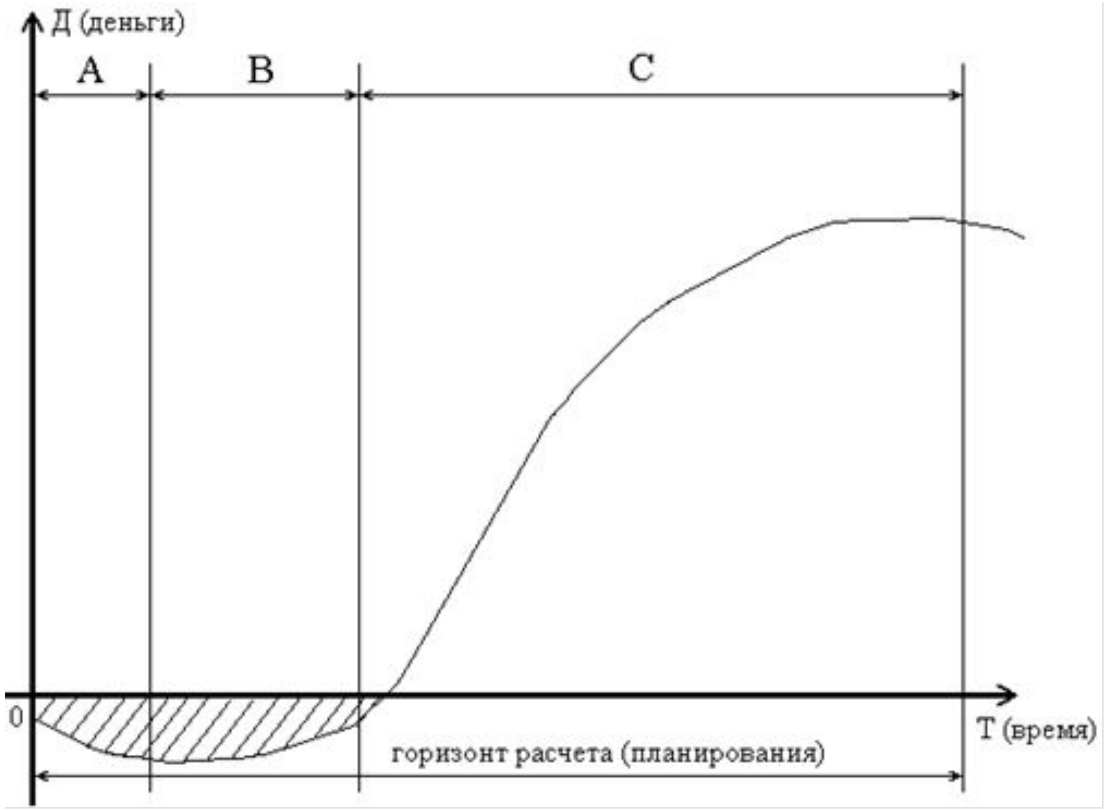
В случае множества осей, новые оси выражаются через линейную комбинацию старых осей, где вклад каждой оси определяется коэффициентом.

X – «старые объекты-признаки», G – « новые объекты-признаки»,
U – матрица перехода.

Таким образом, новые оси представляют собой смесь старых факторов (осей), соответственно, интерпретация новых осей вызывает проблему.

Пример практического применения PCA

Жизненный цикл инвестиционного проекта (ИП)



Для оценки коммерческой эффективности инвестиционных проектов используются следующие показатели:

1) Доход на капитал
2) Срок окупаемости
3) Будущая стоимость проекта
4) NPV – чистая дисконтированная стоимость
5) IRR – внутренняя норма рентабельности
6) PI – индекс доходности
7) PBP – период возврата вложений
…

Смысловая нагрузка показателей

y1 – Доход на капитал - отношение среднегодовой прибыли от реализации проекта к первоначальным вложениям в проект (max)

y2 – Срок окупаемости проекта - период, требуемый для возврата первоначальных инвестиционных расходов посредством накопленных чистых потоков реальных денег, полученных с помощью проекта (min)

y3 – Будущая стоимость проекта - сумма чистых денежных потоков, связанных с реализацией проекта, за весь период его осуществления (max)

y4 – NPV - сумма текущих чистых денежных потоков за весь расчетный период, приведенная к начальному шагу расчета (max)

y5 – IRR - ставка дисконтирования, при которой NPV (чистая дисконтированная стоимость) за весь срок жизни инвестиционного проекта равна нулю (max)

y6 – PI - отношение суммы дисконтированных чистых денежных потоков проекта к дисконтированной величине инвестиций (max)

y7 – PBP - количество периодов, в течении которых дисконтированная денежная прибыль возмещает дисконтированную сумму капитальных вложений (min)

Исходные данные - 7-мерное пространство

№ проекта	Доход на капитал, %	Срок окуп, год	Будущая ст.-сть проекта, тыс.у.е	IRR, %	NPV, тыс.у.е	PI	PBP, год
1	46,6	3,5	5565,9	24	497,4	1,04	5,1
2	37,3	3,02	14293	26	8927	1,51	3,5
3	57,1	1,92	3313,1	26	184,9	1,13	2,3
4	13	6,9	433	3,9	-7125	0,17	9,1
5	17,4	5,6	18000	12,4	1127,2	0,96	7,2
6	38,8	4,58	11959,1	21,4	-11117	0,24	7
7	22	2,5	573,85	27,5	314,53	1,62	2,8
8	120	1,6	133281	115	59723	7,79	2,3
…				…			…

Основная идея PCA (на примере)

От исходного 7-мерного пространства (y1, y2, y3, y4, y5, y6, y7)

переходим к новому пространству - тоже 7-мерному (P1, P2, P3, P4, P5, P6, P7) – это новая ортогональная 7-мерная система координат.

Каждый показатель в новом пространстве (компонента, фактор) – линейная комбинация всех показателей исходного пространства:

$$P1=a11*y1+a12*y2+a13*y3+a14*y4+a15*y5+a16*y6+a17*y7$$

$$P2=a21*y1+a22*y2+a23*y3+a24*y4+a25*y5+a26*y6+a27*y7$$

…

$$P7=a71*y1+a72*y2+a73*y3+a74*y4+a75*y5+a76*y6+a77*y7$$

Постановка задачи

Необходимо описать набор критериев числом главных компонент m<<7 , обеспечивающих долю дисперсии 0,85 и сформировать интегральный показатель на основе матрицы весовых коэффициентов, учитывающих тесноту связи между исходными показателями и главными компонентами.

$$\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ y_{21} & y_{22} & \cdots & y_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nN} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nN} \end{pmatrix} \cdot \begin{pmatrix} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ f_{n1} & f_{n2} & \cdots & f_{nN} \end{pmatrix}$$

$$y_{ji} = a_{j1} f_{1i} + a_{j2} f_{2i} + \ldots + a_{jm} f_{mi}$$

Вклад каждой компоненты неодинаков

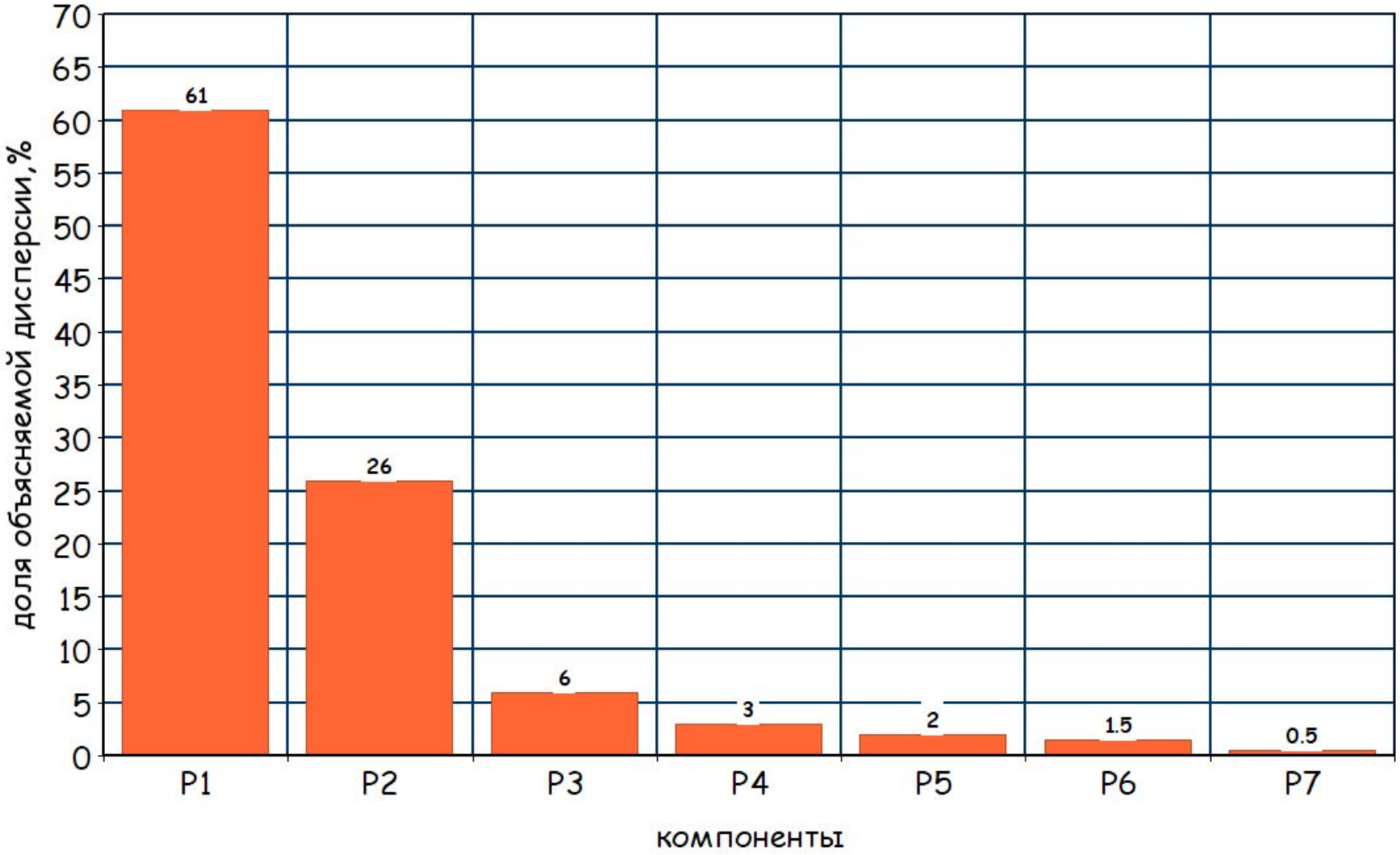
Предполагается, что исходные показатели y1-y7 полностью объясняют дисперсию результирующих факторов P1-P7.

Понижение размерности получается путем отбрасывания компонент P, в наименьшей степени объясняющих дисперсию всех исходных значений.

$$\sigma_j^2 = a_{j1}^2 + a_{j2}^2 + \ldots + a_{jn}^2 = 1$$

Слева записана полная дисперсия, а справа – доли полной дисперсии, относящиеся к соответствующим главным компонентам. Дисперсия является характеристикой изменчивости случайной величины, ее отклонений от среднего значения. Полный вклад каждого фактора в дисперсию всех исходных признаков определяет ту долю общей дисперсии, которую данная главная компонента объясняет.

Вклад каждой компоненты неодинаков



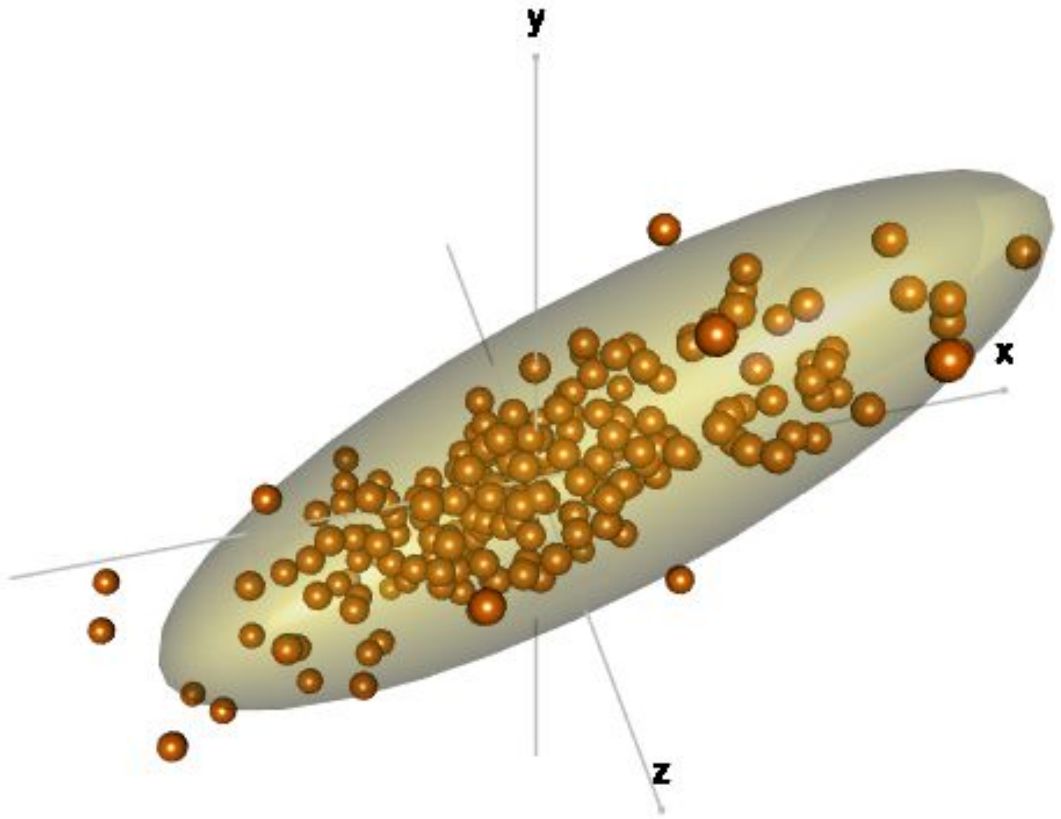
Принцип выбора координатных осей в новом пространстве

В качестве первой главной компоненты избирают направление, вдоль которого массив данных имеет наибольший разброс.

Выбор каждой главной последующей компоненты происходит так, чтобы разброс данных вдоль нее был максимальным, и чтобы эта главная компонента была ортогональна другим главным компонентам, выбранным прежде.

В результате получаем несколько главных компонент, каждая следующая из которых несет все меньше информации из исходного набора.

В качестве первой главной компоненты нужно выбрать такую координату, чтобы соответствующая координатная ось была направлена вдоль того направления, вдоль которого разброс точек самый боль



Результаты использования МГК

Исходные признаки	1 компонента	2 компонента	3 компонента	…	7 компонента
y1	0,858186	0,108055	- 0,278043		0,104445
y2	-0,096511	0,986718	0,610732		0,582818
y3	0,965640	-0,107920	0,804925		-0,148920
y4	0,915716	0,223502	0,283927		0,247292
y5	0,942211	0,046976	0,119485		0,283749
y6	0,966173	0,004762	0,859439		-0,987281
y7	-0,010893	0,990660	0,493820		0,114356
дисперсия, %	61,5	24,1	5,1		0,0031
Суммарная дисперсия, %	61,5	85,6	90,7		100

1 компонента – индекс доходности P1

P1=0,858*y1-*0,096*y2***+0,965*y3+0,915*y4+0,942*y5+0,966*y6-***0,011*y7***

2 компонента – индекс возвратности вложенных средств P2

P2 =*0,108*y1***+0,987*y2-***0,108*y3+0,224*y4+0,047*y5+0,005*y6***+0,991*y7**

Результаты использования МГК

Исходные признаки	1 компонента	2 компонента
y1	0,858186	0,108055
y2	-0,096511	0,986718
y3	0,965640	-0,107920
y4	0,915716	0,223502
y5	0,942211	0,046976
y6	0,966173	0,004762
y7	-0,010893	0,990660
дисперсия,%	61,5	24,1
Суммарная дисперсия,%	85,6	

1 компонента – индекс доходности P1

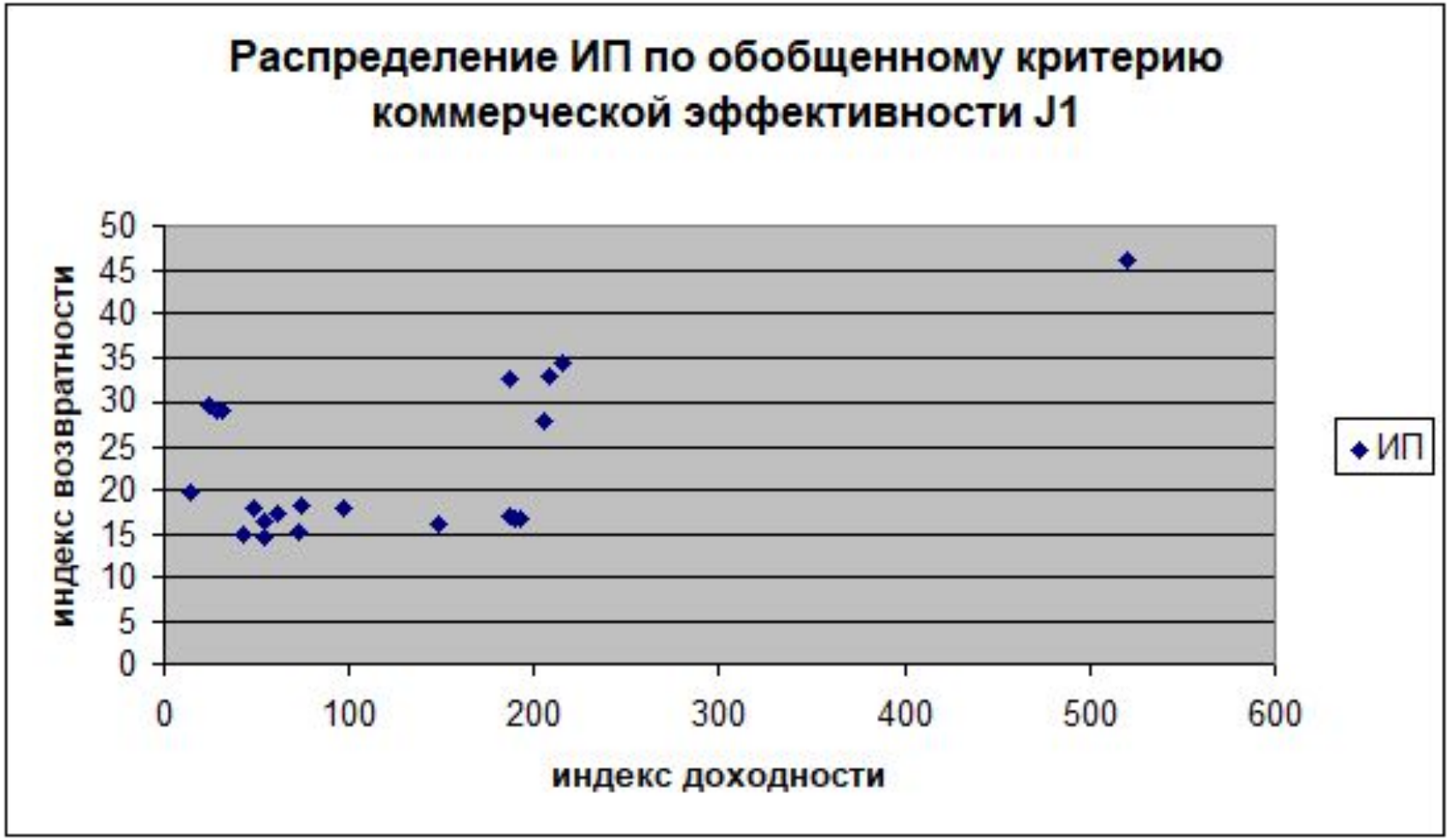
P1=0,858*y1-*0,096*y2***+0,965*y3+0,915*y4+0,942*y5+0,966*y6-***0,011*y7***

2 компонента – индекс возвратности вложенных средств P2

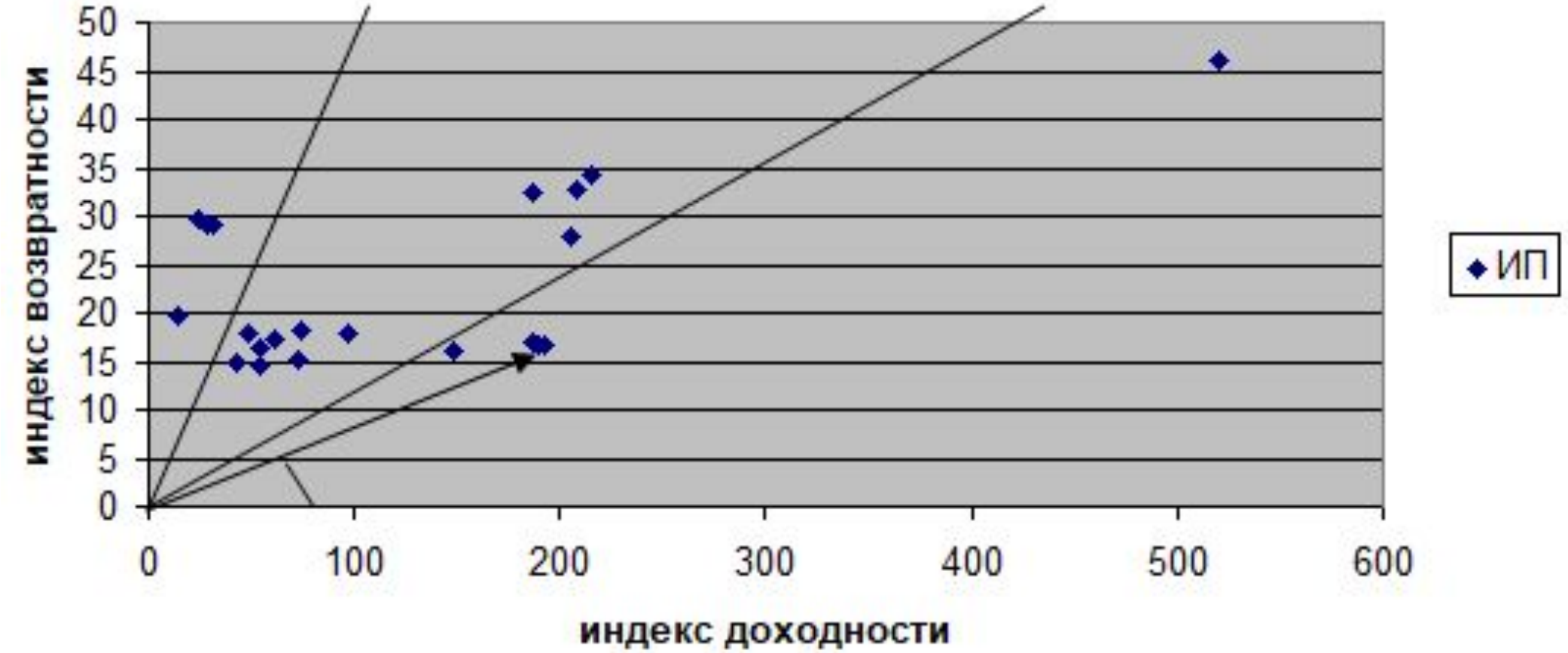
P2 =*0,108*y1***+0,987*y2-***0,108*y3+0,224*y4+0,047*y5+0,005*y6***+0,991*y7**

Таким образом, исходное 7-мерное пространство y1-y7 может быть сведено к 2-мерному ортогональному пространству главных компонент P1-P2 без существенной потери информативности.

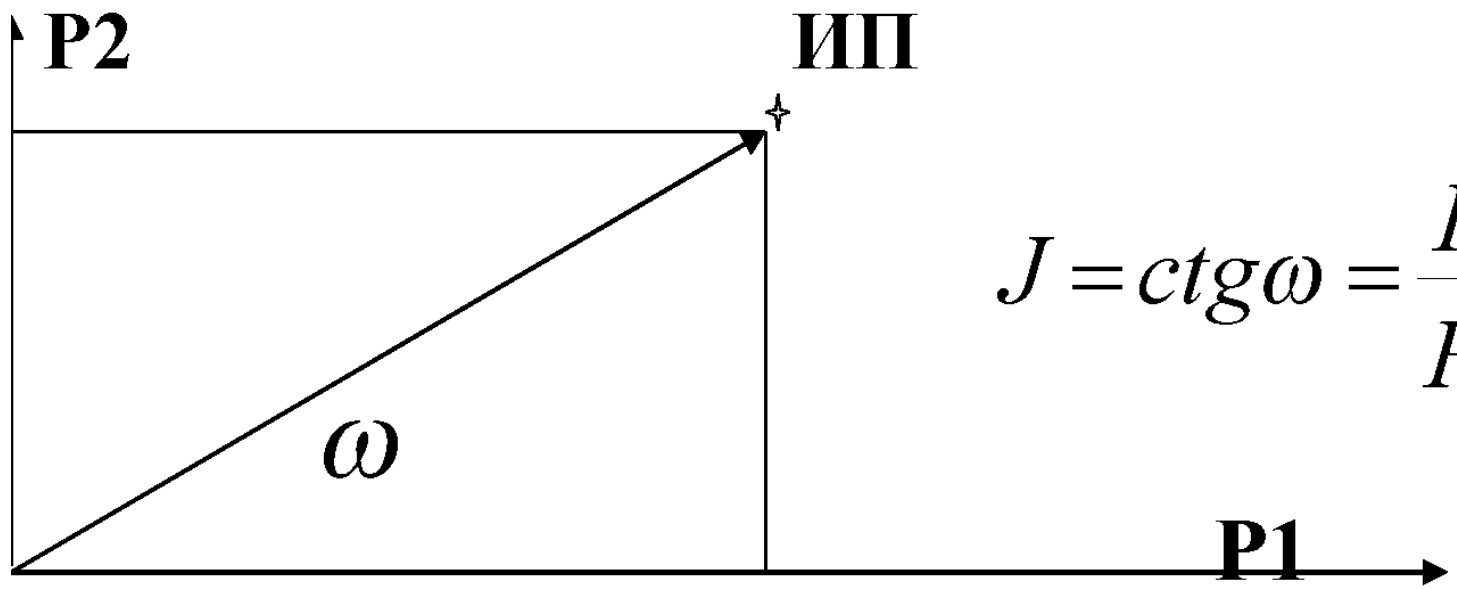
ИП в пространстве двух главных компонент (фрагмент, 20 проектов)



Выделение зон коммерческой эффективности ИП в пространстве двух главных компонент



От двух компонент – к одному обобщенному показателю



$$J = ctg\omega = \frac{P1}{P2}$$

Еще пример – применение МГК для классификации банков

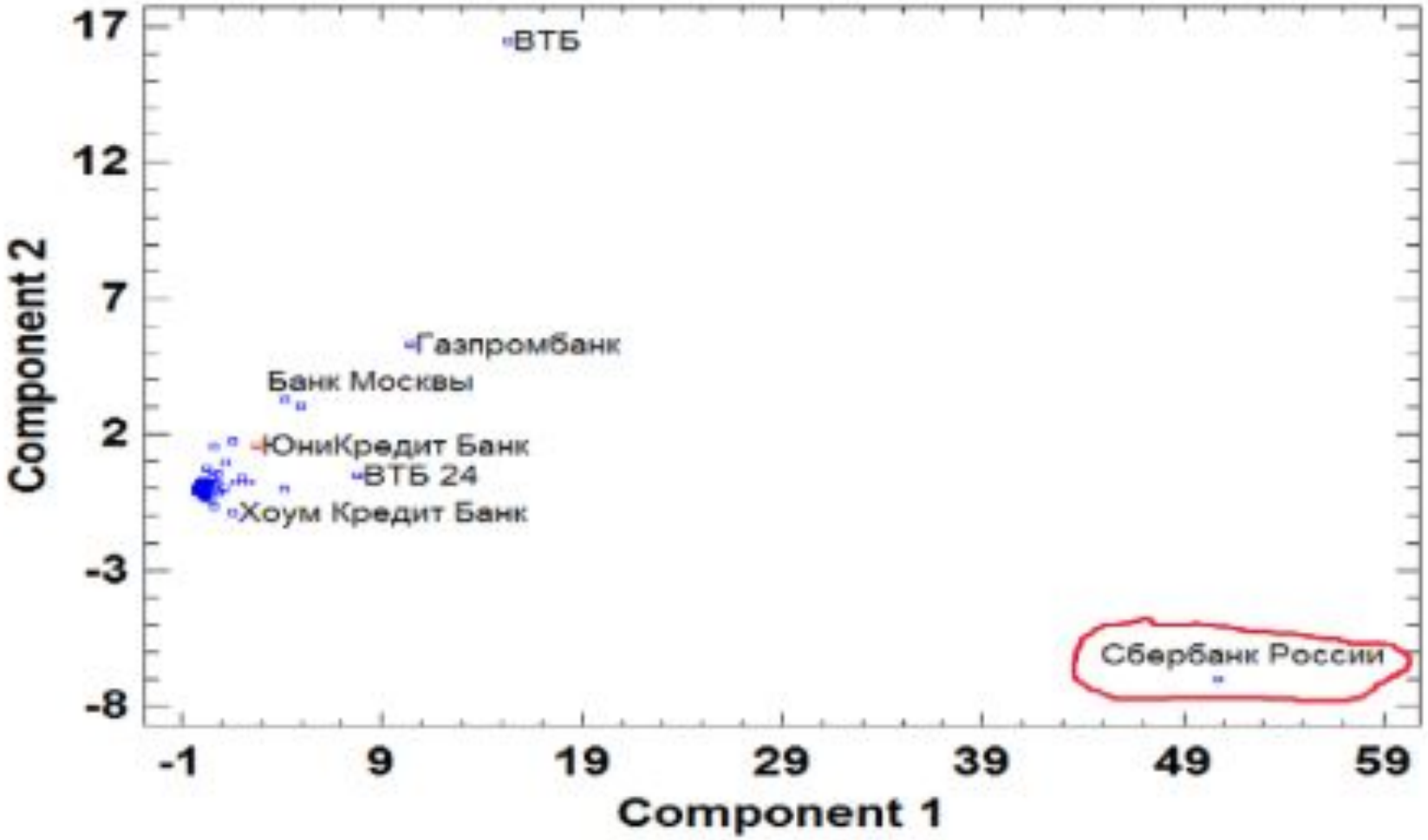
Рассматривалась выборка из 600 коммерческих банков, каждый их которых характеризуется следующими признаками:

- кредиты физическим лицам;
- кредиты предприятиям и организациям;
- вклады физических лиц;
- средства предприятий и организаций;
- чистая прибыль;
- выданные межбанковские кредиты.

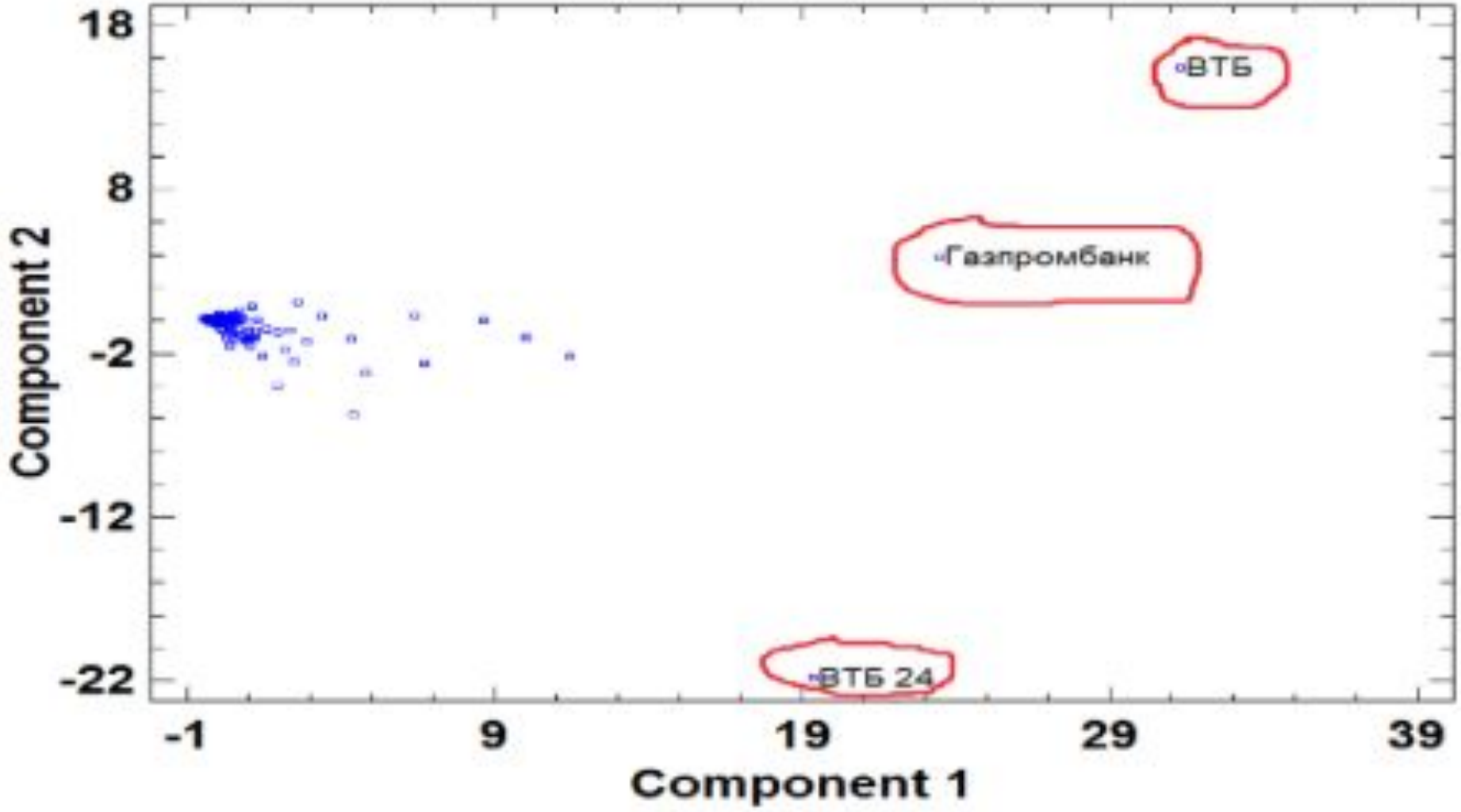
В таблицы – весовые коэффициенты исходных признаков по результатам использования МГК:

Признак	Весовые коэффициенты признака	
	Компонента 1	Компонента 2
Кредиты физическим лицам	0,411412	-0,366944
Кредиты предприятиям и организациям	0,434981	0,0359021
Вклады физических лиц	0,414246	-0,405891
Средства предприятий и организаций	0,408291	0,354798
Чистая прибыль	0,42905	-0,214544
Выданные межбанковские кредиты (МБК)	0,345176	0,726226

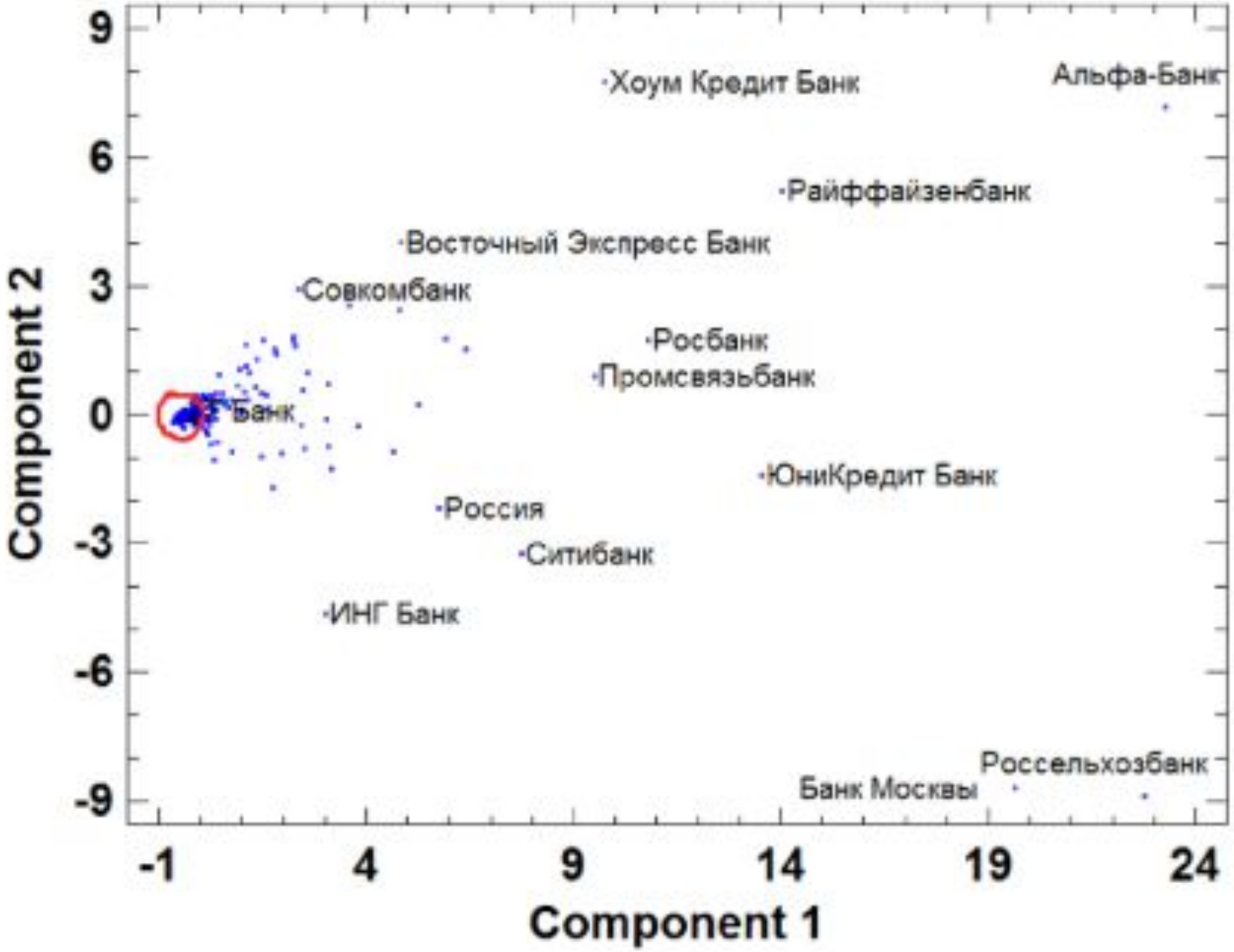
Распределение банков в пространстве двух главных компонент (600 банков)



Распределение банков в пространстве двух главных компонент (599 банков)



Распределение банков в пространстве двух главных компонент (596 банков)



Распределение банков в пространстве двух главных компонент (100 банков)