МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
Государственное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

Кафедра математических методов и моделей в экономике

А.Г. РЕННЕР, О.С. БРАВИЧЕВА

СНИЖЕНИЕ РАЗМЕРНОСТИ ПРИЗНАКОВОГО ПРОСТРАНСТВА

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНОМУ ПРАКТИКУМУ И САМОСТОЯТЕЛЬНОЙ РАБОТЕ СТУДЕНТОВ

Рекомендовано к изданию Редакционно-издательским советом
государственного образовательного учреждения высшего профессионального
образования «Оренбургский государственный университет»

Оренбург 2005

УДК 519.2 (07)
ББК 22.172 я7
Р 39

Рецензент
кандидат экономических наук, доцент С.В. Дьяконова

Реннер А.Г., Бравичева О.С.
Р 39 **Снижение размерности признакового пространства [Текст]: методические указания к лабораторному практикуму и самостоятельной работе студентов / А.Г. Реннер, О.С. Бравичева.– Оренбург: ГОУ ОГУ, 2005. – 29 с.**

Методические указания предназначены для выполнения лабораторных работ и самостоятельной работы студентов специальностей 061800, 061700, 010200 по дисциплине «Многомерные статистические методы» или «Многомерный статистический анализ» на темы «Компонентный анализ», «Метод главных факторов».

ББК 22.172 я7

Содержание

Введение

В случае, когда число признаков, характеризующих объекты исследования, достаточно велико естественно желание исследователя снизить размерность признакового пространства с целью:

а) получения наглядного представления (визуализации) исходных данных;

б) существенного сжатия объема хранимой статистической информации о свойствах объектов исследования, но без существенных потерь информации об объектах.

Принципиальная возможность существенного уменьшения числа признаков, характеризующих объекты исследования, без существенных потерь информации об объектах обусловлена малой вариативностью отдельных признаков или их линейных комбинаций.

Геометрическая суть рассматриваемых в лабораторной работе методе главных компонент и методе главных факторов состоит в линейном преобразовании исходной системы координат (показателей) к такой новой, в которой объекты будут характеризоваться существенно меньшим числом координат (латентных показателей).

Целью лабораторных работ по методам снижения размерности признакового пространства является выработка у студентов навыков статистического исследования и практической реализации алгоритмов метода главных компонент и метода главных факторов в пакете Statistica 6.0.

1 Содержание лабораторной работы

Лабораторная работа включает следующие этапы:

– постановку задачи;

– ознакомление с порядком решения задачи в пакетах прикладных программ;

– выполнение расчетов на компьютере;

– анализ результатов;

– подготовку письменного отчета по лабораторной работе;

– защиту лабораторной работы.

2 Постановка задачи

Исходные данные: выборочные данные по 53 предприятиям машиностроительного комплекса, характеризующимся пятью показателями производственно-хозяйственной деятельности /1/.

На основе выборочных данных из генеральной совокупности $\vec{X} = (X_1, X_2, X_3, X_4, X_5)^T$:

1) с помощью компонентного анализа и метода главных факторов снизить размерность признакового пространства, обеспечив уровень информативности не менее 70%;

2) при необходимости провести вращение пространства новых факторов;

3) дать экономическую интерпретацию факторам;

4) найти матрицу индивидуальных значений факторов.

3 Порядок выполнения работы

Порядок выполнения лабораторной работы рассмотрен на основании данных нулевого варианта таблиц А.1, А.2.

Поскольку исходные признаки отличаются масштабом (имеют разные единицы измерения), то перейдем к центрировано-нормированным признакам и, как следствие, в дальнейшем будем работать с матрицей парных коэффициентов корреляции. Оценка матрицы парных коэффициентов корреляции имеет вид:

$$\hat{R} = \begin{pmatrix} 1 & -0{,}49 & -0{,}29 & -0{,}53 & -0{,}63 \\ -0{,}49 & 1 & -0{,}2 & 0{,}23 & 0{,}22 \\ -0{,}29 & -0{,}2 & 1 & 0{,}32 & 0{,}38 \\ -0{,}53 & 0{,}23 & 0{,}32 & 1 & 0{,}79 \\ -0{,}63 & 0{,}22 & 0{,}38 & 0{,}79 & 1 \end{pmatrix}.$$

Перед проведением компонентного и факторного анализа, предполагая, что выборка произведена из нормально распределенной генеральной совокупности, на уровне значимости $\alpha = 0{,}05$ проверим гипотезу о незначимости (о диагональности) матрицы парных коэффициентов корреляции /2/, /3/.

$H_0 : R = E_n$ (корреляционная матрица диагональна),

$H_1 : R \neq E_n$ (корреляционная матрица отлична от диагональной).

Для проверки нулевой гипотезы используется статистика:

$$\chi^2 = -\left(N - \frac{1}{6}(2n+11) \right) \ln \left| \hat{R} \right|, \qquad (1)$$

где N – объем выборки;

n – число исходных признаков;

$\hat{R}$ – оценка матрицы парных коэффициентов корреляции;

$\left| \hat{R} \right|$ – определитель матрицы $\hat{R}$, равный произведению оценок собственных чисел матрицы $\hat{R}$: $\left| \hat{R} \right| = \hat{\lambda}_1 \cdot \hat{\lambda}_2 \cdot \ldots \cdot \hat{\lambda}_n$.

Статистика (1) при справедливости гипотезы H_0 имеет распределение «Хи-квадрат» с числом степеней свободы $\nu = \frac{n(n-1)}{2}$. Для нахождения наблюденного значения статистики (1) воспользуемся математическим пакетом Mathcad 2001. Порядок расчетов представлен на рисунке 1.

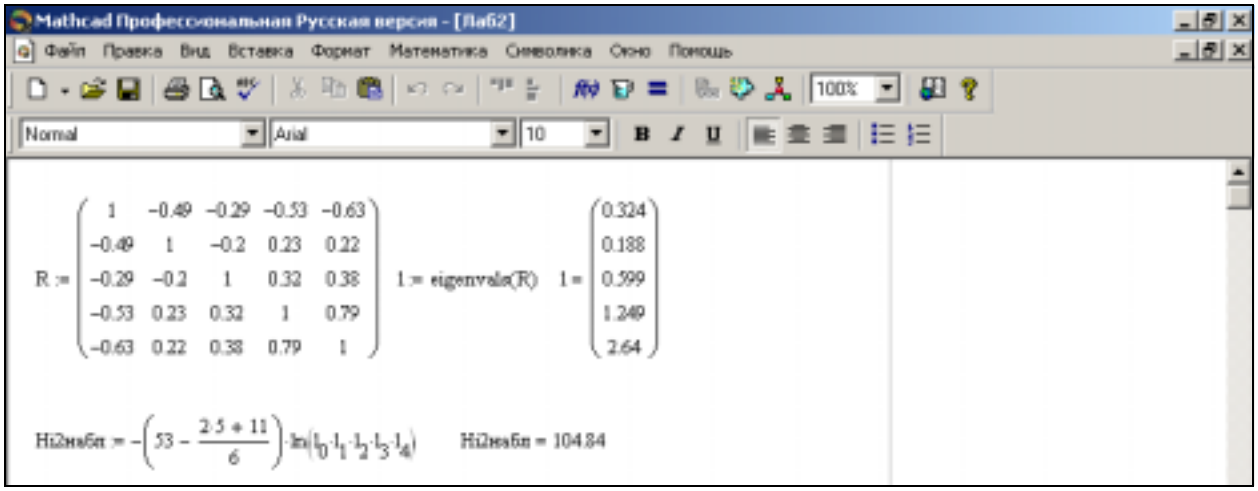


Рисунок 1 – Нахождение наблюденного значения статистики (1)

Критические значения статистики $\chi^2_{кр1}$ и $\chi^2_{кр2}$ определяются из уравнений:

$$P(\chi^2 < \chi^2_{кр1}) = \frac{\alpha}{2},$$
$$P(\chi^2 > \chi^2_{кр2}) = \frac{\alpha}{2}. \qquad (2)$$

Для решения уравнений (2) необходимо воспользоваться либо таблицей критических точек распределения «Хи-квадрат», либо функцией ХИ2ОБР(вероятность, ν) пакета Excel. Критические точки принимают следующие значения:

$$\chi^2_{кр1} = Pi^{-1}(0{,}975;10) = 3{,}25,$$
$$\chi^2_{кр2} = Pi^{-1}(0{,}025;10) = 18{,}31.$$

Так как $\chi^2_{набл} > \chi^2_{кр2}$, то гипотеза H_0 отвергается, матрица парных коэффициентов корреляции значима.

Поскольку матрица парных коэффициентов корреляции значима, можно перейти к решению задачи снижения размерности признакового пространства /2/-/4/.

Снижение размерности признакового пространства методом главных компонент

Для нахождения с помощью пакета Statistica 6.0 оценок собственных чисел корреляционной матрицы $\hat{R}$ после запуска программы и ввода исходных данных (вид экрана представлен на рисунке 2) необходимо выполнить следующие действия:

1) выбрать пункт меню «Статистика» («Statistics»), подпункты «Многомерные исследовательские методы», «Анализ фактора» («Factor Analysis») /1/, /5/, /6/;

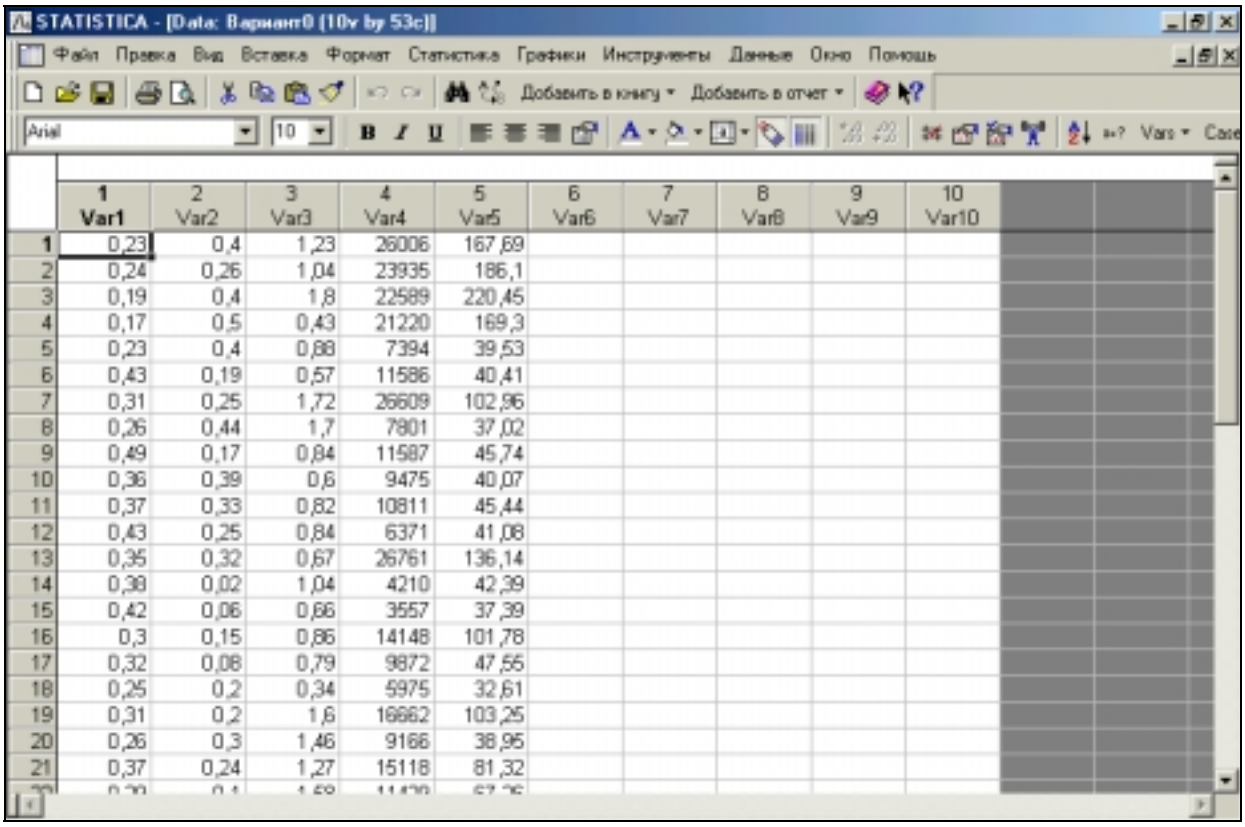
2) в появившейся форме для отбора признаков для анализа нажать кнопку «Variables», выбрать все признаки (1-5) и нажать кнопку «ОК» /5/, /6/;

3) в появившейся форме выбора метода выделения факторов на странице «Advanced» в группе радио-кнопок установить «Principal components» (компонентный анализ), в полях «Максимальное число факторов» и «Минимальное собственное число» ввести значения 5 и 0 соответственно, что даст возможность анализа всех извлеченных главных компонент /1/. Вид формы представлен на рисунке 3;

4) нажать кнопку «ОК».

После выполнения перечисленных шагов на экране появится форма со значениями оценок собственных чисел (eigenvalues), расположенных по убыванию /1/. Вид формы представлен на рисунке 4. Так как расчеты

проводятся на основе оценки матрицы парных коэффициентов корреляции, то $\sum_{i=1}^{5} \hat{\lambda}_i = 5$.



	1 Var1	2 Var2	3 Var3	4 Var4	5 Var5	6 Var6	7 Var7	8 Var8	9 Var9	10 Var10
1	0,23	0,4	1,23	26006	167,69					
2	0,24	0,26	1,04	23935	186,1					
3	0,19	0,4	1,8	22589	220,45					
4	0,17	0,5	0,43	21220	169,3					
5	0,23	0,4	0,88	7394	39,53					
6	0,43	0,19	0,57	11586	40,41					
7	0,31	0,25	1,72	26609	102,96					
8	0,26	0,44	1,7	7801	37,02					
9	0,49	0,17	0,84	11587	45,74					
10	0,36	0,39	0,6	9475	40,07					
11	0,37	0,33	0,82	10811	45,44					
12	0,43	0,25	0,84	6371	41,08					
13	0,35	0,32	0,67	26761	136,14					
14	0,38	0,02	1,04	4210	42,39					
15	0,42	0,06	0,66	3557	37,39					
16	0,3	0,15	0,86	14148	101,78					
17	0,32	0,08	0,79	9872	47,55					
18	0,25	0,2	0,34	5975	32,61					
19	0,31	0,2	1,6	16662	103,25					
20	0,26	0,3	1,46	9166	38,95					
21	0,37	0,24	1,27	15118	81,32					

Рисунок 2 – Исходные данные для анализа

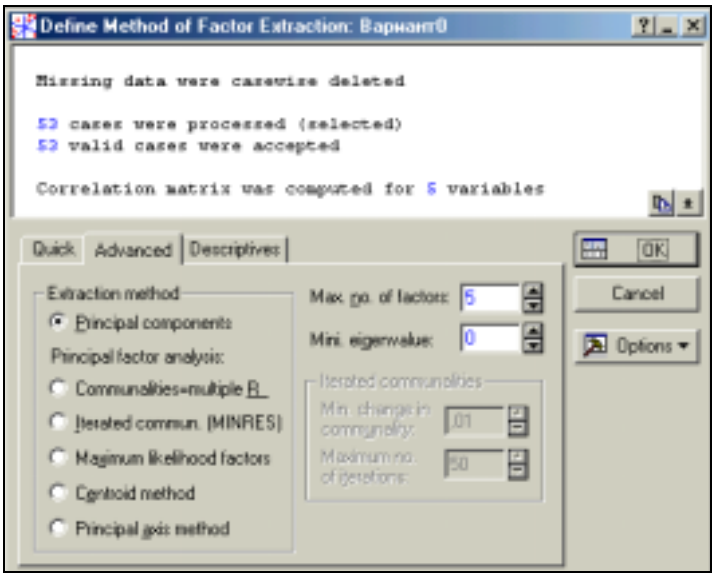


Рисунок 3 – Выбор метода выделения факторов

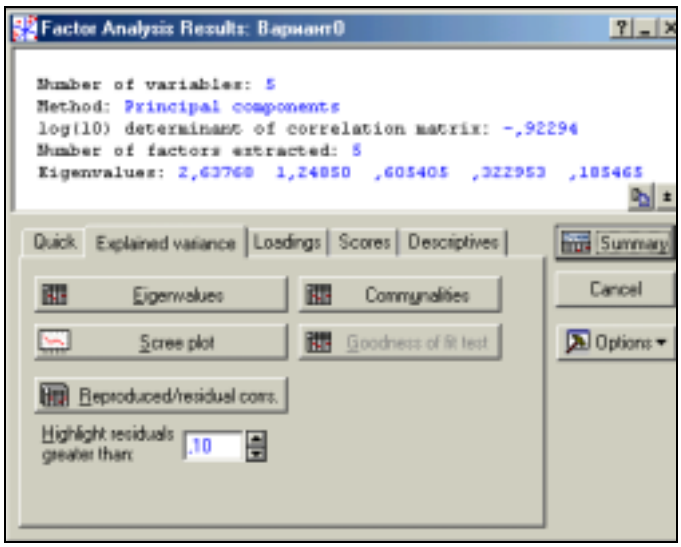


Рисунок 4 – Оценки собственных чисел матрицы парных коэффициентов корреляции

С вероятностью $\gamma = 0{,}95$ построим доверительные интервалы для собственных чисел матрицы парных коэффициентов корреляции /2/. Для построения доверительного интервала для собственного числа λ_i используется статистика $\sqrt{N-1}(\hat{\lambda}_i - \lambda_i)$, имеющая при $N \to \infty$ нормальный закон распределения с параметрами $(0; 2\lambda_i^2)$. В результате решения уравнения $P(|u| < \delta) = \gamma$, где $u = \dfrac{\sqrt{N-1}(\hat{\lambda}_i - \lambda_i)}{\sqrt{2\lambda_i^2}}$ доверительный интервал для i-ого собственного числа λ_i при большом объеме выборки имеет вид:

$$\frac{\hat{\lambda}_i}{1+\delta\sqrt{\dfrac{2}{N-1}}} \le \lambda_i \le \frac{\hat{\lambda}_i}{1-\delta\sqrt{\dfrac{2}{N-1}}}, \qquad (3)$$

где $\hat{\lambda}_i$ – точечная оценка собственного числа λ_i;

δ – q-квантиль стандартного нормального распределения, $q = 1 - \dfrac{\alpha}{2}$;

α – уровень значимости, $\alpha = 1 - \gamma$.

По таблице квантилей стандартного нормального распределения или с помощью функции НОРМСТОБР(q) пакета Excel найдем $\delta = u_{0{,}975} = 1{,}96$.

В результате расчетов доверительные интервалы для собственных чисел имеют вид:

$$1{,}91 \le \lambda_1 \le 4{,}28,$$
$$0{,}90 \le \lambda_2 \le 2{,}03,$$
$$0{,}44 \le \lambda_3 \le 0{,}98,$$
$$0{,}23 \le \lambda_4 \le 0{,}52,$$
$$0{,}13 \le \lambda_5 \le 0{,}30.$$

Так как оценки собственных чисел не попадают в доверительные интервалы других собственных чисел, то нет оснований заподозрить кратность собственных чисел.

Для определения вклада каждой главной компоненты в суммарную дисперсию исходных признаков на странице «Explained variance» формы «Factor analysis results» (рисунок 4) необходимо выбрать кнопку «Eugenvalues» /1/. На экране появится таблица, представленная на рисунке 5.

Eigenvalues (Вариант0)
Extraction: Principal components

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	2,637677	52,75354	2,637677	52,7535
2	1,248500	24,97000	3,886177	77,7235
3	0,605405	12,10811	4,491582	89,8316
4	0,322953	6,45907	4,814535	96,2907
5	0,185465	3,70929	5,000000	100,0000

Рисунок 5 – Вклад главных компонент в суммарную дисперсию исходных признаков

В первом столбце таблицы приведены оценки собственных чисел, в третьем столбце – накопленные значения собственных чисел, во втором и в четвертом столбцах – относительный вклад каждой главной компоненты в суммарную дисперсию и накопленный относительный вклад соответственно. Как видно из рисунка 5 оценка вклада первых двух компонент в суммарную дисперсию исходных признаков составляет 77,72%. На основе доверительных интервалов для собственных числе рассчитаем нижнюю границу уровня информативности: $\frac{1{,}91 + 0{,}9}{5} \cdot 100\% = 56{,}2\%$.

Так как на основании выборочных данных можно рассчитать лишь оценку критерия информативности, то необходимо проверить гипотезу о том, что две главные компоненты (m=2) вносят существенный вклад в дисперсию исходных признаков. Нулевая и альтернативная гипотезы формулируются следующим образом:

H_0 : m главных компонент достаточно,

H_1 : m главных компонент недостаточно.

При $m < \frac{n-1}{2}$ (для рассматриваемого примера это условие не выполнено) для проверки нулевой гипотезы можно воспользоваться χ^2-критерием Бартлетта /2/, /3/:

$$\chi^2 = -\left(N - \frac{1}{6}(2n+5) - \frac{2}{3}m \right) \ln R_{n-m}, \qquad (4)$$

где $R_{n-m} = \dfrac{\left|\hat{R}\right|}{\hat{\lambda}_1 \cdot \hat{\lambda}_2 \cdot \ldots \cdot \hat{\lambda}_m \cdot \left(\dfrac{n - \hat{\lambda}_1 - \hat{\lambda}_2 - \ldots - \hat{\lambda}_m}{n-m} \right)^{n-m}}$.

При справедливости нулевой гипотезы статистика (4) имеет распределение «Хи-квадрат» с числом степеней свободы $\nu = \frac{1}{2}\left((n-m)^2 - n - m - 1\right)$.

Нажатием на кнопку «Scree plot» формы, изображенной на рисунке 4, выводится график собственных чисел, представленный на рисунке 6.

Для расчета коэффициентов линейного преобразования центрировано-нормированных исходных признаков x_j^* ($j = \overline{1...5}$) необходимо на странице «Scores» формы «Factor Analysis Results» выбрать кнопку «Factor scores coefficients» /1/. На экране появится таблица, представленная на рисунке 7.

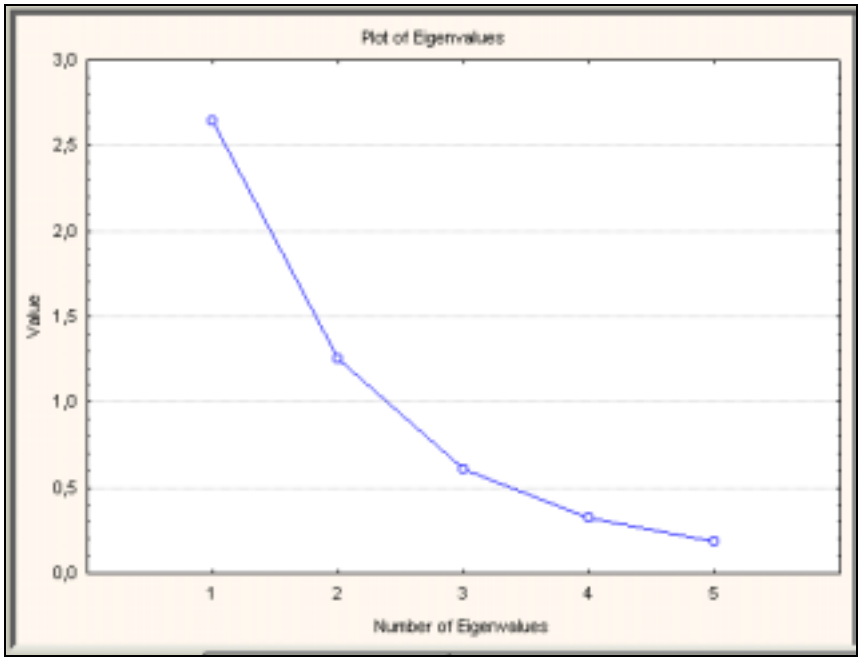


Рисунок 6 – График собственных значений

Variable	Factor Score Coefficients (Вариант0) Rotation: Unrotated Extraction: Principal components Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Var1	0,316944	0,177386	0,495365	1,168615	0,75653
Var2	-0,165116	-0,648961	-0,405568	0,918051	0,40559
Var3	-0,176590	0,575125	-0,782315	0,640966	0,09241
Var4	-0,323224	0,079851	0,666058	0,628449	-1,31425
Var5	-0,340216	0,105829	0,431585	-0,286631	1,70858

Рисунок 7 – Коэффициенты линейного преобразования

Матрица коэффициентов линейного преобразования имеет вид:

$$U^T = \begin{pmatrix} 0{,}317 & 0{,}177 & 0{,}49 & 1{,}17 & 0{,}76 \\ -0{,}165 & -0{,}649 & -0{,}40 & 0{,}92 & 0{,}40 \\ -0{,}177 & 0{,}575 & -0{,}78 & 0{,}64 & 0{,}09 \\ -0{,}323 & 0{,}080 & 0{,}67 & 0{,}63 & -1{,}31 \\ -0{,}340 & 0{,}106 & 0{,}43 & -0{,}29 & 1{,}71 \end{pmatrix}.$$

При снижении размерности признакового пространства до двух главных компонент следует рассматривать только два первых столбца матрицы U^T.

Главные компоненты связаны с центрировано-нормированными исходными признаками следующими линейными комбинациями:

$$z_1 = 0{,}317x_1^* - 0{,}165x_2^* - 0{,}177x_3^* - 0{,}323x_4^* - 0{,}340x_5^*,$$
$$z_2 = 0{,}177x_1^* - 0{,}649x_2^* + 0{,}575x_3^* + 0{,}080x_4^* + 0{,}106x_5^*.$$

Для интерпретации новых признаков необходимо провести анализ матрицы факторных нагрузок. Для этого на странице «Loadings» формы «Factor Analysis Results» следует выбрать кнопку «Factor loadings». На экране появится таблица, изображенная на рисунке 8.

Variable	Factor Loadings (Unrotated) (Вариант0) Extraction: Principal components (Marked loadings are > ,700000) Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Var1	0,835997	0,221467	0,299897	0,377408	0,140309
Var2	-0,435523	-0,810228	-0,245533	0,296488	0,075223
Var3	-0,465787	0,718043	-0,473617	0,207002	0,017138
Var4	-0,852559	0,099694	0,403235	0,202960	-0,243747
Var5	-0,897380	0,132127	0,261284	-0,092569	0,316880
Expl.Var	2,637677	1,249500	0,605405	0,322953	0,185465
Prp.Totl	0,527535	0,249700	0,121081	0,064591	0,037093

Рисунок 8 – Факторные нагрузки

Так как расчеты проводятся на основании матрицы парных коэффициентов корреляции, то элементы матрицы факторных нагрузок являются парными коэффициентами корреляции исходных признаков и

главных компонент /2/. Как видно из таблицы, между исходными признаками и последними тремя главными компонентами не наблюдается тесной связи. Это подтверждает правильность выделения только двух первых главных компонент.

Так как размерность признакового пространства снижена до двух, то матрица факторных нагрузок имеет размерность 5×2:

$$A = \begin{pmatrix} 0{,}836 & 0{,}221 \\ -0{,}436 & -0{,}810 \\ -0{,}466 & 0{,}718 \\ -0{,}853 & 0{,}100 \\ -0{,}897 & 0{,}132 \end{pmatrix}.$$

Первая главная компонента тесно связана (коэффициент корреляции >0,7) с тремя исходными признаками: трудоемкость единицы продукции (X1), среднегодовая численность ППП (X4) и среднегодовая стоимость ОПФ (X5). Поэтому первую главную компоненту можно интерпретировать как «Уровень развития производства». Вторая главная компонента тесно связана (коэффициент корреляции >0,7) с двумя исходными признаками: удельный вес покупных изделий (X2) и премии и вознаграждения на одного работника (X3). Вторую главную компоненту можно интерпретировать «Расходы предприятия».

Расположение признаков в пространстве первых двух главных компонент можно получить нажатием на кнопку «Plot of loadings, 2D». График представлен на рисунке 9.

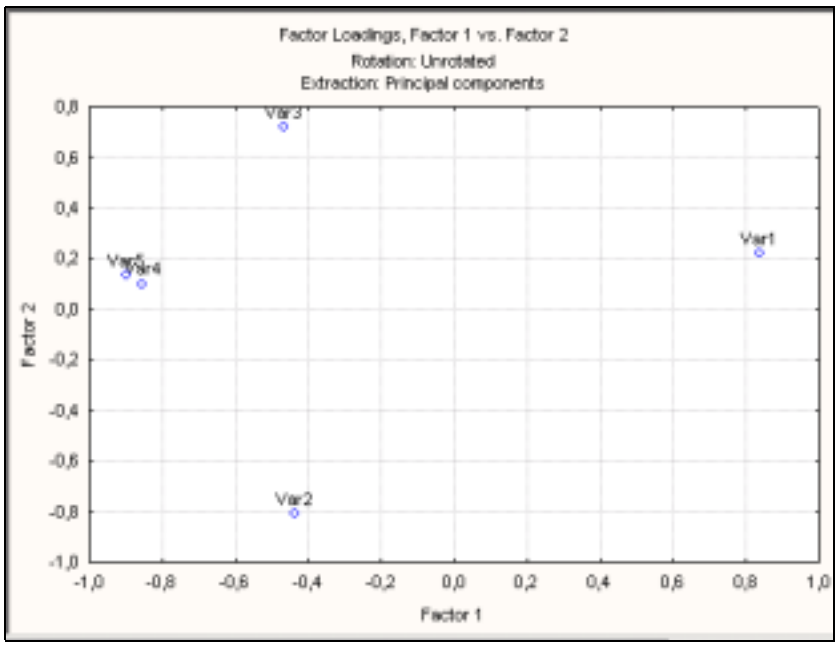


Рисунок 9 – Распределение признаков в пространстве первых двух главных компонент

Центрировано-нормированные исходные признаки связаны с центрировано-нормированными главными компонентами f_1, f_2 следующими выражениями:

$$x_1^* = 0{,}836 f_1 + 0{,}221 f_2;$$
$$x_2^* = -0{,}436 f_1 - 0{,}81 f_2;$$
$$x_3^* = -0{,}466 f_1 + 0{,}718 f_2;$$
$$x_4^* = -0{,}853 f_1 + 0{,}1 f_2;$$
$$x_5^* = -0{,}897 f_1 + 0{,}132 f_2.$$

Для расчета матрицы индивидуальных значений центрировано-нормированных главных компонент необходимо на странице «Scores» формы «Factor Analysis Results» выбрать кнопку «Factor scores» /1/. На экране появится таблица, представленная на рисунке 10.

Factor Scores (Вариант0)
Rotation: Unrotated
Extraction: Principal components

Case	Factor 1	Factor 2
1	-1,01205	-0,22280
2	-0,78367	0,22217
3	-1,36399	0,22779
4	-0,93952	-1,45921
5	0,15865	-0,81743
6	0,92399	0,16806
7	-0,52135	0,88654
8	-0,01201	-0,23618
9	1,03535	0,58841
10	0,56317	-0,77783
11	0,53621	-0,30603
12	0,95730	0,10662
13	-0,32851	-0,19299
14	1,06519	1,13751
15	1,28181	0,70322
16	0,19055	0,45439
17	0,68265	0,62457
18	0,64438	-0,42630

Рисунок 10 – Индивидуальные значения центрировано-нормированных главных компонент

Снижение размерности признакового пространства методом главных факторов

Для реализации метода главных факторов в форме выбора метода выделения факторов, изображенной на рисунке 3, необходимо в группе радио-кнопок установить «Communalities=R?». Вид экрана представлен на рисунке 11. Оценками общностей в данном алгоритме будут служить квадраты оценок множественных коэффициентов корреляции /4/, т.е.:

$$\hat{h}_i^2 = R^2_{i/1,\ldots,i-1,i+1,\ldots,n},\ i = \overline{1\ldots n}. \qquad (5)$$

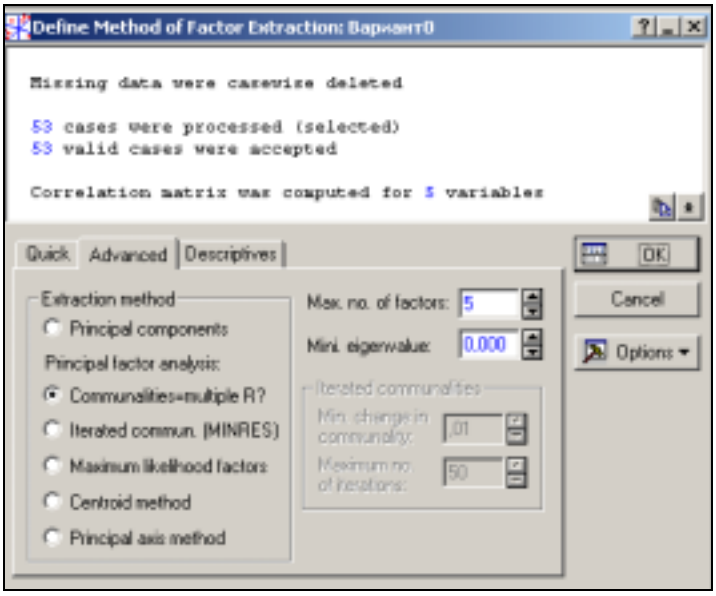


Рисунок 11 – Выбор метода выделения факторов

После нажатия на кнопку «ОК» на экране появится форма результатов факторного анализа, представленная на рисунке 12. В последней строке информационной части экранной формы приведены оценки положительных собственных чисел редуцированной матрицы.

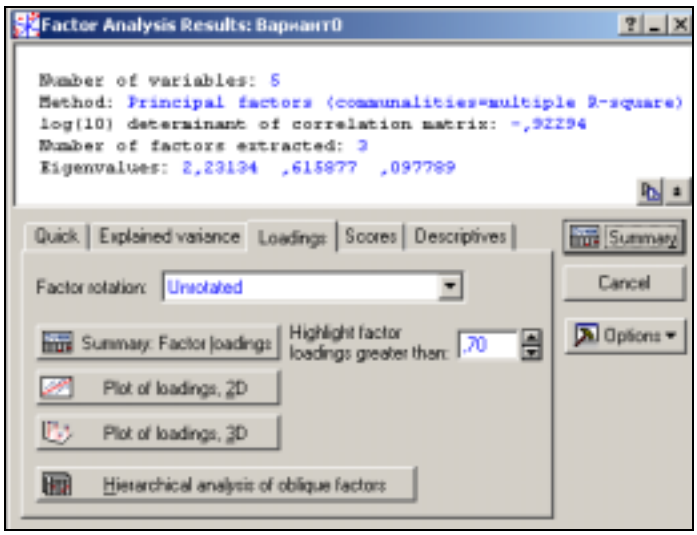


Рисунок 12 – Результаты факторного анализа

Для вывода на экран оценок общностей, рассчитанных по формуле (5), необходимо на форме результатов факторного анализа выбрать страницу «Explained variance». Вид экранной формы представлен на рисунке 13. После нажатия на кнопку «Communalities» на экране появится таблица, представленная на рисунке 14.

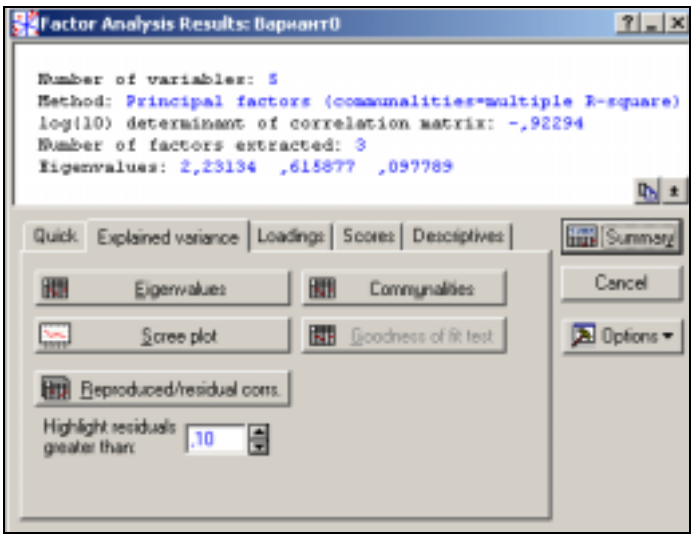


Рисунок 13 – Страница «Explained variance»

Communalities (Вариант0)
Extraction: Principal factors (comm.=multiple R-square)
Rotation: Unrotated

Variable	From 1 Factor	From 2 Factors	From 3 Factors	Multiple R-Square
Var1	0,572255	0,620575	0,653052	0,568090
Var2	0,132684	0,477566	0,477627	0,382801
Var3	0,138398	0,326351	0,356553	0,282302
Var4	0,638979	0,652401	0,681281	0,626404
Var5	0,749026	0,770326	0,776494	0,698305

Рисунок 14 – Результаты расчета общностей

В первом, втором и третьем столбцах таблицы содержатся вклады одного, двух и трех главных факторов в дисперсию признаков. Оценки общностей приведены в четвертом столбце таблицы, изображенной на рисунке 14. На основе оценки матрицы парных коэффициентов корреляции и оценок общностей можно составить оценку редуцированной матрицы $\hat{R}_h$:

$$\hat{R}_h = \begin{pmatrix} 0{,}57 & -0{,}49 & -0{,}29 & -0{,}53 & -0{,}63 \\ -0{,}49 & 0{,}38 & -0{,}2 & 0{,}23 & 0{,}22 \\ -0{,}29 & -0{,}2 & 0{,}28 & 0{,}32 & 0{,}38 \\ -0{,}53 & 0{,}23 & 0{,}32 & 0{,}63 & 0{,}79 \\ -0{,}63 & 0{,}22 & 0{,}38 & 0{,}79 & 0{,}7 \end{pmatrix}.$$

Суммарная общность составляет: $\sum_{i=1}^{5} \hat{h}_i^2 \approx 2{,}558$. Так как сумма оценок первых двух собственных чисел редуцированной матрицы больше суммарной общности ($\hat{\lambda}_1 + \hat{\lambda}_2 = 2{,}231 + 0{,}616 = 2{,}847 > 2{,}558$), то размерность признакового пространства можно снизить до двух общих факторов.

Проведем повторный запуск формы, представленной на рисунке 11, указав в поле «Максимальное число факторов» цифру 2. Результаты представлены на рисунке 15.

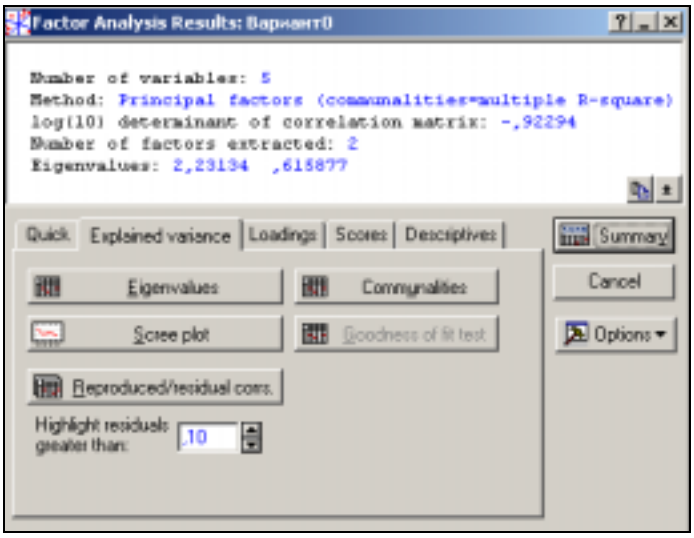


Рисунок 15 – Результаты расчетов для двух общих факторов

Для определения вклада каждого главного фактора в суммарную дисперсию исходных признаков на странице «Explained variance» необходимо выбрать кнопку «Eugenvalues». На экране появится таблица, представленная на рисунке 16.

Eigenvalues (Вариант0)
Extraction: Principal factors (comm.=multiple R-square)

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	2,231342	44,62685	2,231342	44,62685
2	0,615877	12,31754	2,847219	56,94438

Рисунок 16 – Вклад главных факторов в суммарную дисперсию исходных признаков

Вид таблицы вкладов главных факторов аналогичен таблице, представленной на рисунке 5. Вклад двух главных факторов в суммарную дисперсию исходных признаков (в дисперсию процесса) составляет 56,94%. С помощью кнопки «Scree plot» формы, изображенной на рисунке 15, построен график собственных чисел, представленный на рисунке 17.

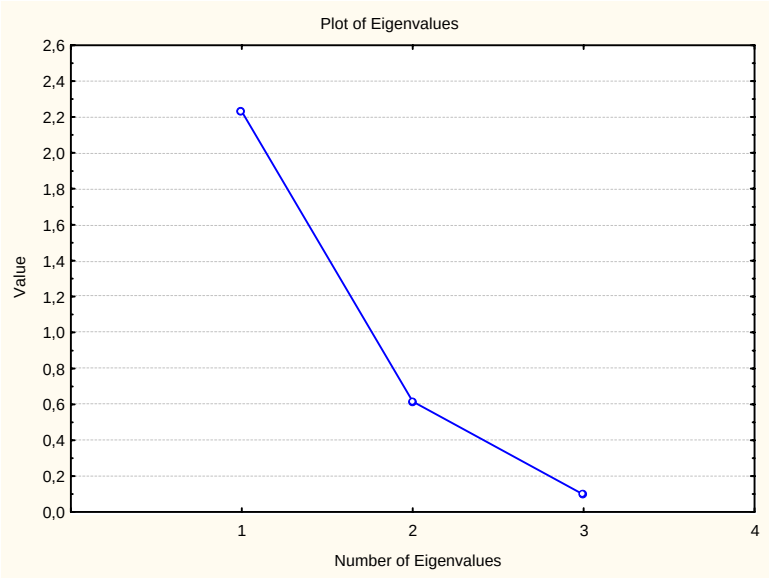
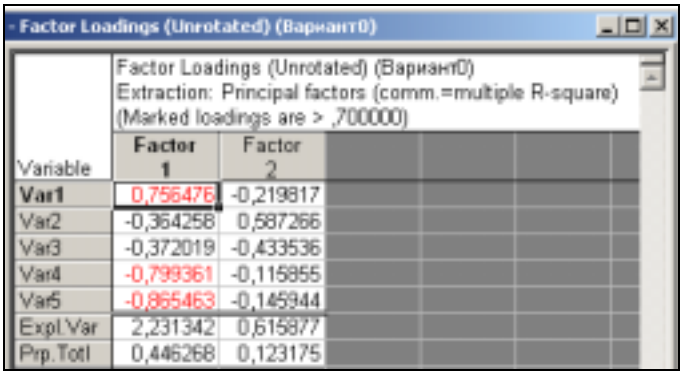


Рисунок 17 – График собственных значений

Таблица весовых коэффициентов при главных факторах выводится на экран при выборе на странице «Loadings» формы результатов факторного анализа (рисунок 15) кнопки «Summary: Factor loadings». Результаты представлены на рисунке 18.



Variable	Factor 1	Factor 2
Var1	0,756476	-0,219817
Var2	-0,364258	0,587266
Var3	-0,372019	-0,433536
Var4	-0,799361	-0,115855
Var5	-0,865463	-0,145944
Expl.Var	2,231342	0,615877
Prp.Totl	0,446268	0,123175

Рисунок 18 – Весовые коэффициенты при общих факторах

Матрица факторных нагрузок имеет вид:

$$A = \begin{pmatrix} 0,76 & -0,22 \\ -0,36 & 0,59 \\ -0,37 & -0,43 \\ -0,80 & -0,12 \\ -0,86 & -0,15 \end{pmatrix}. \qquad (6)$$

После расчета матрицы факторных нагрузок можно проверить гипотезу о достаточности выделения m (m=2) главных факторов. Нулевая и альтернативная гипотезы формулируются следующим образом:

H_0 : m главных факторов достаточно,

H_1 : m главных факторов недостаточно.

При $m < \frac{n-1}{2}$ (для рассматриваемого примера это условие не выполнено) для проверки нулевой гипотезы можно воспользоваться χ^2-критерием Лоули /2/, /3/:

$$\chi^2 = (N-1)\ln\frac{\left|AA^T\right|}{\left|\hat{R}\right|}. \qquad (7)$$

При справедливости нулевой гипотезы статистика (7) имеет распределение «Хи-квадрат» с числом степеней свободы $\nu = \frac{1}{2}\left((n-m)^2 - n - m\right)$.

Так как главные факторы некоррелированы между собой, то элементы матрицы A являются коэффициентами корреляции между исходными признаками и главными факторами. Как видно из рисунка 18 второй главный фактор не имеет тесной связи ни с одним из признаков. Это затруднит его интерпретацию. Расположение исходных признаков на плоскости, образованной двумя главными факторами можно получить с помощью кнопки «Plot of loadings, 2D». График представлен на рисунке 19. Попробуем упростить структуру главных факторов с помощью ортогонального вращения /3/, /4/. Для оценки структуры обобщенных факторов выберем критерий «Квартимакс». Для этого установим в поле «Factor rotation» значение «Quartimax raw». Вид экрана представлен на рисунке 20.

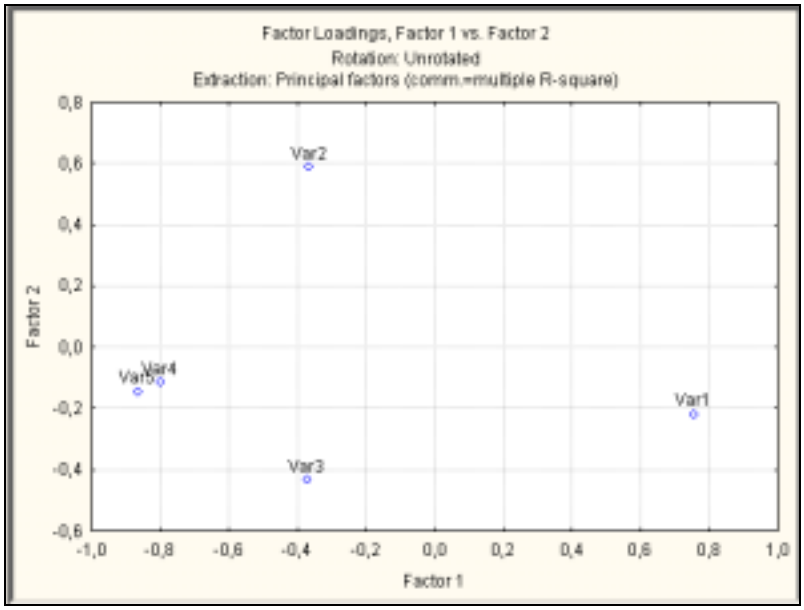


Рисунок 19 – Расположение исходных признаков на плоскости, образованной главными факторами

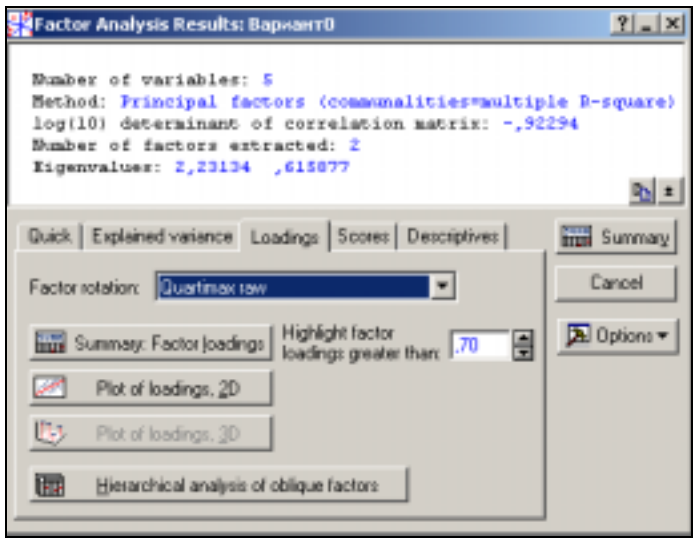
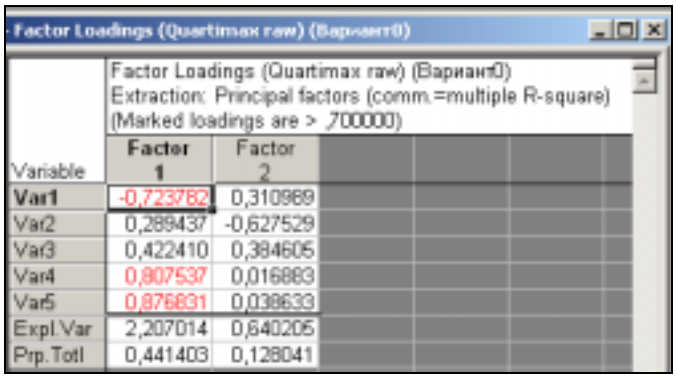


Рисунок 20 – Выбор критерия для оценки структуры обобщенных факторов

С помощью кнопки «Summary: Factor loadings» на экране будет представлена таблица нагрузок после вращения главных факторов. Вид экрана изображен на рисунке 21. График распределения исходных признаков на плоскости, образованной обобщенными факторами представлен на рисунке 22.



Variable	Factor 1	Factor 2
Var1	-0,723762	0,310989
Var2	0,289437	-0,627529
Var3	0,422410	0,384605
Var4	0,907537	0,016883
Var5	0,876831	0,038633
Expl.Var	2,207014	0,640205
Prp.Totl	0,441403	0,128041

Рисунок 21 – Весовые коэффициенты факторов после вращения

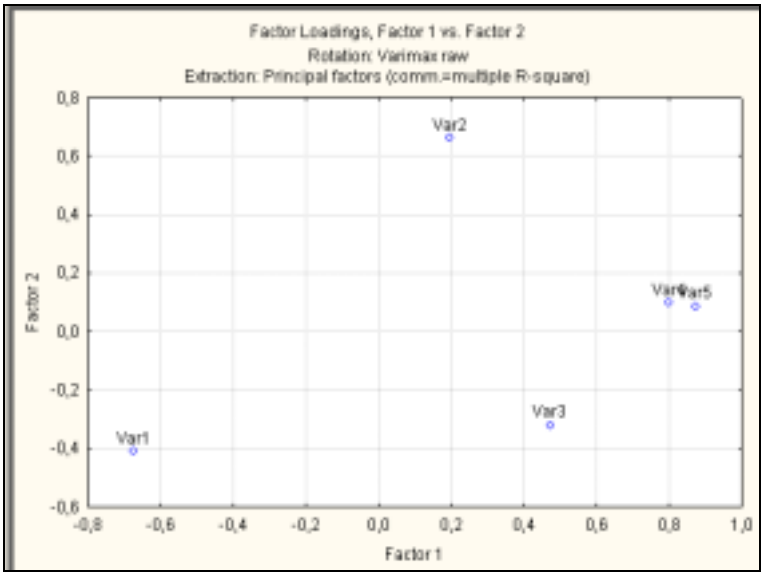


Рисунок 22 – Расположение исходных признаков на плоскости, образованной обобщенными факторами

Матрица факторных нагрузок после вращения имеет вид:

$$B = \begin{pmatrix} -0{,}72 & 0{,}31 \\ 0{,}29 & -0{,}63 \\ 0{,}42 & 0{,}38 \\ 0{,}81 & 0{,}02 \\ 0{,}88 & 0{,}04 \end{pmatrix}. \qquad (8)$$

Сравнивая матрицы A (6) и B (8) можно сделать вывод, что структура обобщенных факторов после вращения улучшилась незначительно. Первый фактор тесно связан (коэффициент корреляции >0,7) с тремя исходными признаками: трудоемкость единицы продукции (X1), среднегодовая численность ППП (X4) и среднегодовая стоимость ОПФ (X5). Связь средней силы наблюдается между первым фактором и признаком «Премии и вознаграждения на одного работника» (X3). Первый главный фактор можно интерпретировать «Уровень развития производства». Связь средней силы наблюдается между вторым фактором и признаками: удельный вес покупных изделий (X2) и премии и вознаграждения на одного работника (X3). Второй главный фактор можно интерпретировать «Расходы предприятия».

Таблица вкладов одного и двух главных факторов в дисперсию исходных признаков выводится на экран при выборе на странице «Explained variance» кнопки «Communalities». Таблица, представленная на рисунке 23.

Communalities (Вариант0)
Extraction: Principal factors (comm.=multiple R-square)
Rotation: Quartimax raw

Variable	From 1 Factor	From 2 Factors	Multiple R-Square
Var1	0,523861	0,620575	0,568090
Var2	0,083774	0,477566	0,382801
Var3	0,178430	0,326351	0,282302
Var4	0,652116	0,652401	0,626404
Var5	0,768833	0,770326	0,698305

Рисунок 23 – Вклады одного и двух главных факторов в дисперсию признаков

Элементы первого столбца таблицы, представленной на рисунке 23, равны квадратам соответствующих элементов первого столбца матрицы B, а элементы второго столбца таблицы – сумме квадратов соответствующих элементов первого и второго столбцов матрицы B, т.е. общностям. Таким образом, во втором столбце таблицы представлены оценки общностей, рассчитанные по матрице B, а в третьем столбце – оценки общностей, рассчитанные по формуле (5).

Так как исходные признаки процентрированы и пронормированы, а главные факторы некоррелированы между собой, то оценки характерностей можно рассчитать следующим образом:

$$\hat{d}_1^2 = 1 - 0{,}62 = 0{,}38;$$
$$\hat{d}_2^2 = 1 - 0{,}48 = 0{,}52;$$
$$\hat{d}_3^2 = 1 - 0{,}33 = 0{,}67;$$
$$\hat{d}_4^2 = 1 - 0{,}65 = 0{,}35;$$
$$\hat{d}_{51}^2 = 1 - 0{,}77 = 0{,}23.$$

Центрировано-нормированные исходные признаки связаны с главными и характерными факторами следующими выражениями:

$$x_1^* = -0{,}72 f_1 + 0{,}31 f_2 + 0{,}62 v_1,$$
$$x_2^* = 0{,}29 f_1 - 0{,}63 f_2 + 0{,}72 v_2,$$
$$x_3^* = 0{,}42 f_1 + 0{,}38 f_2 + 0{,}82 v_3,$$
$$x_4^* = 0{,}81 f_1 + 0{,}02 f_2 + 0{,}59 v_4,$$
$$x_5^* = 0{,}88 f_1 + 0{,}04 f_2 + 0{,}48 v_5.$$

Оценка редуцированной матрицы парных коэффициентов корреляции, рассчитанная по матрице нагрузок B, и оценка остаточной матрицы парных коэффициентов корреляции выводятся на экран с помощью кнопки «Reproduced/residual corrs.» на странице «Explained variance». Данные матрицы представлены на рисунках 24 и 25.

Reproduced Correlations (Вариант0)
Extraction: Principal factors (comm.=multiple R-square)

Variable	Var1	Var2	Var3	Var4	Var5
Var1	0,62	-0,40	-0,19	-0,58	-0,62
Var2	-0,40	0,48	-0,12	0,22	0,23
Var3	-0,19	-0,12	0,33	0,35	0,39
Var4	-0,58	0,22	0,35	0,65	0,71
Var5	-0,62	0,23	0,39	0,71	0,77

Рисунок 24 – Оценка редуцированной матрицы

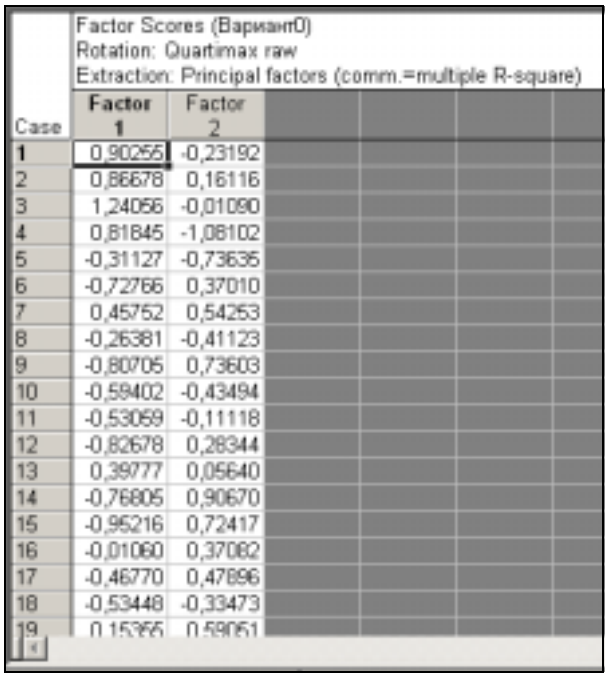
Residual Correlations (Вариант0)
Extraction: Principal factors (comm.=multiple R-square)
(Marked residuals are > ,100000)

Variable	Var1	Var2	Var3	Var4	Var5
Var1	0,38	-0,09	-0,11	0,05	-0,01
Var2	-0,09	0,52	-0,08	0,01	-0,01
Var3	-0,11	-0,08	0,67	-0,03	-0,01
Var4	0,05	0,01	-0,03	0,35	0,08
Var5	-0,01	-0,01	-0,01	0,08	0,23

Рисунок 25 – Оценка остаточной матрицы парных коэффициентов корреляции

На главной диагонали матрицы, представленной на рисунке 25, расположены оценки характерностей $\hat{d}_i^2$, $i = \overline{1...5}$.

Для расчета матрицы индивидуальных значений обобщенных факторов необходимо на странице «Scores» формы «Factor Analysis Results» выбрать кнопку «Factor scores». На экране появится таблица, представленная на рисунке 26.



Factor Scores (Вариант0)
Rotation: Quartimax raw
Extraction: Principal factors (comm.=multiple R-square)

Case	Factor 1	Factor 2
1	0,90255	-0,23192
2	0,86678	0,16116
3	1,24056	-0,01090
4	0,81845	-1,08102
5	-0,31127	-0,73636
6	-0,72766	0,37010
7	0,45752	0,54253
8	-0,26381	-0,41123
9	-0,80705	0,73603
10	-0,59402	-0,43494
11	-0,53059	-0,11118
12	-0,82678	0,28344
13	0,39777	0,05640
14	-0,76805	0,90670
15	-0,95216	0,72417
16	-0,01060	0,37082
17	-0,46770	0,47896
18	-0,53448	-0,33473
19	0,15366	0,59051

Рисунок 26 – Индивидуальные значения обобщенных факторов

Для последующей обработки матрицы индивидуальных значений обобщенных факторов ее можно сохранить в файле с помощью кнопки «Save factor scores».

Выводы

В результате реализации метода главных компонент и метода главных факторов размерность признакового пространства снижена с пяти признаков до двух. Вклад двух главных компонент в суммарную дисперсию исходных признаков составил 77,72%, вклад двух главных факторов в суммарную дисперсию исходных признаков – 56,94%. Проведено ортогональное вращение плоскости, образованной двумя главными факторами. Интерпретация главных компонент и главных факторов совпадает.

4 Содержание письменного отчета

Отчет должен быть оформлен на листах формата А4 с титульным листом, оформленным соответствующим образом, и содержать следующее:

1) исходные данные для анализа;
2) постановку задачи;
3) краткое изложение теории;
4) результаты выполнения лабораторной работы.

5 Вопросы к защите

1) В чем заключаются необходимость снижения размерности признакового пространства?
2) Каковы предпосылки, обуславливающие возможность снижения размерности признакового пространства?
3) Что понимается под «мерой информативности» в методе главных компонент?
4) Сформулировать определение *к*-ой главной компоненты
5) Сформулировать оптимизационную задачу для построения первой главной компоненты
6) В чем заключается алгоритм нахождения коэффициентов линейного преобразования исходных признаков?
7) Каковы основные числовые характеристики главных компонент?
8) Каковы свойства матрицы факторных нагрузок?
9) Чему равен вклад главных компонент в дисперсию *i*-го признака?
10) Чему равен вклад *j*-ой главной компоненты в суммарную дисперсию признаков?
11) Как определить матрицу индивидуальных значений главных компонент?
12) Описать процедуру статистического анализа при реализации метода главных компонент
13) Привести математическую модель главных факторов
14) Что представляют собой составляющие в разложении дисперсии *i*-го признака?
15) В чем заключается фундаментальная теорема факторного анализа?
16) Определить понятие факторного отображения
17) Определить понятие факторной структуры
18) Определить понятия пространства главных факторов и полного факторного пространства
19) Какие существуют методы оценки общностей?
20) Сформулировать оптимизационную задачу для построения первого главного фактора
21) В чем заключается алгоритм метода главных факторов?
22) Дать понятия простой ортогональной, косоугольной и случайной структуры
23) В чем идея и суть вращения пространства главных факторов?
24) Какие существуют методы оценки индивидуальных значений главных факторов?
25) Описать процедуру статистического анализа при реализации метода главных факторов

Список использованных источников

1 **Дуброва Т.А.** Факторный анализ с использованием ППП «STATISTICA» [Текст]: учебное пособие / Т.А. Дуброва, Д.Э. Павлов, Н.П. Осипова. – М.: Московский государственный университет экономики, статистики и информатики, 2000. – 64 с.

2 **Айвазян С.А.** Прикладная статистика и основы эконометрики [Текст]: учебник для вузов / С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ, 1998. – 1022с.

3 **Сошникова Л.А.** Многомерный статистический анализ в экономике [Текст]: учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г.Е. Уебе, М. Шефер. – М.: ЮНИТИ, 1999. – 598 с.

4 **Дубров А.М.** Многомерные статистические методы [Текст]: учебник / А.М.Дубров, В.С. Мхитарян, Л.И. Трошин. – М.: Финансы и статистика, 1998. – 352 с.

5 **Боровиков В.П.** STATISTICA – Статистический анализ и обработка данных в среде Windows [Текст] / В.П. Боровиков, И.П. Боровиков. – М.: Инф. изд. дом «Филин», 1998. – 608 с.

6 **Тюрин Ю.Н.** Статистический анализ данных на компьютере [Текст] / Ю.Н. Тюрин, А.А. Макаров; под ред. В.Э. Фигурнова. – М.: ИНФРА-М, 1998. – 528 с.

Приложение А
(обязательное)

Исходные данные для анализа

Таблица А.1 – Выборочные данные

№ объекта	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1	0,23	0,78	0,40	1,37	1,23	0,23	1,45	26006	167,69	47750	6,40	166,32	10,08	17,72
2	0,24	0,75	0,26	1,49	1,04	0,39	1,30	23935	186,10	50391	7,80	92,88	14,76	18,39
3	0,19	0,68	0,40	1,44	1,80	0,43	1,37	22589	220,45	43149	9,76	158,04	6,48	26,46
4	0,17	0,70	0,50	1,42	0,43	0,18	1,65	21220	169,30	41089	7,90	93,96	21,96	22,37
5	0,23	0,62	0,40	1,35	0,88	0,15	1,91	7394	39,53	14257	5,35	173,88	11,88	28,13
6	0,43	0,76	0,19	1,39	0,57	0,34	1,68	11586	40,41	22661	9,90	162,30	12,60	17,55
7	0,31	0,73	0,25	1,16	1,72	0,38	1,94	26609	102,96	52509	4,50	88,56	11,52	21,92
8	0,26	0,71	0,44	1,27	1,70	0,09	1,89	7801	37,02	14903	4,88	101,16	8,28	19,52
9	0,49	0,69	0,17	1,16	0,84	0,14	1,94	11587	45,74	25587	3,46	166,32	11,52	23,99
10	0,36	0,73	0,39	1,25	0,60	0,21	2,06	9475	40,07	16821	3,60	140,76	32,40	21,76
11	0,37	0,68	0,33	1,13	0,82	0,42	1,96	10811	45,44	19459	3,56	128,52	11,52	25,68
12	0,43	0,74	0,25	1,10	0,84	0,05	1,02	6371	41,08	12973	5,65	177,84	17,28	18,13
13	0,35	0,66	0,32	1,15	0,67	0,29	1,85	26761	136,14	50907	4,28	114,48	16,20	25,74
14	0,38	0,72	0,02	1,23	1,04	0,48	0,88	4210	42,39	6920	8,85	93,24	13,32	21,21
15	0,42	0,68	0,06	1,39	0,66	0,41	0,62	3557	37,39	5736	8,52	126,72	17,28	22,97
16	0,30	0,77	0,15	1,38	0,86	0,62	1,09	14148	101,78	26705	7,19	91,80	9,72	16,38
17	0,32	0,78	0,08	1,35	0,79	0,56	1,60	9872	47,55	20068	4,82	69,12	16,20	13,21

Продолжение таблицы А.1

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
18	0,25	0,78	0,20	1,42	0,34	1,76	1,53	5975	32,61	11487	5,46	66,24	24,84	14,48
19	0,31	0,81	0,20	1,37	1,60	1,31	1,40	16662	103,25	32029	6,20	67,68	14,76	13,38
20	0,26	0,79	0,30	1,41	1,46	0,45	2,22	9166	38,95	18946	4,25	50,40	7,56	13,69
21	0,37	0,77	0,24	1,35	1,27	0,50	1,32	15118	81,32	28025	5,38	70,56	8,64	16,66
22	0,29	0,78	0,10	1,48	1,58	0,77	1,48	11429	67,26	20968	5,88	72,00	8,64	15,06
23	0,34	0,72	0,11	1,24	0,68	1,20	0,68	6462	59,92	11049	9,27	97,20	9,00	20,09
24	0,23	0,79	0,47	1,40	0,86	0,21	2,30	24628	107,34	45893	4,36	80,28	14,76	15,98
25	0,17	0,77	0,53	1,45	1,98	0,25	1,37	49727	512,60	99400	10,31	51,48	10,08	18,27
26	0,29	0,80	0,34	1,40	0,33	0,15	1,51	11470	53,81	20719	4,69	105,12	14,76	14,42
27	0,41	0,71	0,20	1,28	0,45	0,66	1,43	19448	80,83	36813	4,16	128,52	10,44	22,76
28	0,41	0,79	0,24	1,33	0,74	0,74	1,82	18963	59,42	33956	3,13	94,68	14,76	15,41
29	0,22	0,76	0,54	1,22	0,03	0,32	2,62	9185	36,96	17016	4,02	85,32	20,52	19,35
30	0,29	0,78	0,40	1,28	0,99	0,89	1,75	17478	91,43	34873	5,23	76,32	14,40	16,83
31	0,51	0,62	0,20	1,47	0,24	0,23	1,54	6265	17,16	11237	2,74	153,00	24,84	30,53
32	0,36	0,75	0,64	1,27	0,57	0,32	2,25	8810	27,29	17306	3,10	107,64	11,16	17,98
33	0,23	0,71	0,42	1,51	1,22	0,54	1,07	17659	184,33	39250	10,44	90,72	6,48	22,09
34	0,26	0,74	0,27	1,46	0,68	0,75	1,44	10342	58,42	19074	5,65	82,44	9,72	18,29
35	0,27	0,65	0,37	1,27	1,00	0,16	1,40	8901	59,40	18452	6,67	79,92	3,24	26,05
36	0,29	0,66	0,38	1,43	0,81	0,24	1,31	8402	49,63	17500	5,91	120,96	6,48	26,20
37	0,01	0,84	0,35	1,50	1,27	0,59	1,12	32625	391,27	7888	11,99	84,60	5,40	17,26
38	0,02	0,74	0,42	1,35	1,14	0,56	1,16	31160	258,62	58947	8,30	85,32	6,12	18,83
39	0,18	0,75	0,32	1,41	1,89	0,63	0,88	46461	75,66	94697	1,63	101,52	8,64	19,70
40	0,25	0,75	0,33	1,47	0,67	1,10	1,07	13833	123,68	29626	8,94	107,64	11,88	16,87
41	0,31	0,79	0,29	1,35	0,96	0,39	1,24	6391	37,21	11688	5,82	85,32	7,92	14,63

Продолжение таблицы А.1

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
42	0,38	0,72	0,30	1,40	0,67	0,73	1,49	11115	53,37	21955	4,80	131,76	10,08	22,17
43	0,24	0,70	0,56	1,20	0,98	0,28	2,03	6555	32,87	12243	5,01	116,64	18,72	22,62
44	0,31	0,66	0,42	1,15	1,16	0,10	1,84	11085	45,63	20193	4,12	138,24	13,68	26,44
45	0,42	0,69	0,26	1,09	0,54	0,68	1,22	9484	48,41	20122	5,10	156,96	16,56	22,26
46	0,51	0,71	0,16	1,26	1,23	0,87	1,72	3967	13,58	7612	3,49	137,52	14,76	19,13
47	0,31	0,73	0,45	1,36	0,78	0,49	1,75	15283	63,99	27404	4,19	135,72	7,92	18,28
48	0,37	0,65	0,31	1,15	1,16	0,16	1,46	20874	104,55	39648	5,01	155,52	18,36	28,23
49	0,16	0,82	0,08	1,87	4,44	0,85	1,60	19418	222,11	43799	11,44	48,60	8,28	12,39
50	0,18	0,80	0,68	1,17	1,06	0,13	1,47	3351	25,76	6235	7,67	42,84	14,04	11,64
51	0,43	0,83	0,03	1,61	2,13	0,49	1,38	6338	29,52	11524	4,66	142,20	16,92	8,62
52	0,40	0,70	0,02	1,34	1,21	0,09	1,41	9756	41,99	17309	4,30	145,80	11,16	20,10
53	0,31	0,74	0,22	1,22	2,20	0,79	1,39	11795	78,11	22225	6,62	120,52	14,76	19,41

X1 – трудоемкость единицы продукции;
X2 – удельный вес рабочих в составе ППП;
X3 – удельный вес покупных изделий;
X4 – коэффициент сменности оборудования;
X5 – премии и вознаграждения на одного работника;
X6 – удельный вес потерь от брака;
X7 – фондоотдача;
X8 – среднегодовая численность ППП;
X9 – среднегодовая стоимость ОПФ;
X10 – среднегодовой фонд заработной платы ППП;
X11 – фондовооруженность труда;
X12 – оборачиваемость нормируемых оборотных средств;
X13 – оборачиваемость ненормируемых оборотных средств;
X14 – непроизводственные расходы.

Таблица А.2 – Варианты заданий

№ варианта	Номера признаков
0	1, 3, 5, 8, 9
1	1, 3, 5, 8, 10
2	1, 5, 8, 9, 10
3	1, 3, 5, 10, 11
4	1, 5, 8, 10, 11
5	1, 3, 5, 9, 10
6	1, 4, 8, 9, 10
7	1, 4, 6, 9, 10
8	3, 5, 8, 9, 10
9	3, 5, 6, 10, 11
10	1, 2, 3, 4, 6
11	1, 2, 4, 6, 8
12	1, 2, 3, 9, 10
13	1, 2, 4, 7, 11
14	1, 2, 3, 7, 11
15	5, 7, 12, 13, 14
16	2, 3, 7, 12, 14
17	2, 3, 4, 8, 9
18	5, 6, 7, 8, 14
19	5, 6, 7, 9, 14
20	1, 2, 3, 5, 6