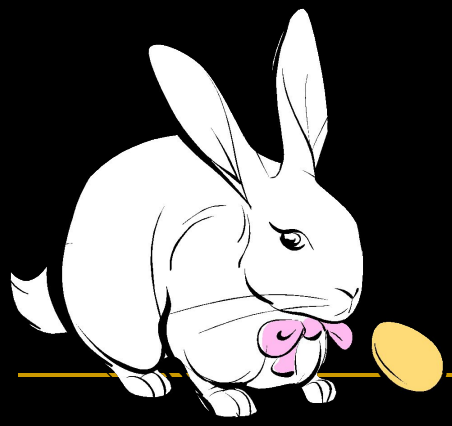
Занятие 9

Основы многомерных методов анализа. Факторный анализ.



Методы многомерного анализа (multivariate analyses)

Предназначены для анализа многомерных данных

Много независимых переменных –

✔Многофакторная ANOVA

✔Множественная регрессия

Много зависимых переменных (или переменных, которые нельзя разделить на зависимые и независимые) –

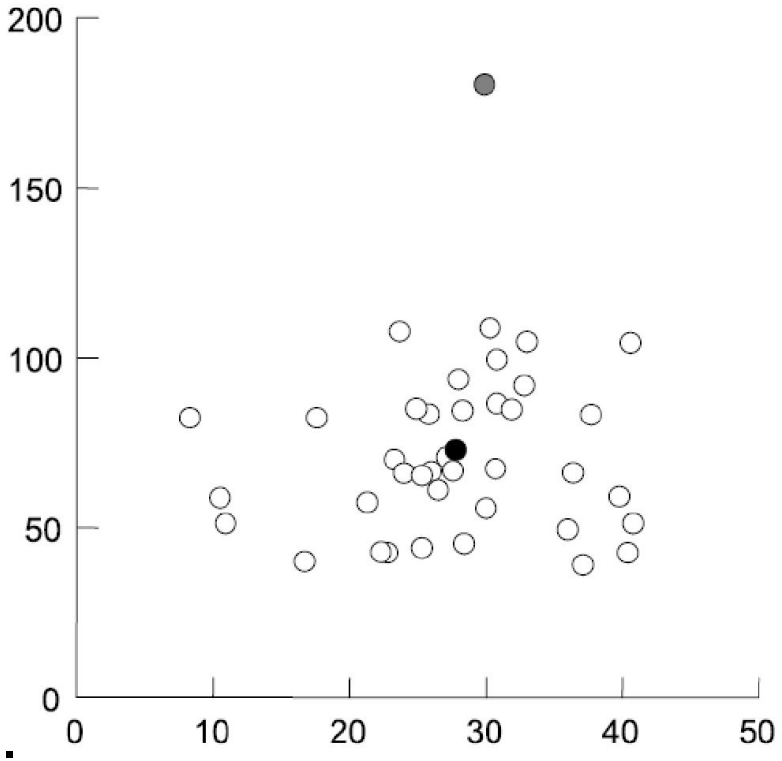
✔multivariate analyses

В массиве данных *n* объектов, для каждого измерено *p* переменных.

multivariate analyses

Многомерное распределение

- Его «центр» - центроид (в одномерном - среднее значение).
- Как оценить разброс в нём? (в одномерном – суммы квадратов и дисперсия).



У нас есть: 1) изменчивость внутри каждой переменной; 2) взаимозависимость переменных.

Как же работать с этими разными изменчивостями?

Многомерные методы в большой степени описательны, но если предполагается тестирование гипотез, надо чтобы данные соответствовали многомерному нормальному распределению.

multivariate analyses

Используют особые таблицы - матрицы.

Одна матрица у нас уже есть – матрица исходных данных (Y).

Clevenger & Waltho изучали, сколько раз и как (на велосипеде-верхом-пешком) люди переходят дорогу в заповеднике на разных 11 переходах.

	Raw		
Underpass	Bicycle	Horse	Foot
1	0	6	7
2	5	3	45
3	6	6	14
4	21	5	20
5	189	42	34
6	8	138	77
7	462	186	129
8	19	12	80
9	595	58	241
10	1	10	10
11	0	10	29

multivariate analyses

Матрица (*p* x *p*) с суммами квадратов на диагонали (sums-of-squares-and-cross-products, SSCP)

$$\begin{bmatrix} \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2 & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{i1}-\bar{y}_1) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i1}-\bar{y}_1) \\ \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2) & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2 & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i2}-\bar{y}_2) \\ \cdots & \cdots & \sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2 & \cdots \\ \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{ip}-\bar{y}_p) & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{ip}-\bar{y}_p) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)^2] \end{bmatrix}$$

Матрица дисперсий и ковариаций (covariances, C) – предыдущая матрица, где все элементы поделили на число степеней свободы (n-1). Сумма элементов её диагонали – сумма дисперсии.

$$\begin{bmatrix} s_1^2 & s_{12}^2 & \cdots & s_{p1}^2 \\ s_{12}^2 & s_2^2 & \cdots & s_{p2}^2 \\ \cdots & \cdots & s_j^2 & \cdots \\ s_{1p}^2 & s_{2p}^2 & \cdots & s_p^2 \end{bmatrix}$$

multivariate analyses

Матрица корреляций (correlation matrix, R) – получится, если в предыдущей матрице каждый элемент поделить на его стандартное отклонение.

На главной диагонали – единицы, все остальные элементы – коэффициенты корреляции

$$\begin{bmatrix} 1 & r_{21} & \dots & r_{p1} \\ r_{12} & 1 & \dots & r_{p2} \\ \dots & \dots & 1 & \dots \\ r_{1p} & r_{2p} & \dots & 1 \end{bmatrix}$$

$$r = \frac{\sum z_X z_Y}{n-1} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y}$$

multivariate analyses

Фундаментальная процедура в многомерном анализе – получение линейных комбинаций исходных переменных, так, что общая изменчивость по-новому распределяется между ними.

Для каждого *i*-го (от 1 до *n*) объекта и *p* исходных переменных можно рассчитать значение новой *k*-той переменной как

$$z_{ik} = c_1 y_{i1} + c_2 y_{i2} + \ldots + c_j y_{ij} + \ldots + c_p y_{ip}$$

Здесь y – значения исходных переменных для данного объекта, c – коэффициенты, показывающие величину вклада данной исходной переменной в новую переменную. В некоторых моделях добавляют ещё константу - intercept

Новые переменные называются дискриминантными функциями, каноническими функциями, главными компонентами (principal components) или факторами (в зависимости от типа анализа). Линейная комбинация аналогична уравнению линейной регрессии.

multivariate analyses

Новые переменные формируют так, чтобы первая объясняла максимум изменчивости исходных переменных, вторая – максимум оставшейся изменчивости, и.т.д., но так, чтобы новые переменные не коррелировали друг с другом.
Так можно получить *p* новых переменных, но большая часть дисперсии должна сосредоточиться в нескольких первых.

Собственное значение (λ) = **eigenvalue** – показатель того, какая доля общей изменчивости приходится на компоненту.
Это популяционные параметры, у них есть выборочные оценки – l
Их сумма = сумме дисперсий (если мы их строим на основе матрицы ковариаций), или = числу исходных переменных (для матрицы корреляций).
Собственный вектор = **eigenvector** – просто список коэффициентов при исходных переменных для каждой компоненты.

multivariate analyses

Выделим новые компоненты для переходов:

В примере используется матрица ковариаций

	Bicycle	Horse	Foot
Bicycle	44 906.018		
Horse	7336.382	3862.018	
Foot	13 084.709	2205.191	4903.655

Значения собственных значений для новых переменных

Eigenvector	1	2	3
Eigenvalue	50 075.681	2592.350	1003.660
Percentage of total variance	93.300	4.830	1.870

	1	2	3
Bicycle	0.945	0.160	0.284
Horse	0.164	−0.986	0.011
Foot	0.282	0.036	−0.959

Коэффициенты для новых переменных (столбец = eigenvector)

multivariate analyses

Теперь можно для каждого конкретного перехода посчитать значения новых переменных = компонент. и, например, использовать в дальнейшем анализе.

Мы рассмотрели способ получения компонент (и их значений для объектов) из матриц ковариаций или корреляций (*p* x *p*). – R-mode analysis.

Есть другой способ: построить матрицу «корреляций» = «дистанций» между объектами (*n* x *n*) в исходных переменных, и из линейных комбинаций объектов рассчитать значения новых компонент, и затем найти eigenvectors - Q-mode analysis.

Разные пути используются в разных типах многомерного анализа, но вообще-то они алгебраически связаны.

Матрица «дистанций» меду объектами (dissimilarity matrix):

Objects	Variables 1	2	3	etc.
1	y_{11}	y_{12}	y_{13}	
2	y_{21}	y_{22}	y_{23}	
3	y_{31}	y_{32}	y_{33}	
etc.				

Variables	Variables 1	2	3	etc.
1	1	r_{12}	r_{13}	
2	r_{21}	1	r_{23}	
3	r_{31}	r_{32}	1	
etc.				

Objects	Objects 1	2	3	etc.
1	0	d_{12}	d_{13}	
2	d_{21}	0	d_{23}	
3	d_{31}	d_{32}	0	
etc.				

multivariate analyses

Есть много показателей «дистанции» между объектами (самый очевидный – евклидовы расстояния).

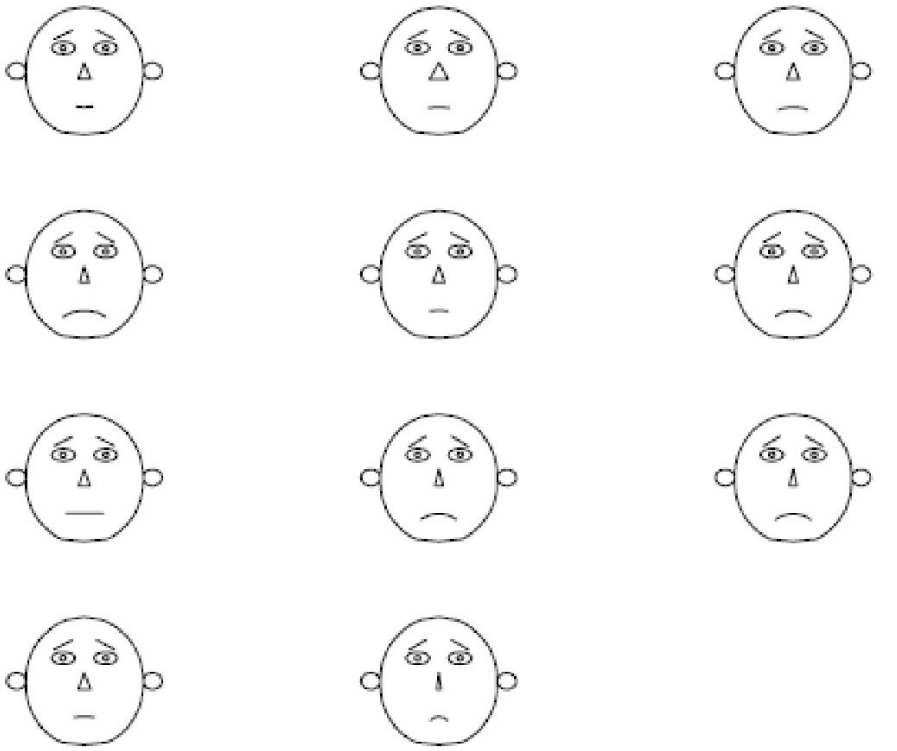
$$\sqrt{\sum_{j=1}^{p}(y_{1j}-y_{2j})^2}$$

Дистанции можно посчитать между объектами с любыми переменными, в т.ч. Качественными и даже бинарными!

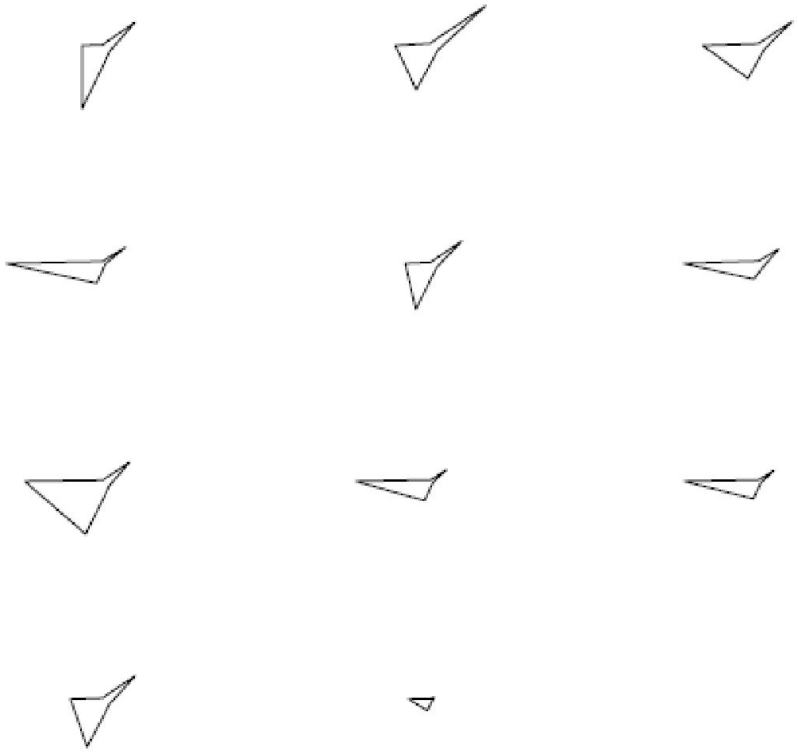
multivariate analyses
Подготовка данных для многомерного анализа

- ✔ Трансформация данных: нормализует распределения и делает отношения между переменными линейными (важно для выделения компонент). Логарифмическая, квадратного корня и пр.
- ✔ можно предварительно построить картинки и оценить сходство – различие между объектами (лица Чернова, «звёздный» график).
- ✔ важно избавить от аутлаеров! Многомерные аутлаеры: их можно найти с помощью дистанций Махаланобиса (квадрат расстояния от объекта до центроида).
- ✔ если переменные измерены в разных шкалах, принципиально использовать матрицу корреляций (не ковариаций) для получения компонент. Если нет – лучше пробовать оба варианта.
- ✔ пропущенные измерения – не casewise, a pairwise deletion.

multivariate analyses



Лица Чернова



«звёздный» график – star plot

ФАКТОРНЫЙ АНАЛИЗ

У нас в руках измерения большого числа переменных для выборки объектов.

Наши цели:

1. Уменьшить число исходных переменных с минимальными потерями исходной информации (что, например, уменьшит эффект множественных сравнений);
2. Обнаружить скрытые закономерности в данных, которые не выявляются при анализе отдельных переменных, путём помещения в пространство новых переменных (scaling). Например, выявление реальных действующих факторов (причинно-следственных связей), или просто выявление структуры взаимосвязи переменных.

Factor analysis

Анализ главных компонент (principal component analysis, PCA)

У нас есть ***n*** объектов и ***p*** переменных. Мы собираемся трансформировать переменные в ***k*** (от 1 до *p*) новых главных компонент = факторов.

Для каждого объекта мы получим значения этих компонент – z-значения.

В анализе – 6 этапов.

Factor analysis

Этап 0. Подготовка данных к анализу.

- ✔ Проверка распределений на соответствие нормальному;
- ✔ Трансформация данных (напр., логарифмирование некоторых переменных);
- ✔ Исключение аутлаеров.

Этап 1. Получение eigenvalues для новых компонент

В программе их получают из матрицы корреляций, их сумма = числу переменных. Разумно использовать компоненты, для которых eigenvalues > 1. т.е., число компонент будет меньше числа исходных переменных.

Напоминание: они независимы между сбой, т.е., ортогональны.

Этап 2. получение коэффициентов для каждой компоненты.

(Factor Score Coefficients). Они показывают вклад каждой переменной в компоненты. Необязательный этап.

Factor analysis

Этап 3. получение factor loadings

Это показатели корреляции (Пирсона) компонент с каждой из исходных переменных. Если какие-то переменные почти одинаково коррелируют с несколькими компонентами, можно улучшить структуру компонент:

Этап 4. вращение выбранных компонент для получения более чётких связей с исходными переменными. (чтобы loadings приблизились к 0, 1 или -1). Varimax rotation – самый распространённый и удобный метод.

Этап 5. получение factor loadings после вращения

Рассмотрение корреляций новых, повёрнутых компонент с исходными переменными, понимание их биологического смысла.

Этап 6. получение значений новых переменных для каждого объекта (для дальнейшего анализа.)

Factor analysis

Несколько слов о компонентах (факторах):

- ✔ В многомерном пространстве первая компонента располагается вдоль наибольшей дисперсии, т.е., это почти аналог линии линейной регрессии.
- ✔ Компоненты взаимно перпендикулярны
- ✔ Компоненты – линейные комбинации исходных переменных
- ✔ Если исходные переменные не коррелируют между собой, не получится собрать много дисперсии в первых компонентах, т.е., уменьшить их число.
- ✔ Сколько компонент оставлять? Это решает исследователь так, чтобы обеспечить биологическую интерпретируемость результатов. Нет смысла оставлять компоненты, с которыми не коррелирует сильно ни одна исходная переменная. Правило «eigenvalue =1».

Factor analysis

Вращение компонент (факторов)

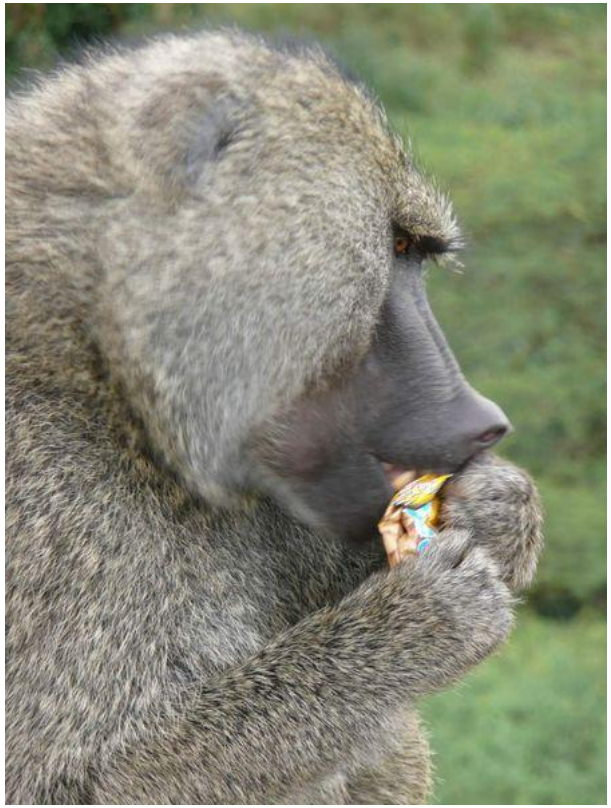
Выбранные нами факторы (их мало) поворачивают для получения более чёткой структуры переменных. Обычно используют ортогональное вращение – факторы остаются перпендикулярными друг другу. Например, varimax.

Не ортогональное вращение – oblique rotation, у него есть свои поклонники, но этот метод не прост.

Анализ остатков – residuals – имеет смысл посмотреть, насколько много информации мы потеряли при сокращении числа переменных. На основе наших факторов генерируются корреляции между исходными переменными и сравниваются с реальными корреляциями. Если разница где-то велика, мы взяли слишком мало факторов.

Factor analysis

Мы изучаем пищевые предпочтения павианов и разработали комплексные оценки привлекательности разных типов пищи для каждой особи.
Павианы едят разную еду, поэтому типов пищи – 10. особей в анализе – 100.
Но реальных факторов, определяющих эти предпочтения, наверняка меньше.



Мы хотим узнать, сколько (и каких) факторов определяют пищевые предпочтения павианов.

Factor analysis

Итак,

Мы хотим

Найти те факторы, которые определяют изменчивость (объясняют действие) большого количества измеренных нами реальных переменных.

Подразумевается, что таких факторов гораздо меньше, чем исходных переменных.

Factor analysis

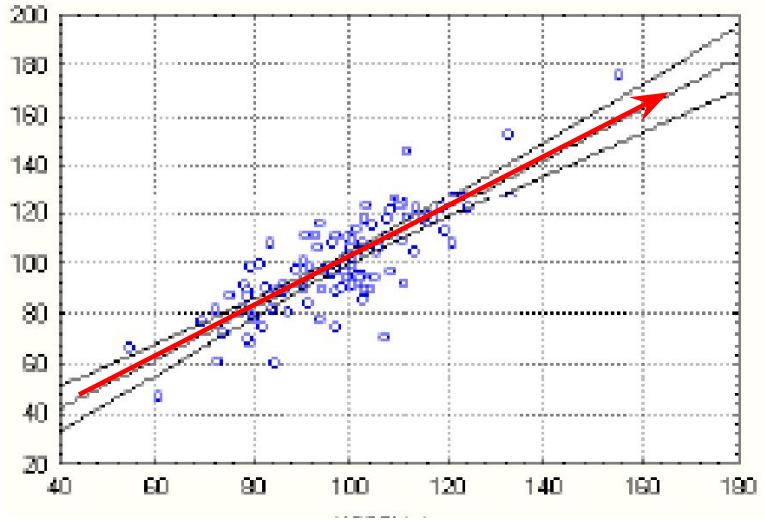
Поясняющий пример:

Мы изучаем кроликов. Сначала взвешиваем каждого из 100 кроликов на безмене, потом на весах с гирьками, потом на электронных кухонных весах.

Потом мы хотим исследовать влияние питания на вес кроликов.
Неужели мы возьмём в анализ все три переменные? Ведь, очевидно, вес кролика – только **одна** его характеристика, а не три. Скорее всего, мы захотим превратить все переменные в одну.

Factor analysis

Подразумевается, что наши реально измеренные переменные являются линейными комбинациями этих подлежащих факторов.



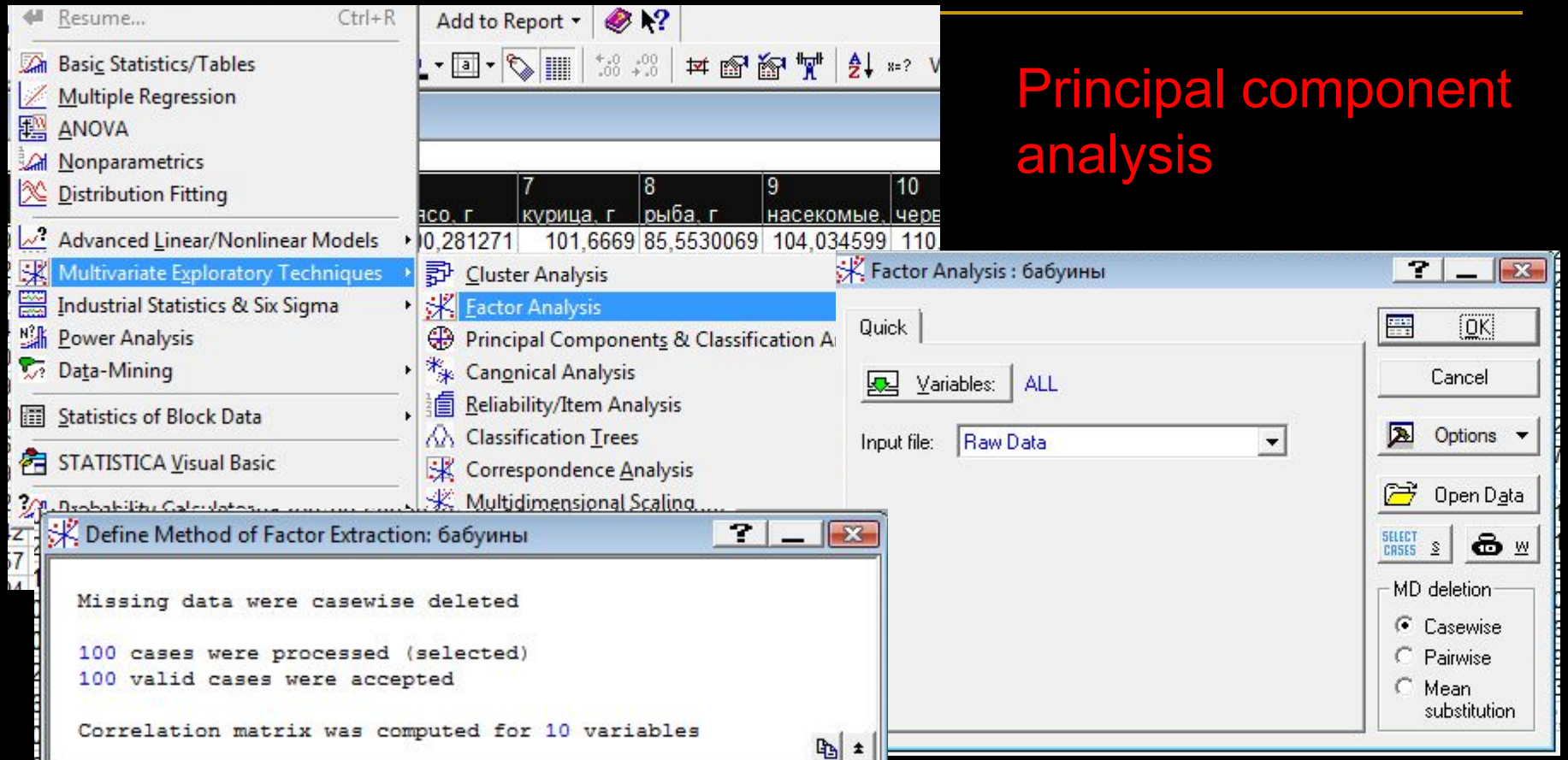
Примерно так будет проходить новая ось OX – первая компонента.

Итак, мы изучаем питание павианов. Типов пищи у павианов 10:

апельсины,
бананы,
яблоки,
помидоры,
огурцы,
мясо,
курица,
рыба,
насекомые,
червяки.

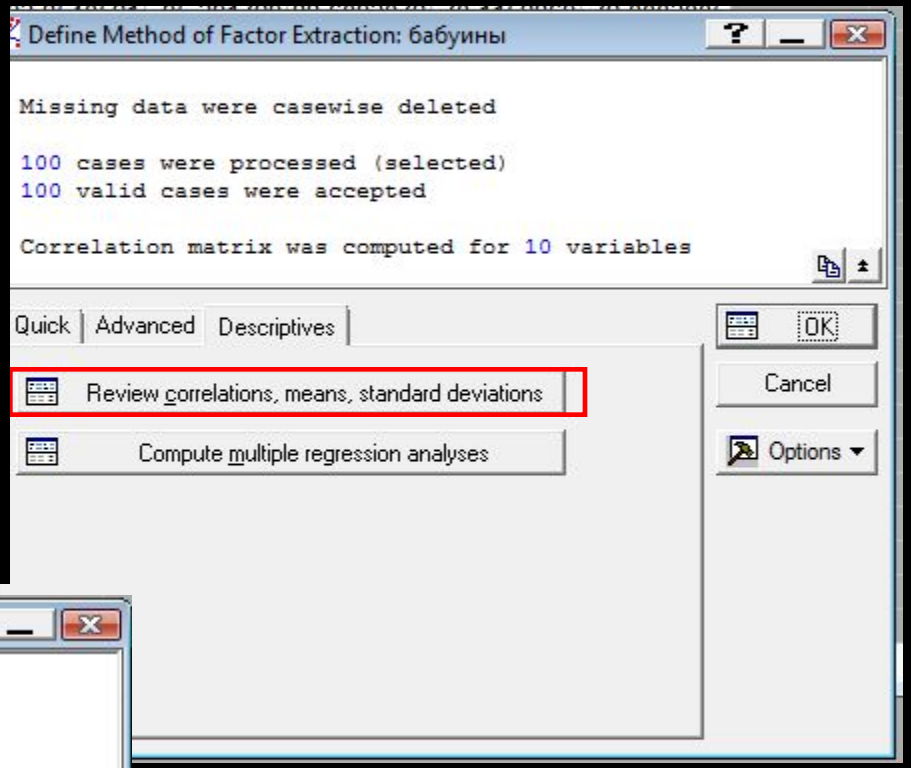
Мы измеряем привлекательность пищи каждого типа, для каждого зверя. Сколько факторов скрывается за разными предпочтениями павианов в еде?

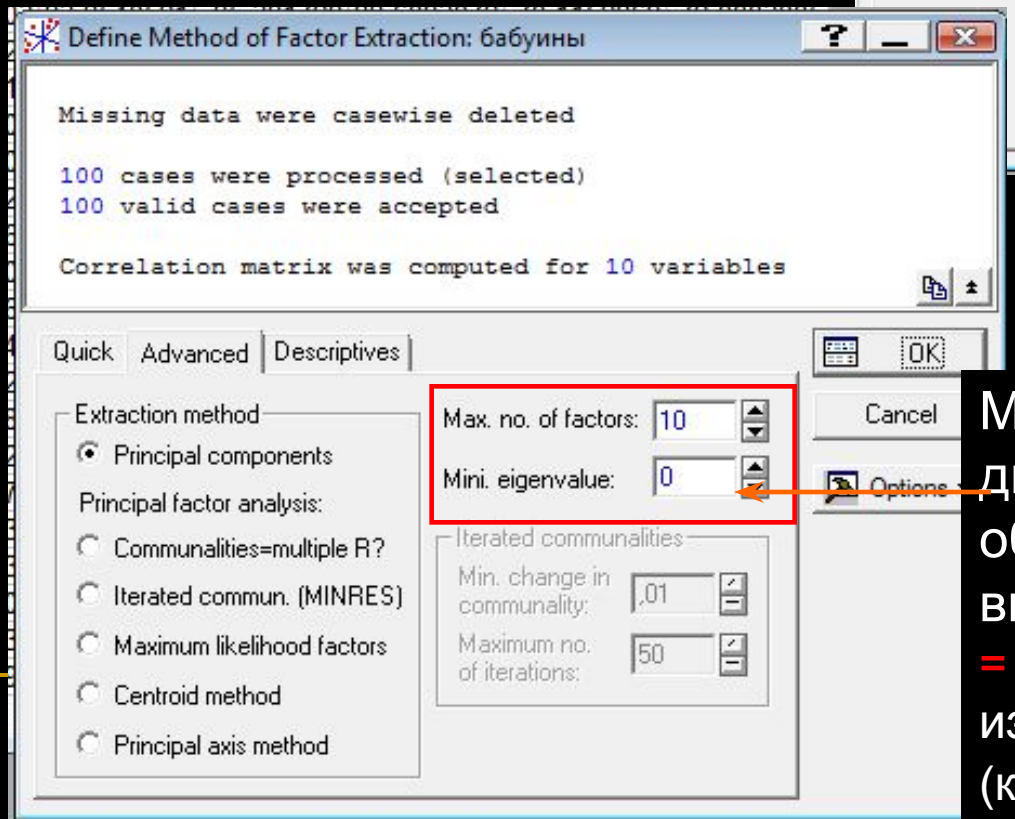
Principal component analysis



(прежде, чем проводить факторный анализ, рекомендуется построить матрицу корреляций: исключить переменные, слишком сильно коррелирующие с другими)

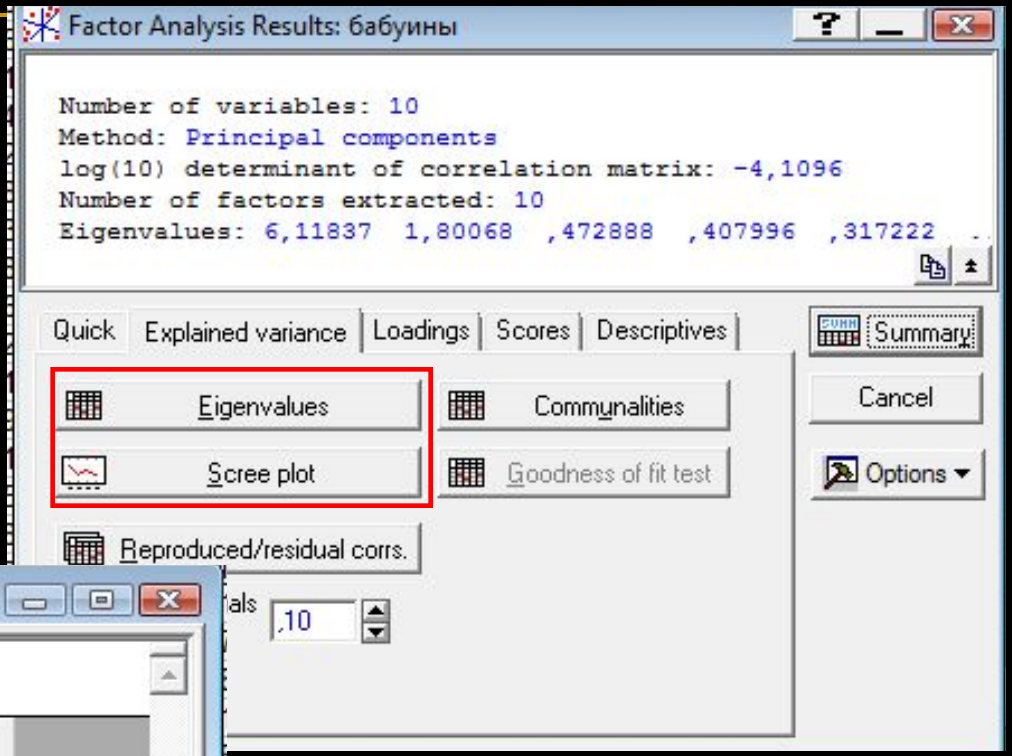
Просмотрим матрицу корреляций:
Не должно быть слишком сильно коррелирующих друг с другом переменных (иначе матрица не может быть транспонирована: *matrix ill-conditioning*)

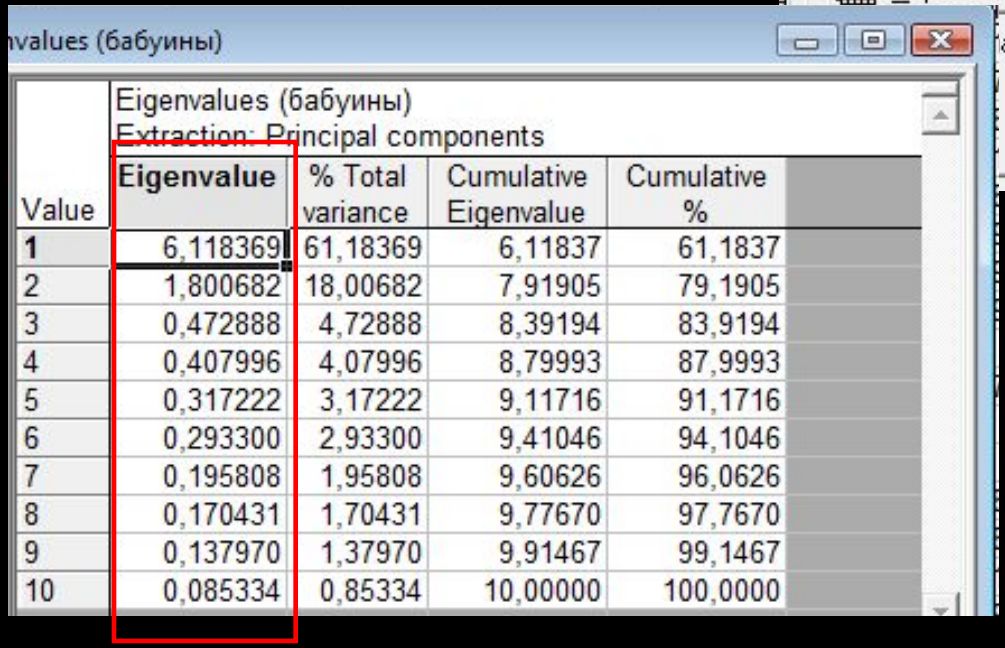




Можно задать min количество дисперсии, которое должен объяснять фактор, чтобы его включили в анализ (обычно min = 1, что соответствует случайной изменчивости одной переменной (критерий Кайзера))

Собственные значения (eigenvalues)– определяют, какую долю общей дисперсии объясняет данный фактор.

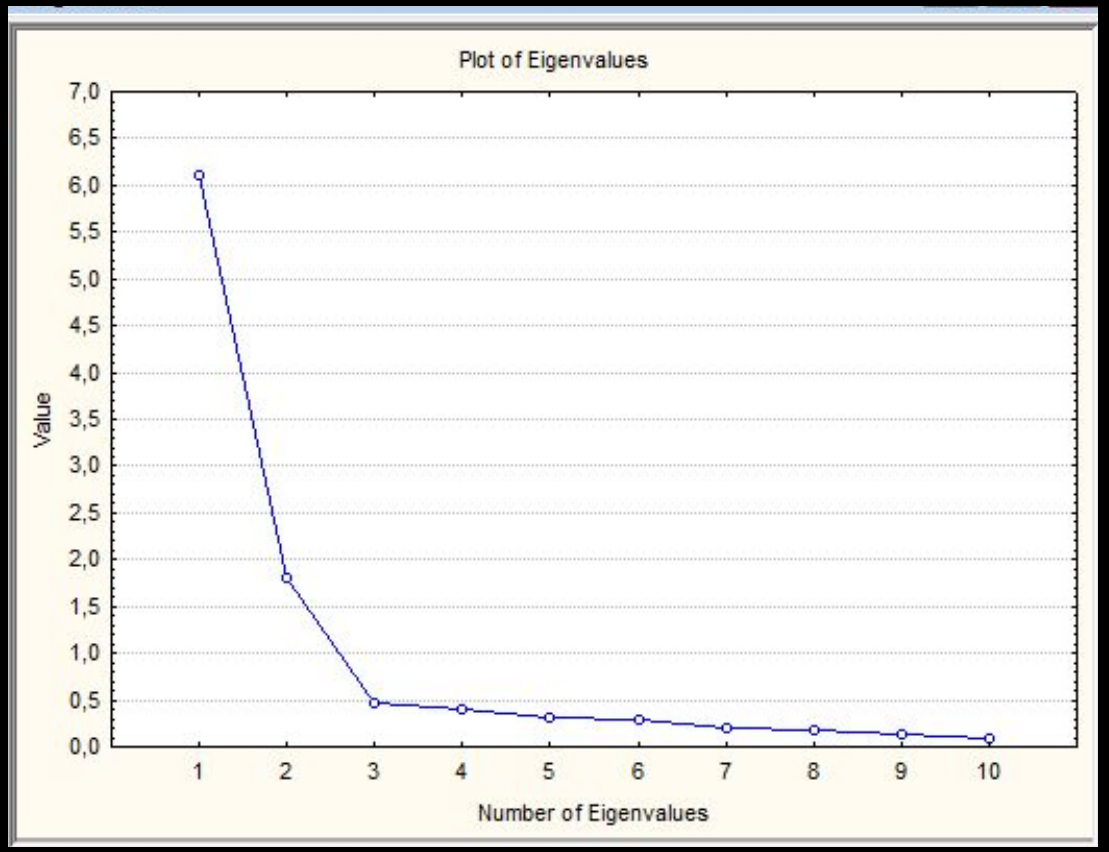




nvalues (бабуины)

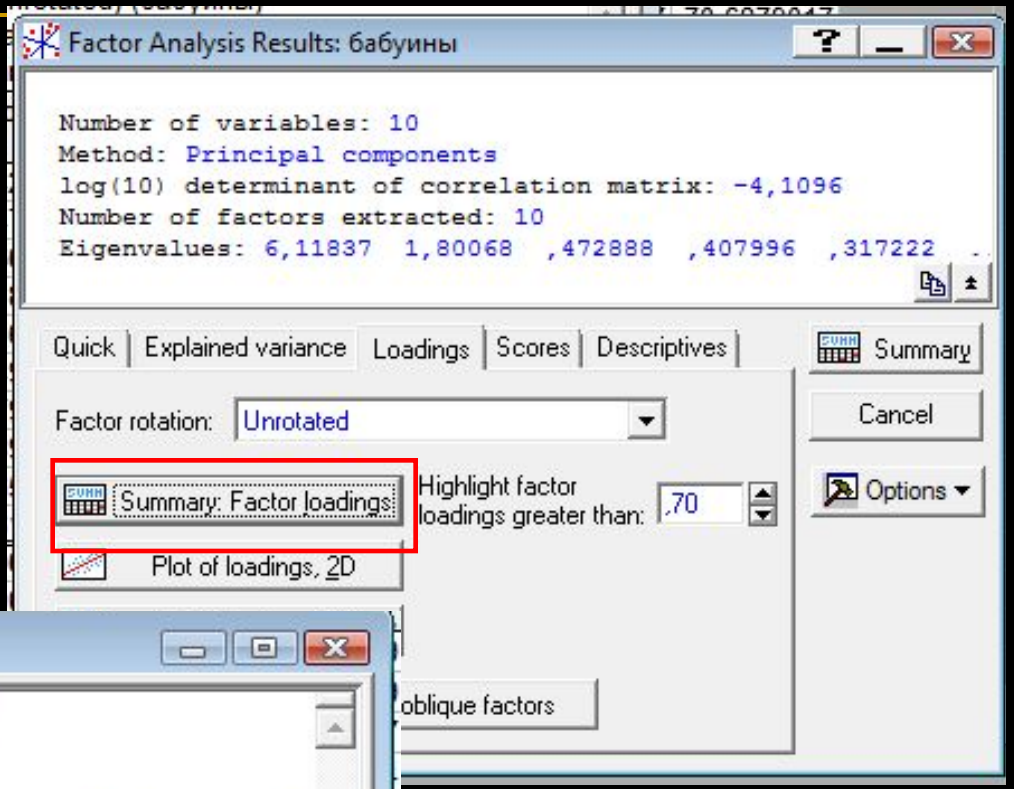
Eigenvalues (бабуины)
Extraction: Principal components

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	6,118369	61,18369	6,11837	61,1837
2	1,800682	18,00682	7,91905	79,1905
3	0,472888	4,72888	8,39194	83,9194
4	0,407996	4,07996	8,79993	87,9993
5	0,317222	3,17222	9,11716	91,1716
6	0,293300	2,93300	9,41046	94,1046
7	0,195808	1,95808	9,60626	96,0626
8	0,170431	1,70431	9,77670	97,7670
9	0,137970	1,37970	9,91467	99,1467
10	0,085334	0,85334	10,00000	100,0000



Этот график показывает, что первые два фактора лучше остальных, они объясняют большую часть общей изменчивости (the scree test).

Посмотрим, как полученные факторы связаны с реальными переменными

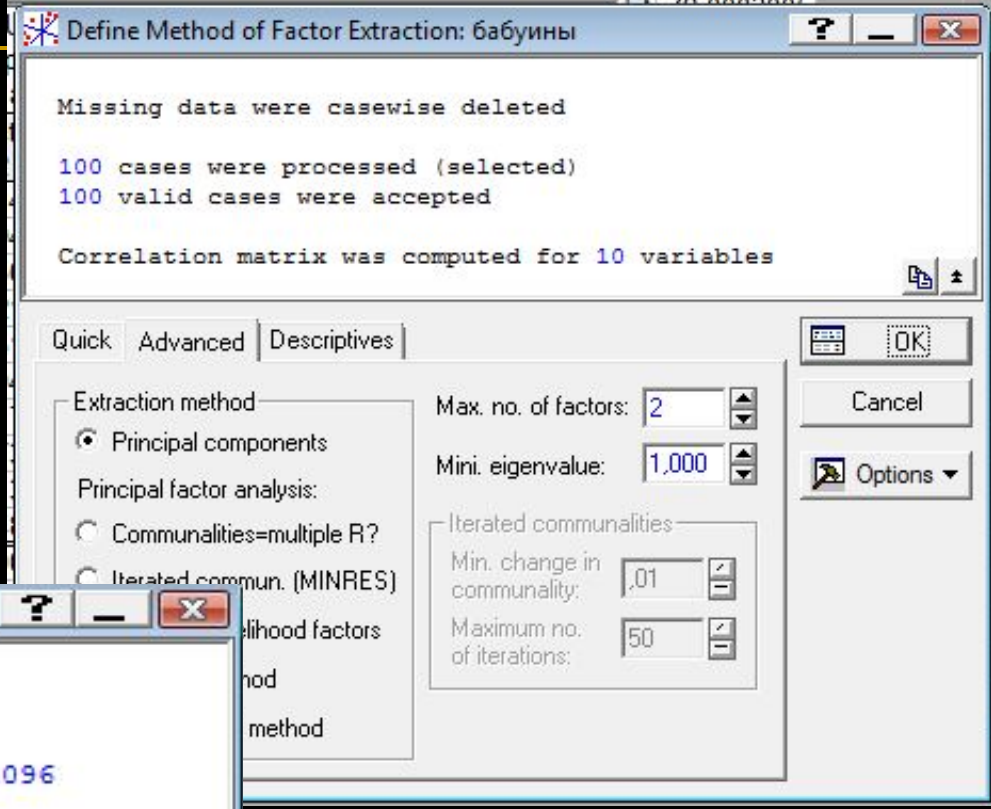


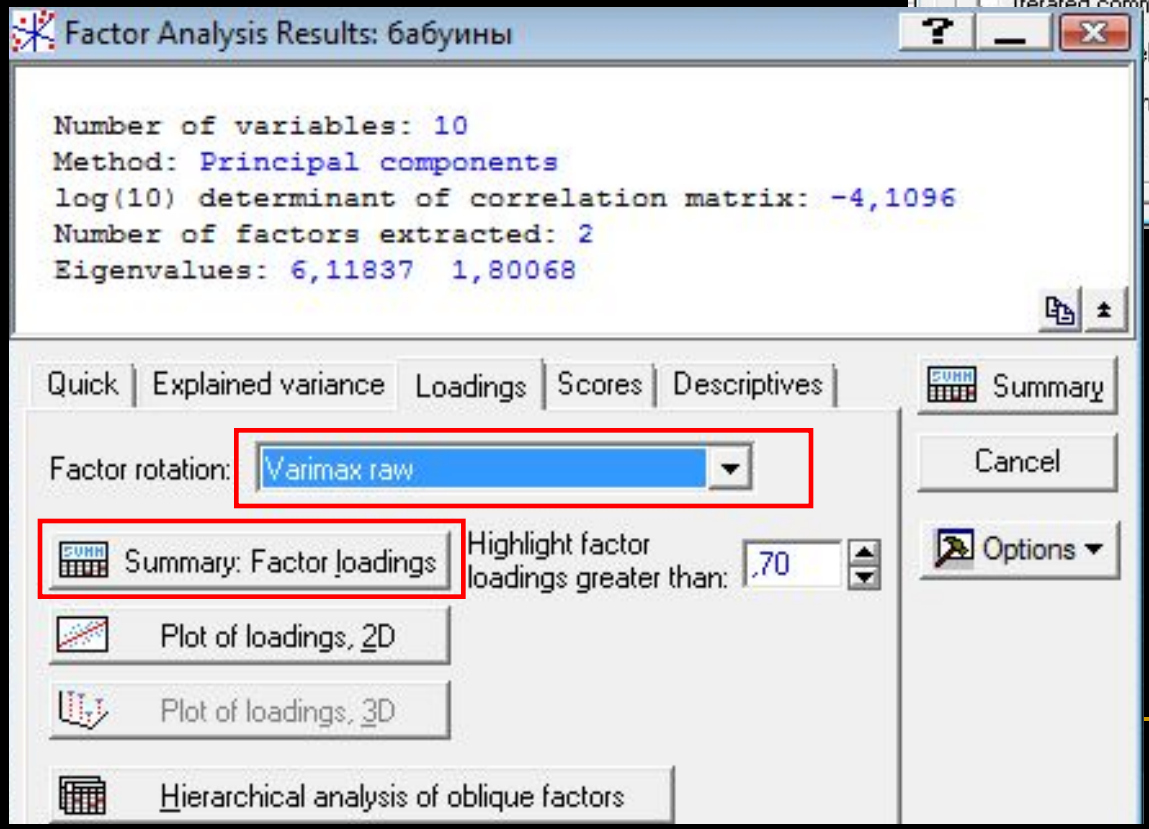
tor Loadings (Unrotated) (бабуины)

Factor Loadings (Unrotated) (бабуины)
Extraction: Principal components
(Marked loadings are > ,700000)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Fa
апельсины, г	-0,652601	0,514217	0,301687	0,439108	-0,013701	0,1
бананы, г	-0,756976	0,494770	-0,078826	-0,211795	-0,090859	0,1
яблоки, г	-0,745706	0,456680	-0,104749	0,030826	-0,204913	-0,4
помидоры, г	-0,941630	-0,021835	0,012653	0,001861	0,120655	0,0
огурцы, г	-0,875615	0,051643	0,099675	-0,324541	-0,015852	0,0
мясо,г	-0,576062	-0,604977	0,490999	-0,114927	-0,112513	-0,1
курица, г	-0,671289	-0,617962	-0,125776	0,159963	0,225012	-0,1
рыба,г	-0,641532	-0,573925	-0,268572	0,152709	-0,362524	0,1
насекомые, г	-0,951516	0,013513	-0,050164	0,026706	0,076795	0,0
червяки, г	-0,900333	0,048154	-0,151805	-0,034832	0,226647	-0,0
Expl.Var	6,118369	1,800682	0,472888	0,407996	0,317222	0,2
Prp.Totl	0,611837	0,180068	0,047289	0,040800	0,031722	0,0

оставим две компоненты и проведём вращение, чтобы улучшить их структуру.





После вращения факторов их структура становится более ясной:

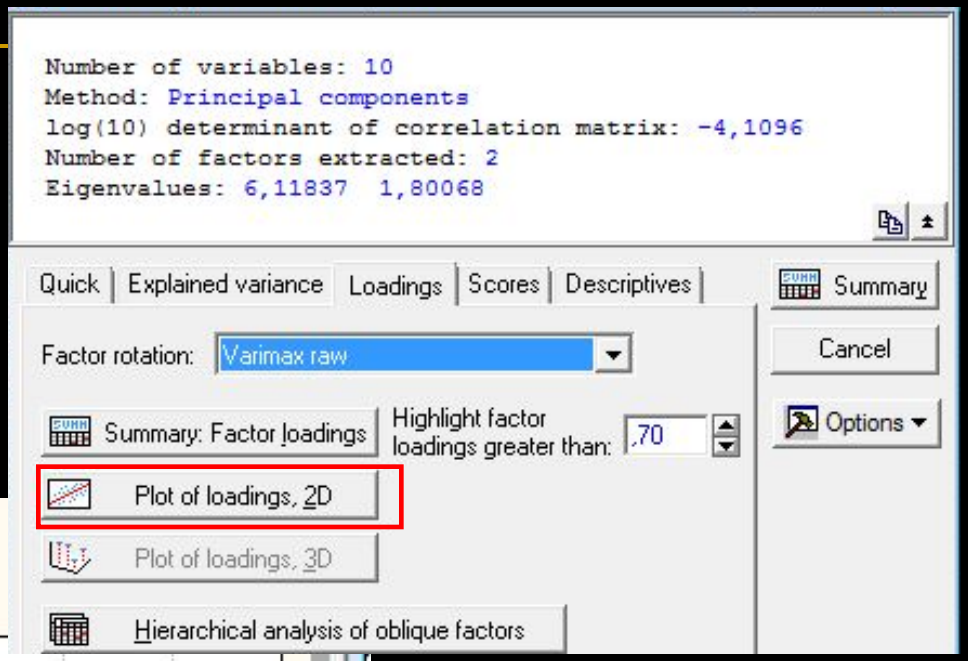
r Loadings (Varimax raw) (бабуины)

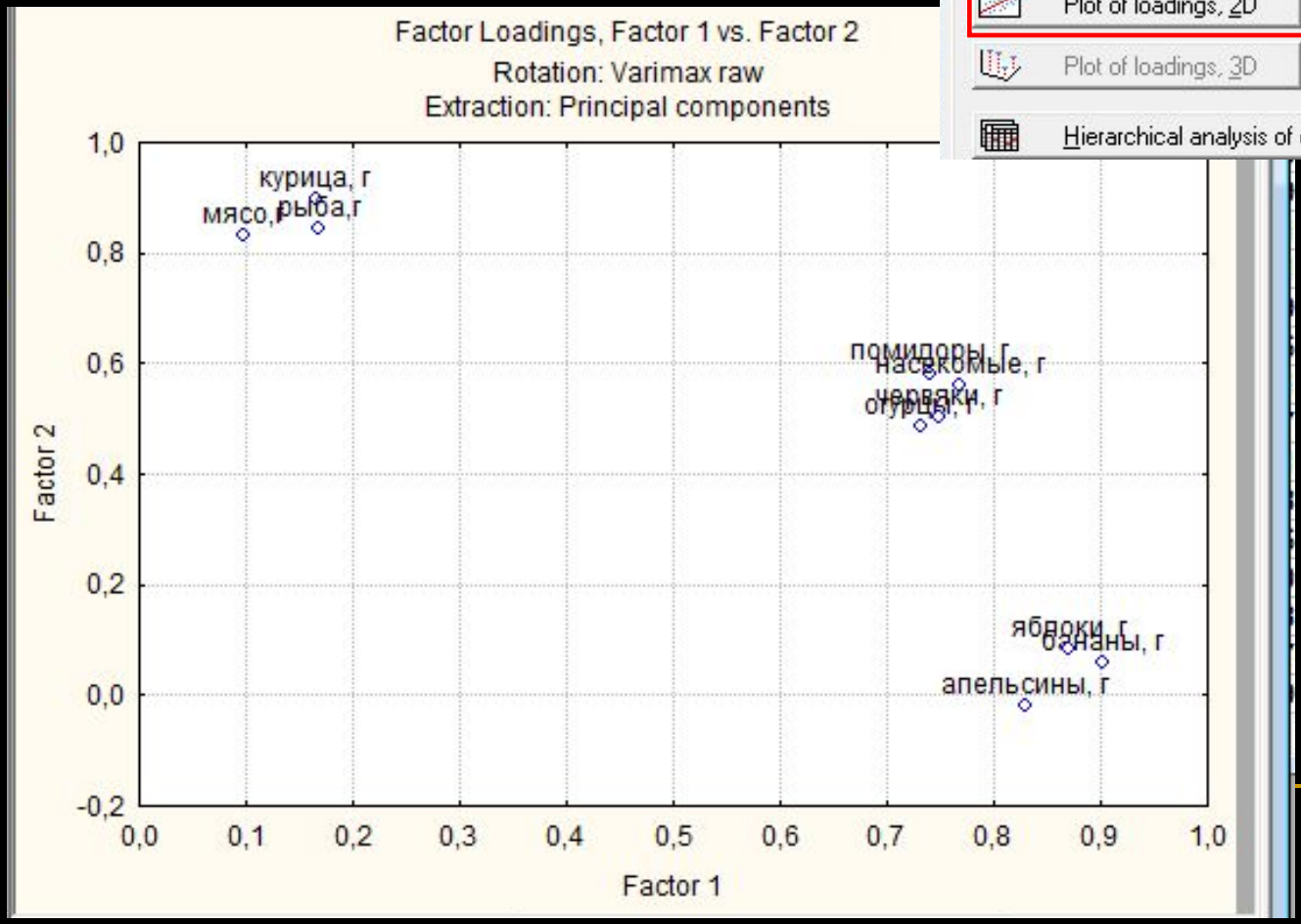
Variable	Factor Loadings (Varimax raw) (ба Extraction: Principal components (Marked loadings are > ,700000) Factor 1	Factor 2
апельсины, г	0,830623	-0,019320
бананы, г	0,902408	0,058905
яблоки, г	0,870524	0,082595
помидоры, г	0,739857	0,582885
огурцы, г	0,731191	0,484489
мясо,г	0,097371	0,829676
курица, г	0,165722	0,897242
рыба,г	0,168370	0,844159
насекомые, г	0,768988	0,560555
червяки, г	0,748861	0,502121
Expl.Var	4,561544	3,357507
Prp.Totl	0,456154	0,335751

Фактор 1 в основном связан с растительной пищей, фактор 2 – с животной.

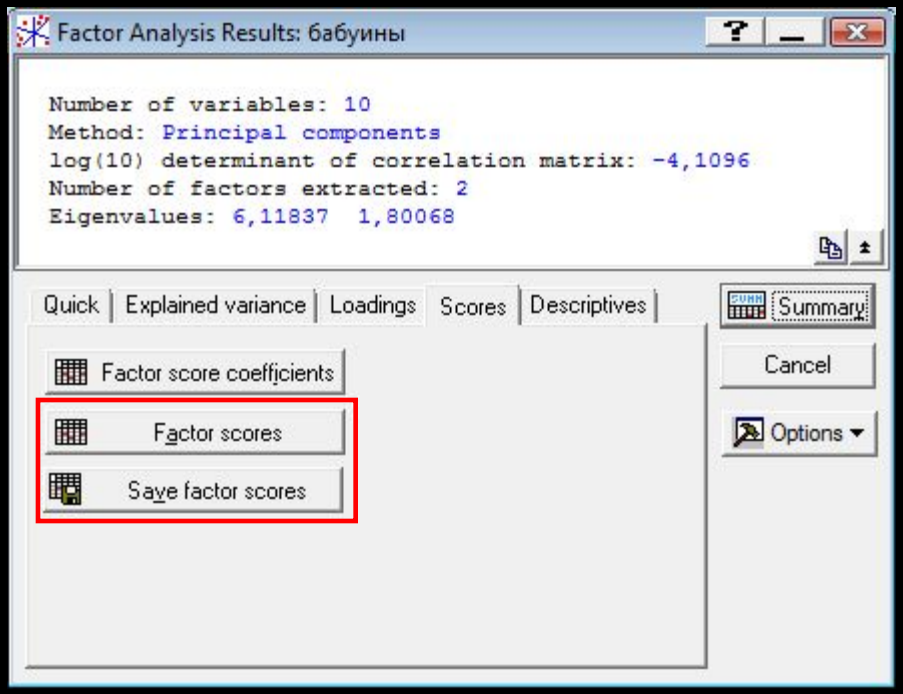
Итак, пищевые предпочтения павианов составлены из двух основных факторов – отношением к животной и растительной пище.

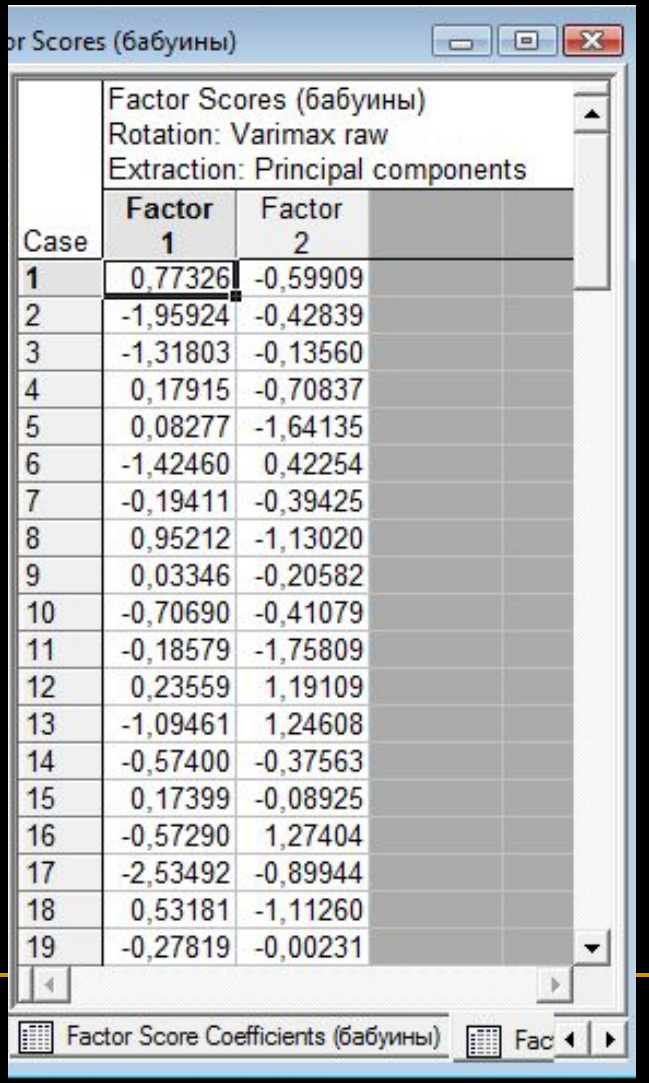
Посмотрим, как исходные переменные расположились в пространстве новых факторов



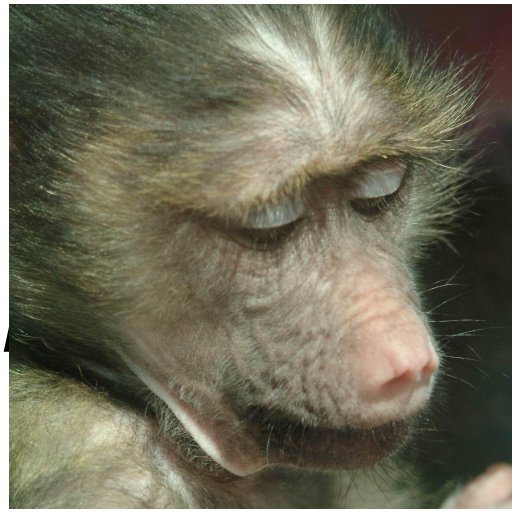


Если мы в дальнейшем хотим проводить анализ связи питания павианов с другими переменными, мы можем заменить наши 10 переменных на полученных два фактора.





Требования к выборкам для проведения факторного анализа



1. Внутри групп должно быть многомерное *распределение* (оценка – на основе гистограмм частот);
2. Гомогенность *дисперсий* (для метода главных компонент; не очень критичное требование);
3. Связь переменных должна быть *линейной*;
4. Размер выборки не должен быть меньше 50, оптимальный – ≥100 наблюдений.
5. Между переменными должна быть *ненулевая корреляция*, но коэффициентов корреляции, близких *единице*, тоже быть не должно.

Factor analysis

Связь с MANOVA и регрессионным анализом.

1. Если мы на самом деле хотим сравнить группы (из объектов с многими переменными) можно провести MANOVA (это тоже многомерный анализ, но он генерирует только одну переменную), а можно сначала факторный анализ, а потом – однофакторные ANOVA (у второго варианта есть преимущества).
2. Если мы хотим провести множественный регрессионный анализ, можно сначала сделать факторный анализ для независимых переменных (можно - без сокращения их числа), а потом – регрессионный анализ, убрав проблему скоррелированности исходных переменных.

Другие многомерные методы, близкие анализу главных компонент

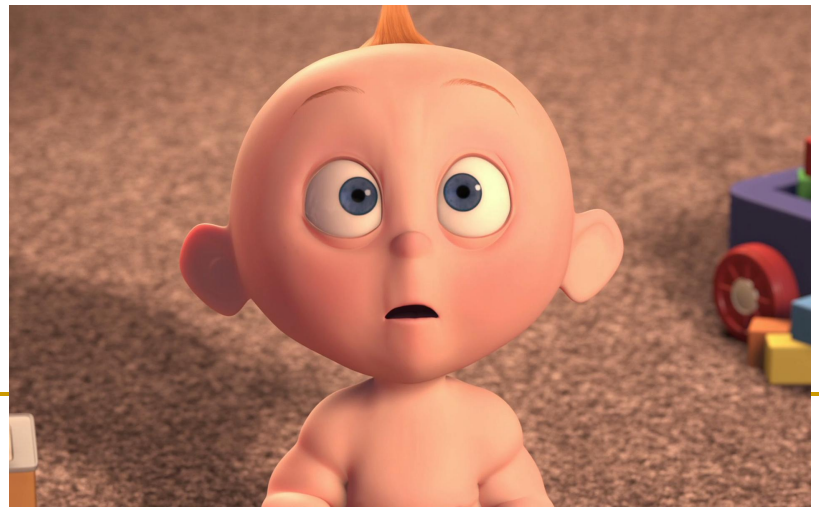
1. Principal factor analysis – если PCA генерирует компоненты, объясняющие изменчивость исходных переменных, то PFA генерирует common factors, объясняющие корреляции между переменными.
2. Correspondence analysis – для анализа таблиц сопряжённости (большого числа качественных переменных) . Сумма eigenvalues = общей статистике χ^2 (называется total intertia).
3. Canonical correlation analysis – если у нас есть два блока переменных и мы хотим анализировать корреляции между ними. Генерирует пары переменных из этих блоков (canonical variates) так, чтобы между ними была максимальная корреляция.

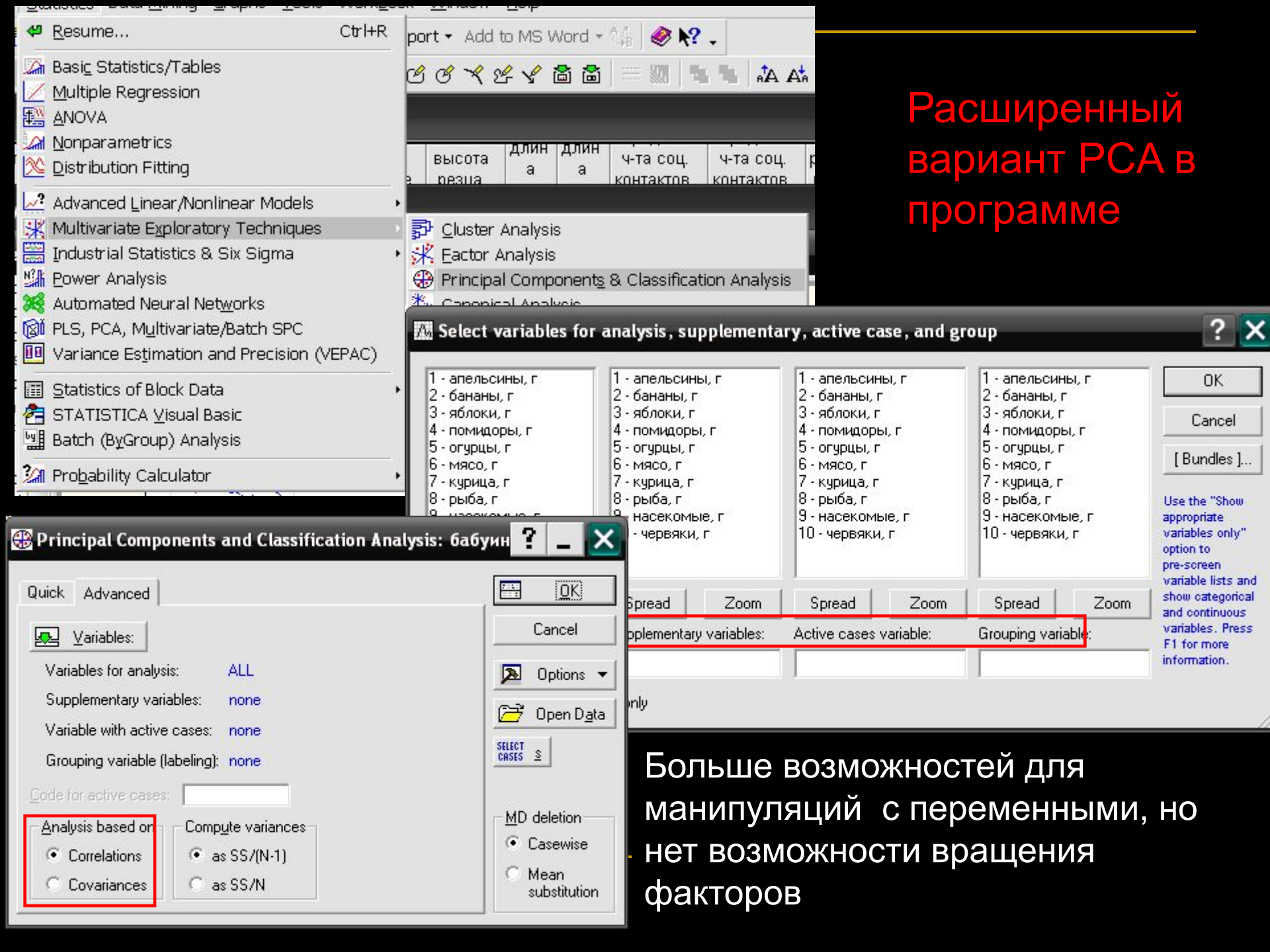
communality of a variable is the portion that can be reproduced from the respective number of factors

Другие многомерные методы, близкие анализу главных компонент

4. Redundancy analysis – усложнённая версия Canonical correlation analysis, предсказывает линейную комбинацию зависимых переменных из комбинации независимых.
5. Canonical correspondence analysis – расширенный вариант Correspondence analysis, в котором дополнительно учитывается влияние добавочных количественных переменных.

На свете много многомерных методов!





Расширенный вариант PCA в программе

Больше возможностей для манипуляций с переменными, но нет возможности вращения факторов