Занятие 7

Основы многомерных методов анализа. Факторный анализ.

Методы многомерного анализа (multivariate analysis)

Предназначены для анализа многомерных данных

Многомерные данные: несколько переменных регистрируются для каждого объекта в выборке (индивидуума, особи, образца, ...

Много **независимых переменных** –

✓Многофакторная ANOVA

✓Множественная регрессия

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

✓multivariate analysis

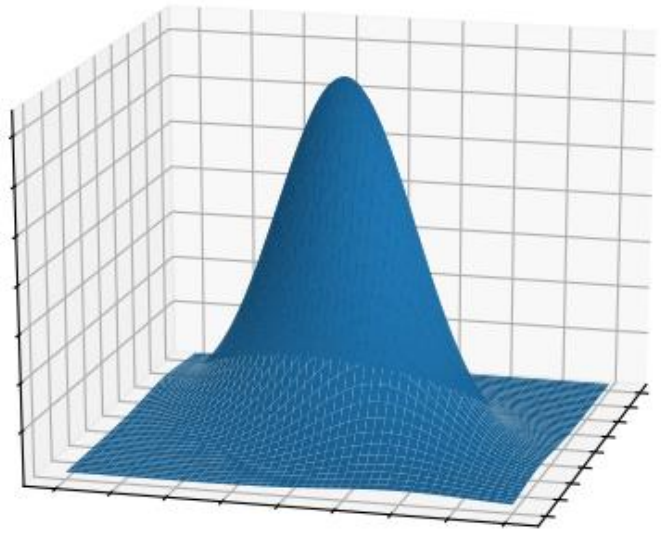
Рассмотрим ситуации, когда проверяется влияние одной или нескольких независимых переменных на несколько зависимых переменных.

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

Описание многомерных данных

1. **Многомерное распределение**
- распределение многомерных данных

При тестировании гипотез в многомерном анализе требуется многомерное нормальное распределение (все переменные и их линейные комбинации имеют нормальное распределение).

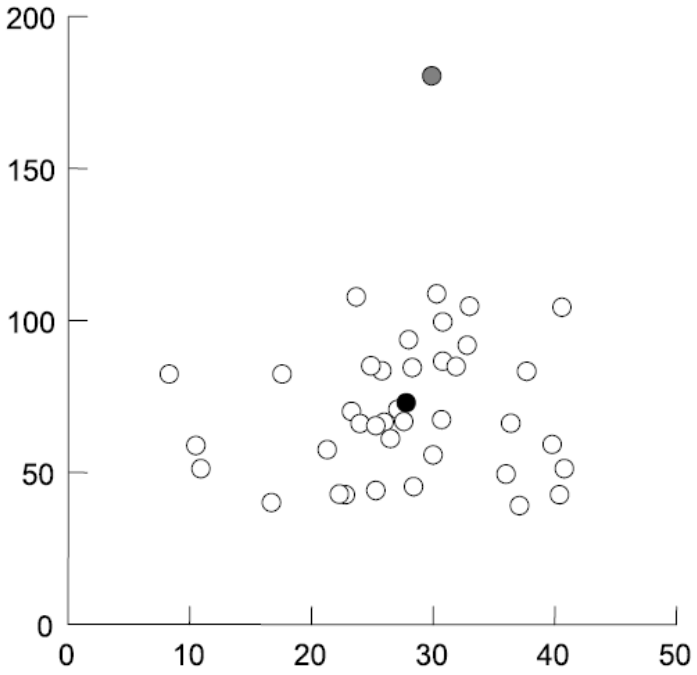


Чем больше отклонение от многомерного нормального распределения, тем больше будет неточности в оценке параметров (коэффициентов и пр.)

2. Показатель «середины» распределения

В одномерном анализе т.е. для одной переменной – *среднее значение.*

Для многомерного распределения – **центроид** (точка, координаты который – средние значения для каждой переменной).



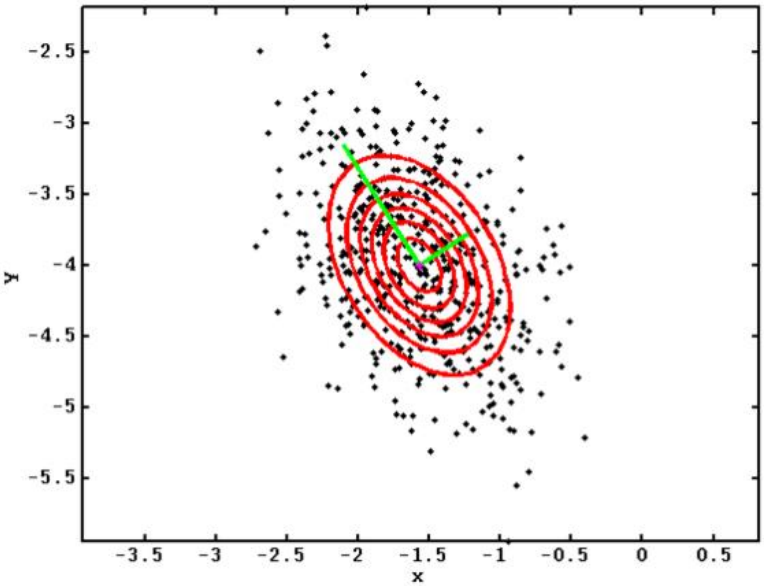
Для каждого объекта можно посчитать его «расстояние» до центроида (дистанция Махаланобиса).

3. Показатели разброса распределения

В одномерном распределении – *сумма квадратов отклонений (SS), дисперсия, стандартное отклонение*.

В многомерных данных два источника изменчивости:

а) изменчивость внутри самих переменных;
б) изменчивость, обусловленная взаимным влиянием переменных.



Поэтому, изменчивость в многомерных данных представляется сложно – в виде таблицы (**матрицы**).

Несколько слов о матрицах (основе многомерного анализа)

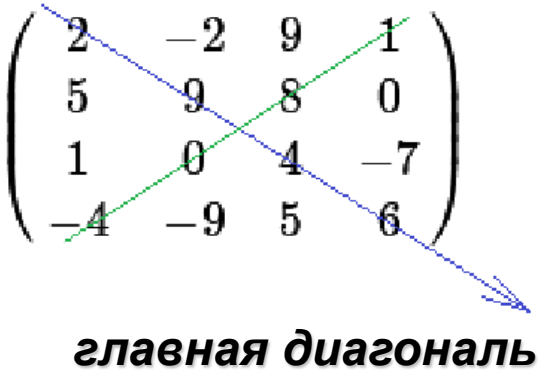
<u>Матрицы</u> - это прямоугольные таблицы, которые состоят из чисел (элементов).

В матрицах есть строки и столбцы (нумеруются слева-направо, сверху-вниз)

m x n матрица – прямоугольная;
n x n – квадратная;

Каждый элемент имеет № строки и столбца, в которых он стоит.

У квадратных матриц есть диагональ.



С матрицами можно производить всякие <u>действия</u>:
<u>*менять*</u> строки/столбцы местами, умножать на число, прибавлять число; <u>*складывать*</u> и <u>*умножать*</u> матрицы, переворачивать относительно диагонали <u>*(транспонировать)*</u>.

Таблицу исходных данных можем рассматривать как матрицу (есть столбцы и строки) – <u>матрица исходных данных</u> (Y).

Иследователи 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

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdots & \cdots & y_{ij} & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

Для описания изменчивость многомерных данных, нам понадобится матрица, т. к. нам надо показать и изменчивость внутри переменных, и их взаимодействие – каждой переменной с каждой.

Матрица (*p* x *p*) с суммами квадратов на диагонали (sums-of-squares-and-cross-products, SSCP) (неудобна, т.к. сильно зависит от абсолютных значений):

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

На главной диагонали – SS, остальные – произведения отклонений в парах переменных

Основные матрицы

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

- ✓ на главной диагонали стоят дисперсии для каждой переменной – показатели разброса внутри переменных;
- ✓ остальные элементы – ковариации (covariances, C) между переменными – показатели взаимосвязи между переменными.

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

$$s_1^2 = \frac{\sum_{i=1}^{n}(Y_{i1} - \overline{Y}_1)^2}{n-1}$$

$$s_{12}^2 = \frac{\sum_{i=1}^{n}(Y_{i1} - \overline{Y}_1)^2(Y_{i2} - \overline{Y}_2)^2}{n-1}$$

Дисперсия 1-й переменной

Ковариация 1-й и 2-й переменных

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.

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

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

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.

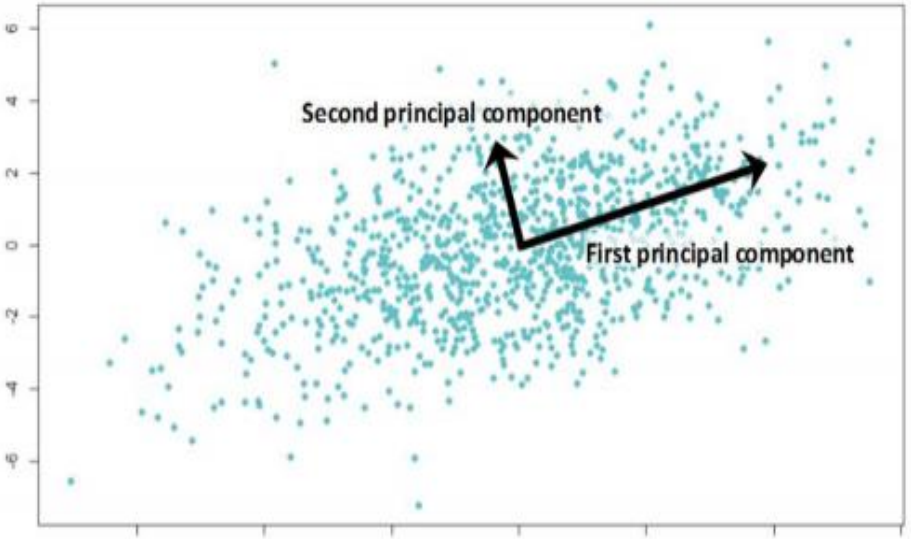
Новые переменные называются дискриминантными функциями, каноническими функциями, главными компонентами (principal components) или факторами (в зависимости от типа анализа).

Линейная комбинация аналогична уравнению линейной регрессии.

Свойства новых переменных

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.



У новых переменных есть:

Собственное значение ( ) = **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.

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

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)

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):

Variables

Objects	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.				

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!

Подготовка данных для многомерного анализа

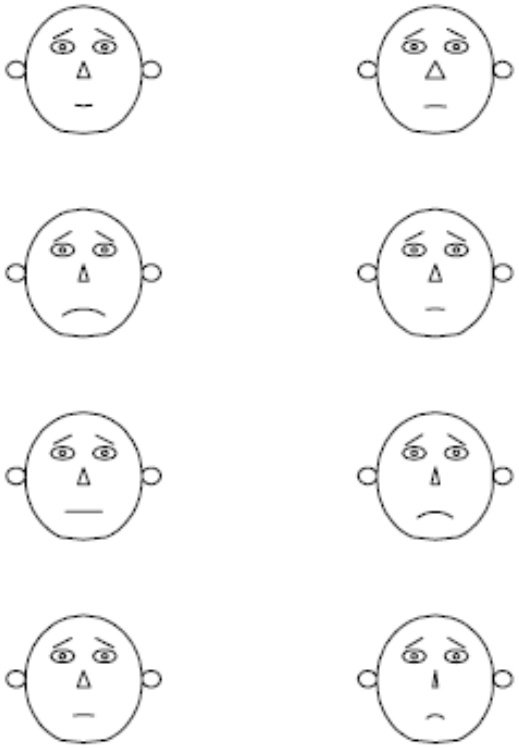
✓Трансформация данных: нормализует распределения и делает отношения между переменными линейными (важно для выделения компонент). Логарифмическая, квадратного корня и пр.

✓ можно предварительно построить картинки и оценить сходство – различие между объектами (лица Чернова, «звёздный» график).

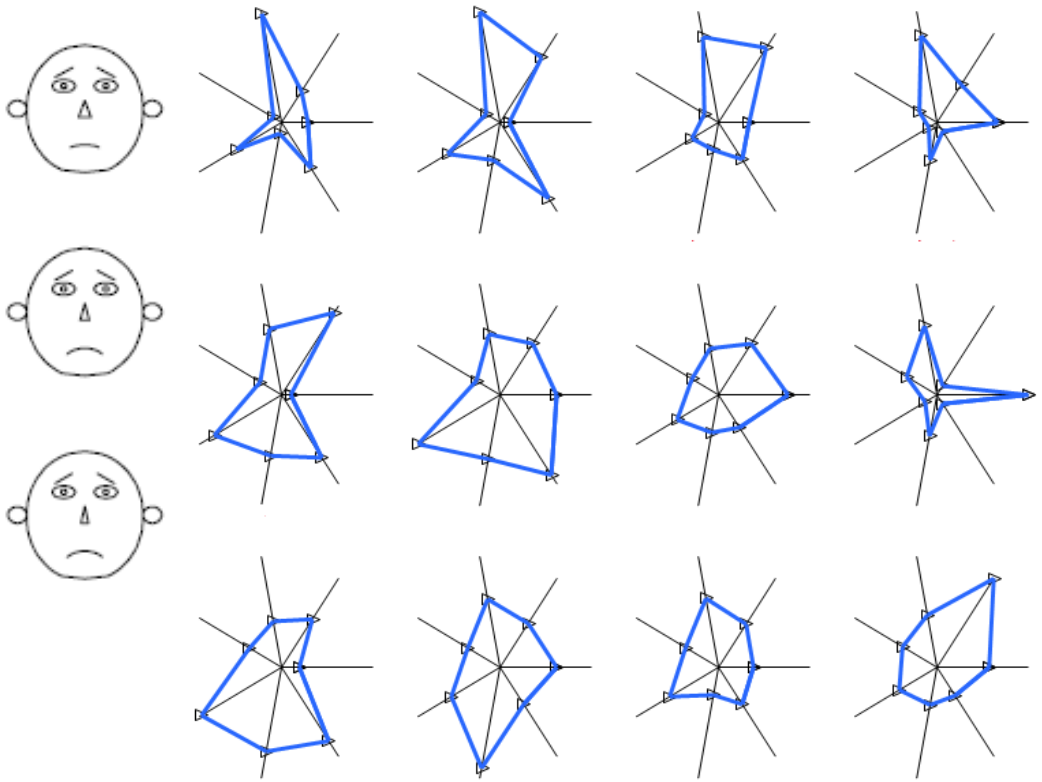
✓ важно избавить от аутлаеров! Многомерные аутлаеры: их можно найти с помощью дистанций Махаланобиса (квадрат расстояния от объекта до центроида).

✓ 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, а pairwise deletion.



Лица Чернова



«звёздный» график – star plot

Сравнение групп объектов

Мы имеем *МНОГОМЕРНЫЕ* данные. Они классифицируются на *ГРУППЫ* (какой-то группирующей переменной или переменными).

Как сравнить группы между собой?

Если зависимая *переменная одна*, используем *ANOVA*.

Почему бы не выполнить отдельные дисперсионные анализы для каждой из переменных?

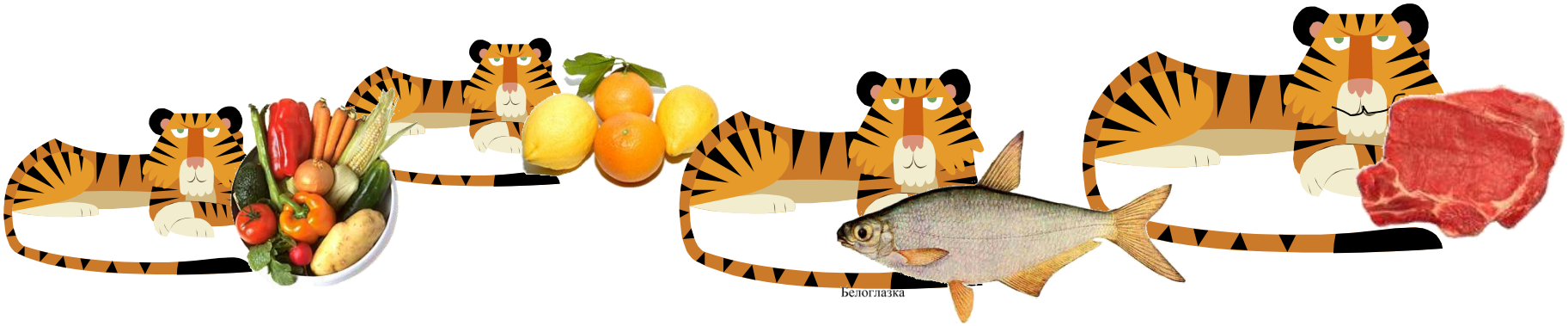
1. Вероятность ошибки 1-го рода превысит 5%;
2. Не будет учтена возможная корреляция между переменными;
3. Средние различия групп по каждой переменной могут быть малы, но по всем переменным совместно различия могут быть очевидными.

Multivariate factorial ANOVA = MANOVA

(многомерный дисперсионный анализ)

Например, нам интересно, как вид пищи (фактор) влияет на массу тигров и длину тела тигров (две зависимые переменные).

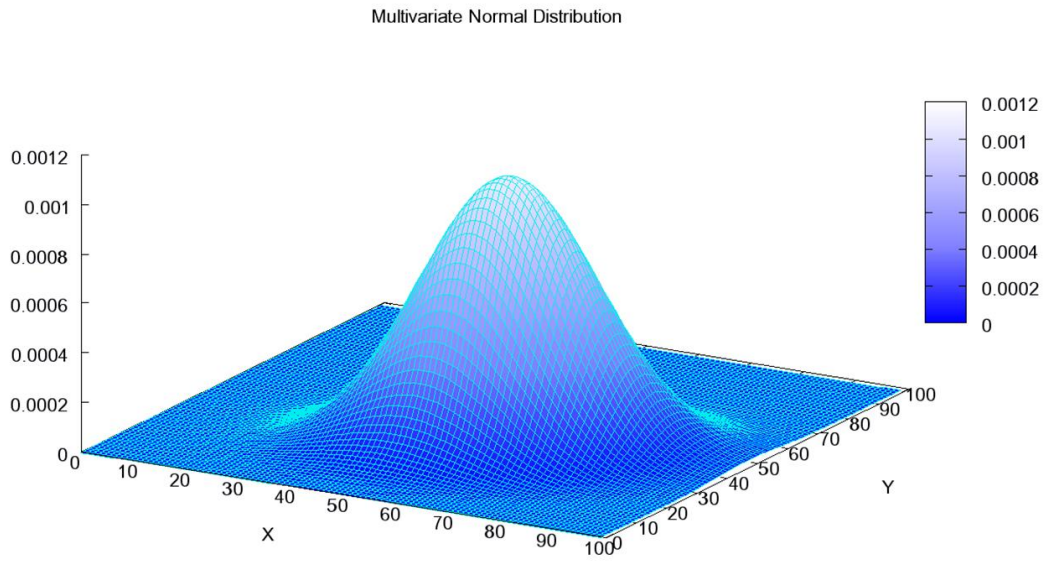
H_0: о влиянии группирующей переменной на комбинацию зависимых переменных = о равенстве центроидов в группах



Multivariate ANOVA = MANOVA
многомерный дисперсионный анализ

Multiway ANOVA
Многофакторный дисперсионный анализ

Предполагается многомерное нормальное распределение.



Нулевая гипотеза одна: о равенстве средних значений по одной из переменных И по другой переменной (и по третьей, четвёртой и т.д.)

Существует несколько вариантов статистики критерия для MANOVA.

Все они считаются на основе:

SS_{total}, $SS_{between\ groups}$, $SS_{within\ groups}$

$$\sum_i \sum_j (X_i - \bar{X})(Y_j - \bar{Y})$$

и сумм «векторных произведений»

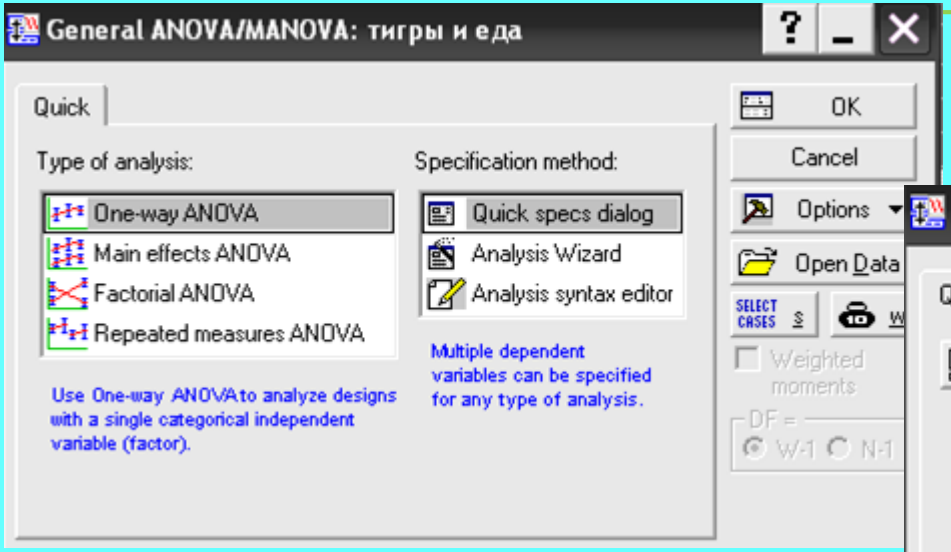
(это понятие уже из регрессионного анализа).

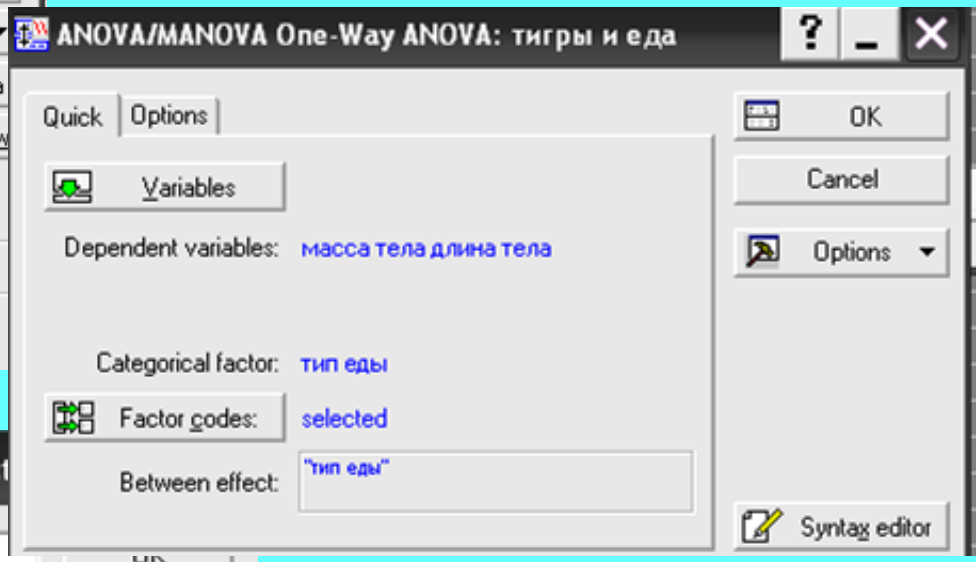
Wilks' lambda λ. Чем она меньше, тем вероятнее отвергнуть H_0. Это мера изменчивости, которая <u>не объясняется</u> действием факторов;

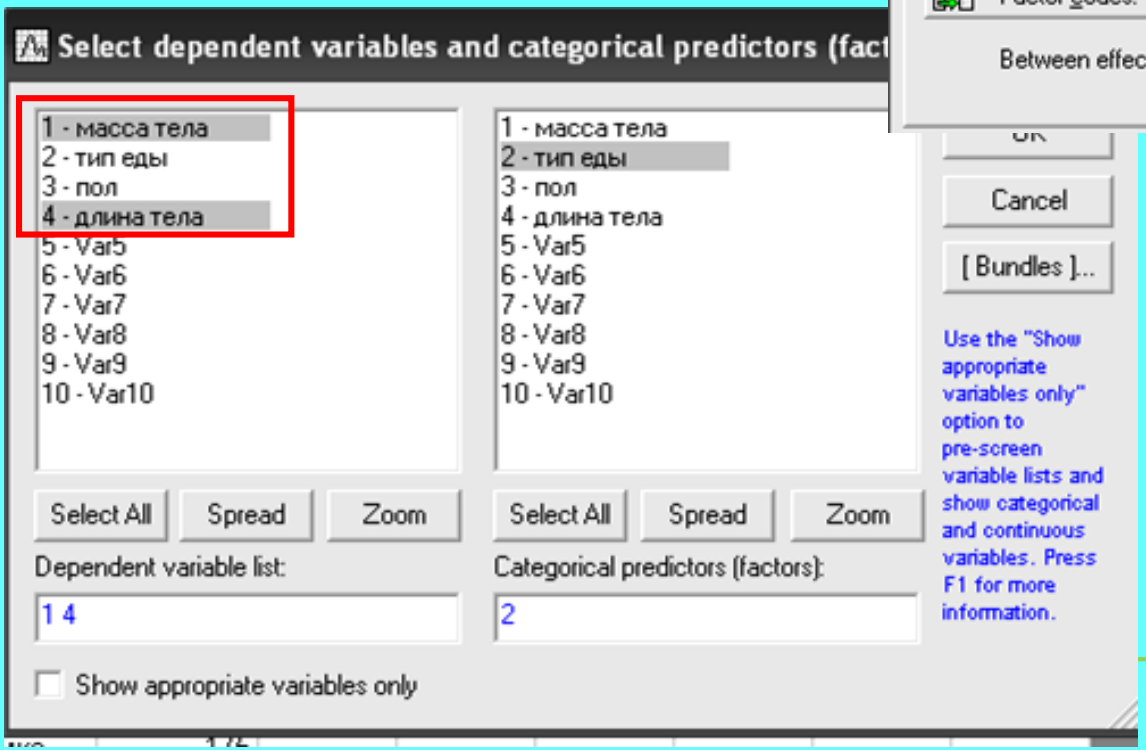
Hotelling trace – чем больше, тем больше различия групп;

Pillai's trace наиболее <u>*устойчив*</u> к отклонениям от многомерного нормального распределения и гомогенности дисперсии.

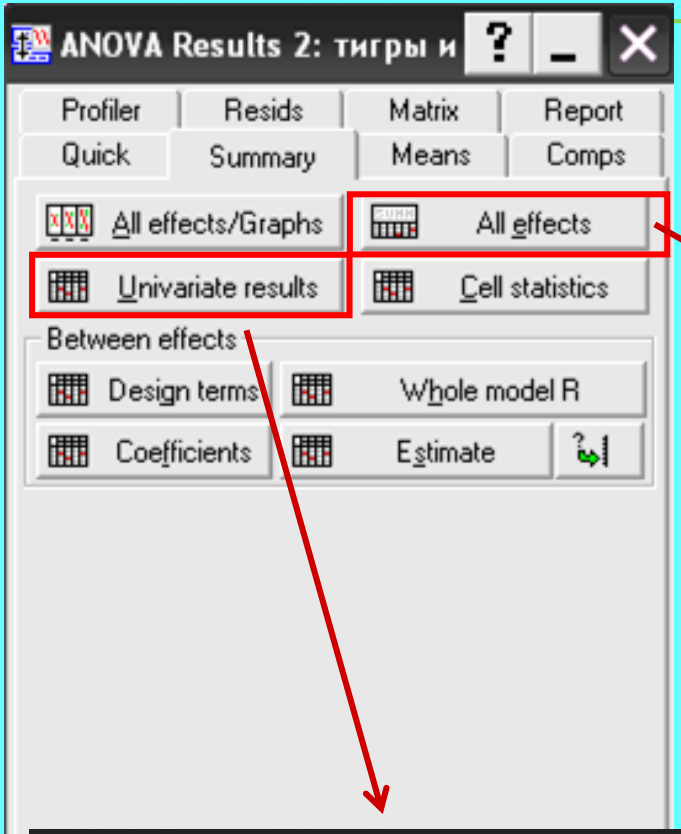
MANOVA

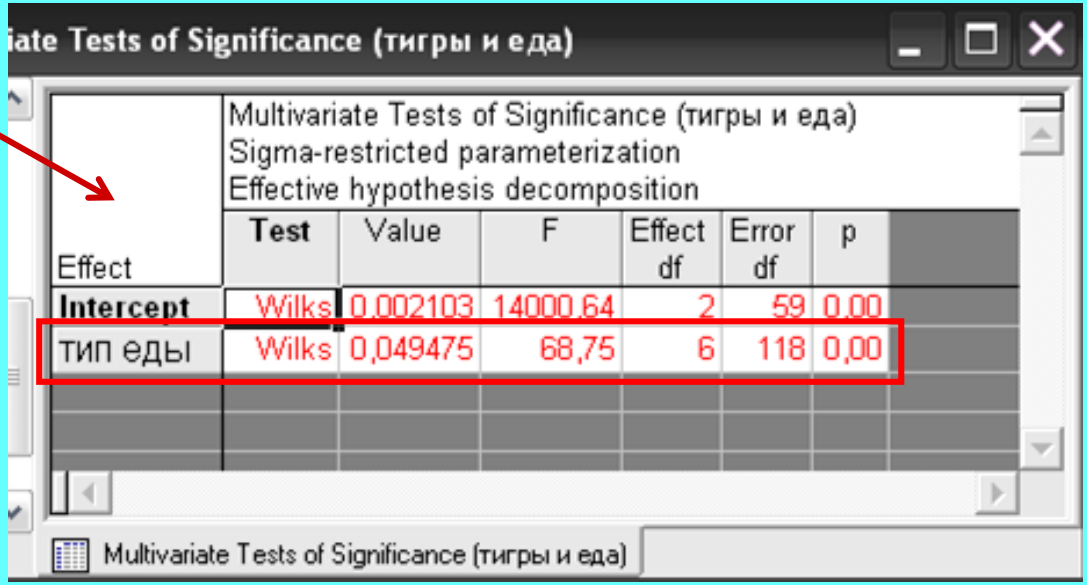

General ANOVA/MANOVA: тигры и еда
Quick
Type of analysis:
One-way ANOVA
Main effects ANOVA
Factorial ANOVA
Repeated measures ANOVA
Use One-way ANOVA to analyze designs with a single categorical independent variable (factor).
Specification method:
Quick specs dialog
Analysis Wizard
Analysis syntax editor
Multiple dependent variables can be specified for any type of analysis.
OK
Cancel
Options
Open Data
SELECT CASES
Weighted moments
DF =
W-1
N-1


ANOVA/MANOVA One-Way ANOVA: тигры и еда
Quick
Options
Variables
Dependent variables: масса тела длина тела
Categorical factor: тип еды
Factor codes: selected
Between effect: "тип еды"
OK
Cancel
Options
Syntax editor


Select dependent variables and categorical predictors (fact
1 - масса тела
2 - тип еды
3 - пол
4 - длина тела
5 - Var5
6 - Var6
7 - Var7
8 - Var8
9 - Var9
10 - Var10
Select All
Spread
Zoom
Dependent variable list:
1 4
Categorical predictors (factors):
2
Cancel
[Bundles]...
Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.
Show appropriate variables only

MANOVA



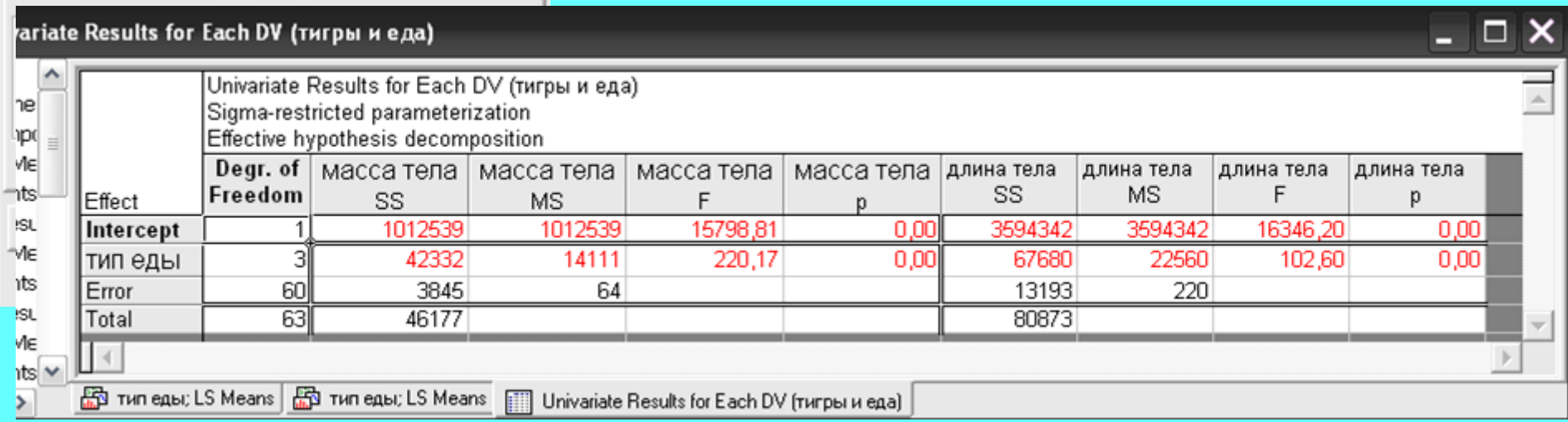


ate Tests of Significance (тигры и еда)

Multivariate Tests of Significance (тигры и еда)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Test	Value	F	Effect df	Error df	p
Intercept	Wilks	0,002103	14000,64	2	59	0,00
тип еды	Wilks	0,049475	68,75	6	118	0,00

Multivariate Tests of Significance (тигры и еда)



ariate Results for Each DV (тигры и еда)

Univariate Results for Each DV (тигры и еда)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Degr. of Freedom	масса тела SS	масса тела MS	масса тела F	масса тела p	длина тела SS	длина тела MS	длина тела F	длина тела p
Intercept	1	1012539	1012539	15798,81	0,00	3594342	3594342	16346,20	0,00
тип еды	3	42332	14111	220,17	0,00	67680	22560	102,60	0,00
Error	60	3845	64			13193	220		
Total	63	46177				80873			

тип еды; LS Means | тип еды; LS Means | Univariate Results for Each DV (тигры и еда)

MANOVA

Требования и рекомендации:

1. Для MANOVA особенно важно, чтобы измерения были случайными и не зависели друг от друга.
2. Многомерное нормальное распределение
3. Гомогенность дисперсий в группах
4. Корреляции между зависимыми переменными должны быть одинаковыми между группами
5. Чем больше число переменных, тем меньше мощность теста, т.е., для большого числа переменных необходимы большие выборки
6. После проведения MANOVA допустимо проводить просто ANOVA с последующими пост-хок тестами.

ФАКТОРНЫЙ АНАЛИЗ

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

Наши цели:

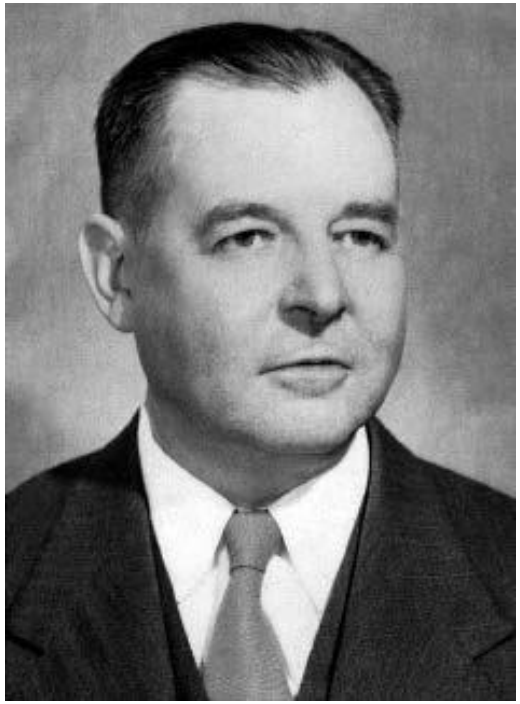
1. Уменьшить число исходных переменных с минимальными потерями исходной информации (что, например, уменьшит эффект множественных сравнений);
2. Обнаружить скрытые закономерности в данных, которые не выявляются при анализе отдельных переменных, путём помещения в пространство новых переменных (scaling).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

Анализ главных компонент
(principal component analysis, PCA)

Исторически первый и наиболее распространенный многомерный метод анализа данных.

Впервые предложен К. Пирсоном в 1901 г., а затем независимо переоткрыт и разработан Г. Хоттелингом в 1930-х гг.

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



Harold Hotelling (1895–1973)

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.

Этапы анализа

Этап 0. Подготовка данных к анализу.

- ✓Проверка распределений на соответствие нормальному;
- ✓Трансформация данных (напр., логарифмирование некоторых переменных);
- ✓Исключение аутлаеров
- ✓Стандартизация данных (если переменные – в разных шкалах);
- ✓проверка, нет ли слишком сильно коррелирующих переменных ($r>0.95$ исключить; иначе невозможны будут операции с матрицами).

Этап 1. Получение и отбор компонент

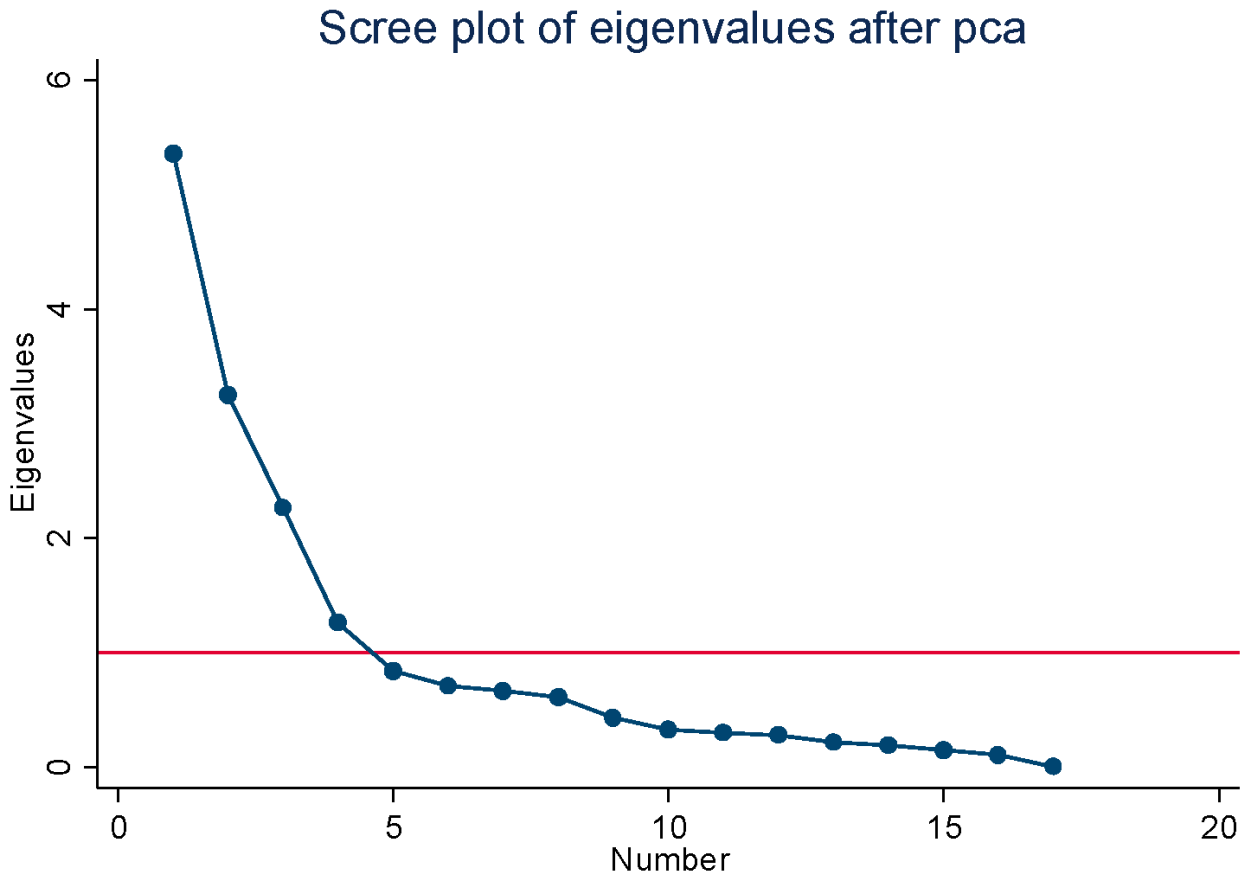
У каждой компоненты есть **eigenvalue** и **eigenvector.**

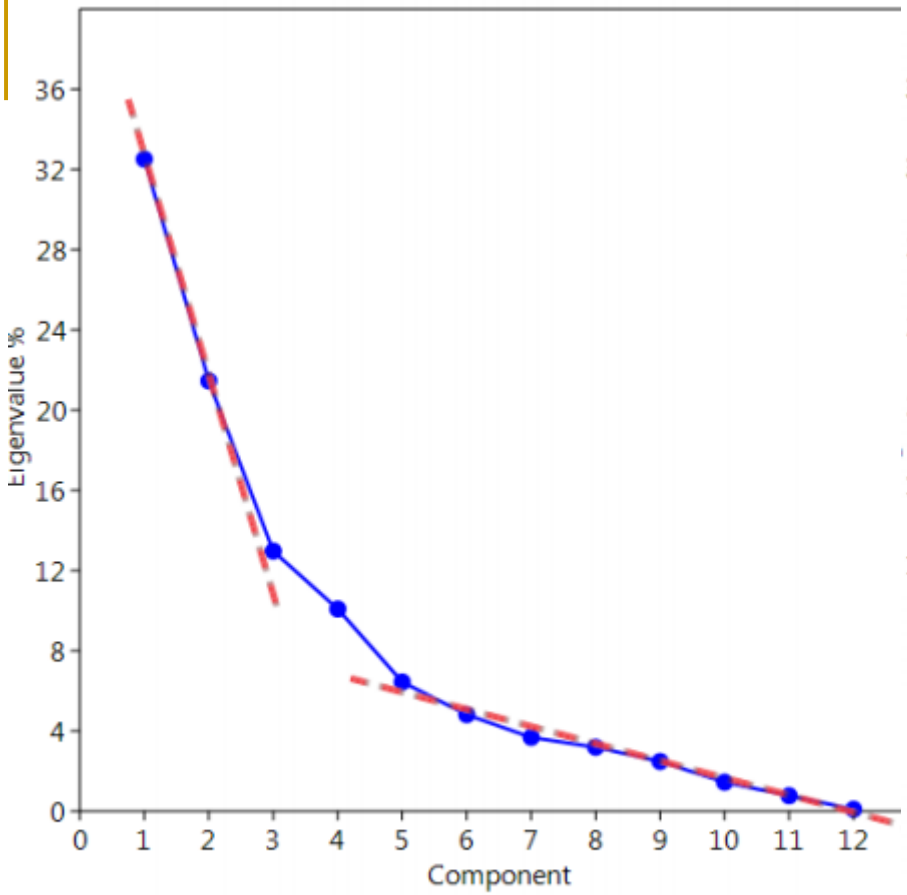
Eigenvalues (собственные значения): получаются из матрицы корреляций, общая сумма = числу переменных. Первая компонента объясняет как можно больше общей дисперсии и имеет максимальное eigenvalue, вторая – максимум оставшейся и т.п. Дисперсия каждой переменной в среднем принимается за 1.

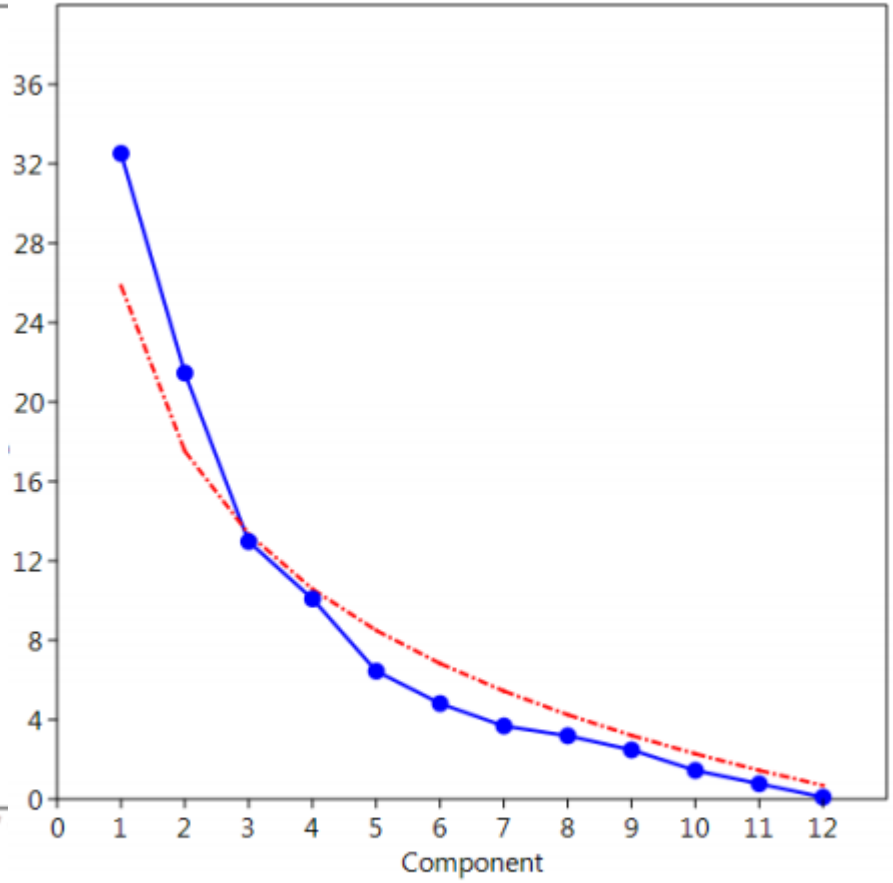
Правило **«eigenvalue = 1»**: оставляем только компоненты, у которых собственные значения >1, т.е., они объясняют больше общей дисперсии, чем одна переменная в среднем.

Eigenvector (Factor Score Coefficients) показывает количественный вклад каждой переменной в компоненты.

Диаграмма каменистой осыпи (scree plot) позволяет принять решение, сколько компонент оставлять. Критерии - «eigenvalue = 1» и наличие перелома на диаграмме.







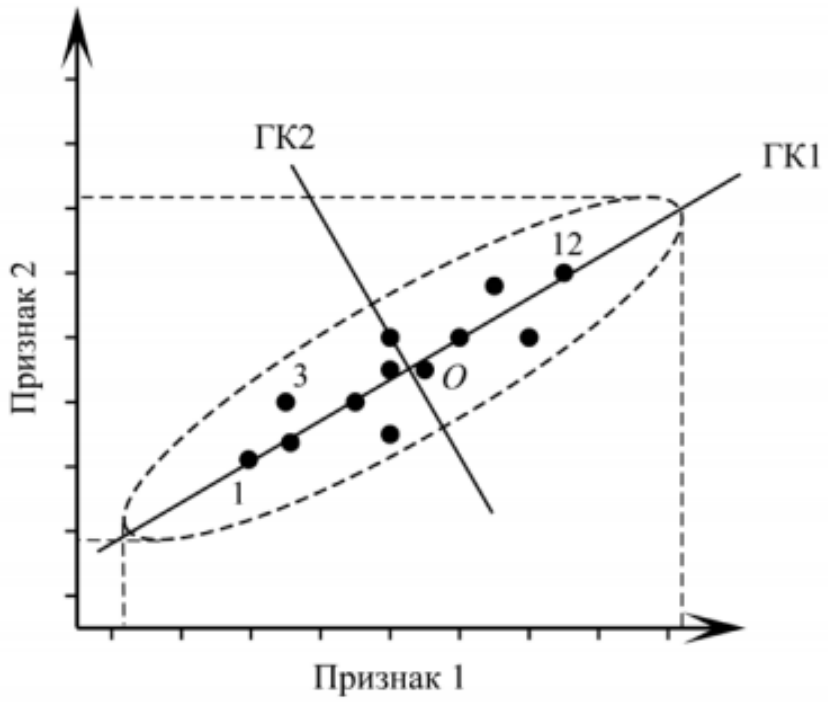
а) *б)*

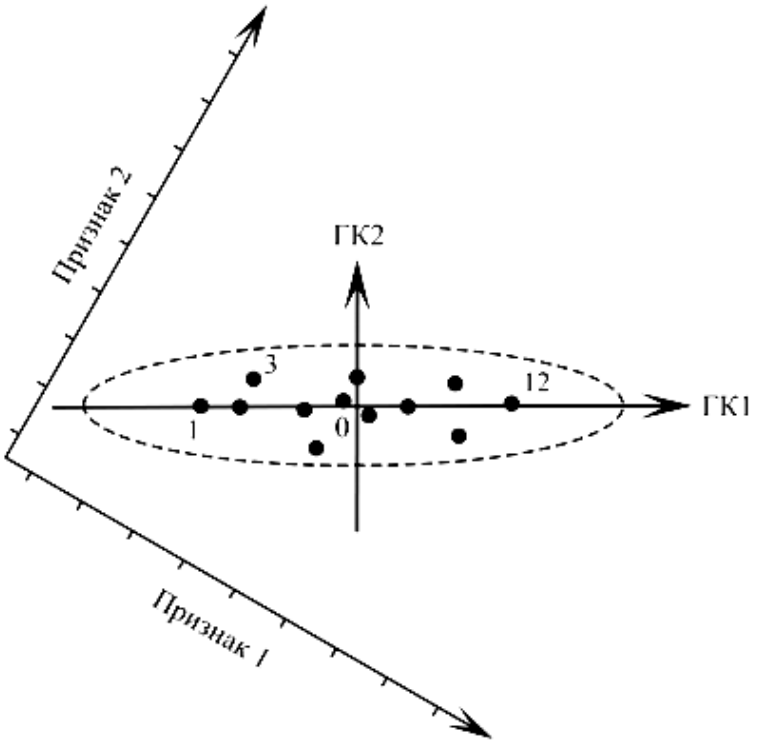
Определение числа наиболее важных компонент на графике «каменистой осыпи»:
а) критерий Кэттелла, б) критерий «сломанной трости»

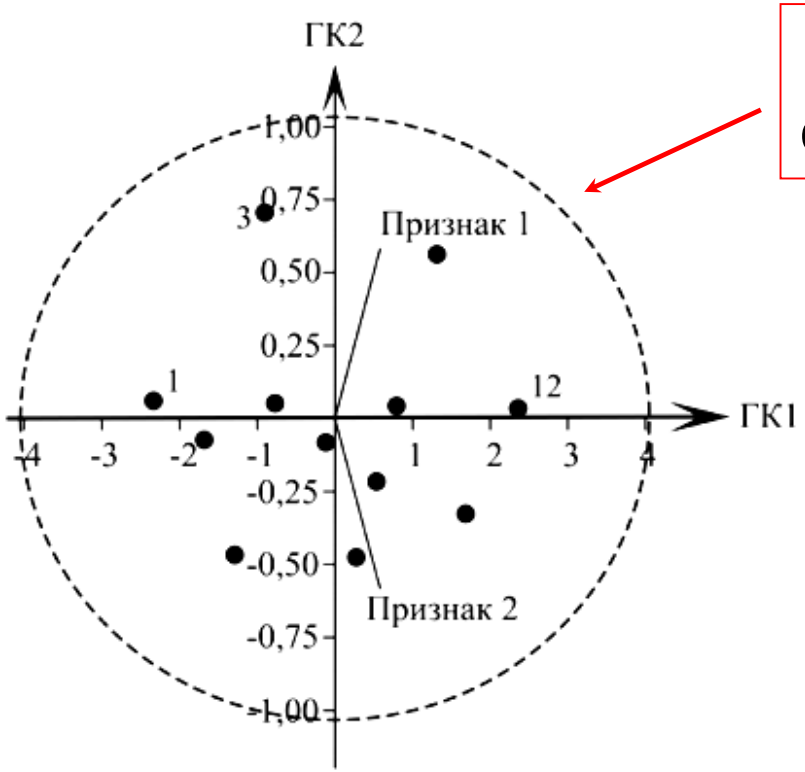
Несколько слов о компонентах (факторах):

✓В многомерном пространстве первая компонента располагается вдоль наибольшей дисперсии, т.е., это почти аналог линии линейной регрессии.

✓Компоненты взаимно перпендикулярны

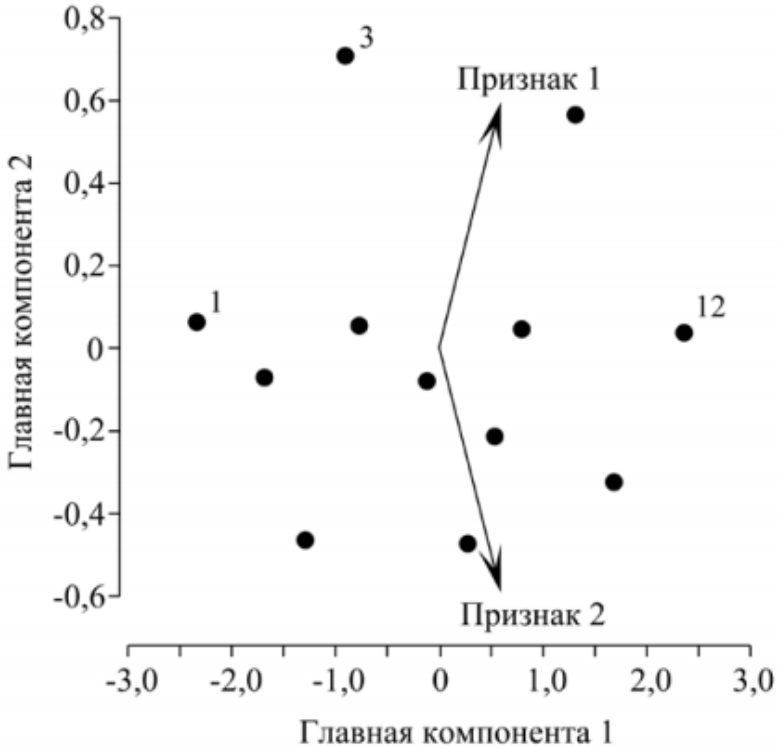






Нормировка на долю объясняемой дисперсии

Биплот (biplot)



Этап 2. Интерпретация компонент: **eigenvectors** и **factor loadings**

1. **Eigenvectors** (коэффициенты): чем коэффициент больше по модулю, тем больше вклад данной переменной в компоненту.
2. Factor **loadings** (корреляции Пирсона для компоненты с каждой из исходных переменных): чем больше по модулю, тем сильнее корреляция компоненты с переменной.

Компоненты **легко интерпретировать** если: каждая исходная переменная коррелирует только с одной компонентой; loadings близки либо 1/-1, либо 0 (так получается, если исходно корреляция переменных есть).
Сложно интерпретировать если: среди factor loadings много невысоких значений; некоторые переменные почти одинаково коррелируют с несколькими компонентами.

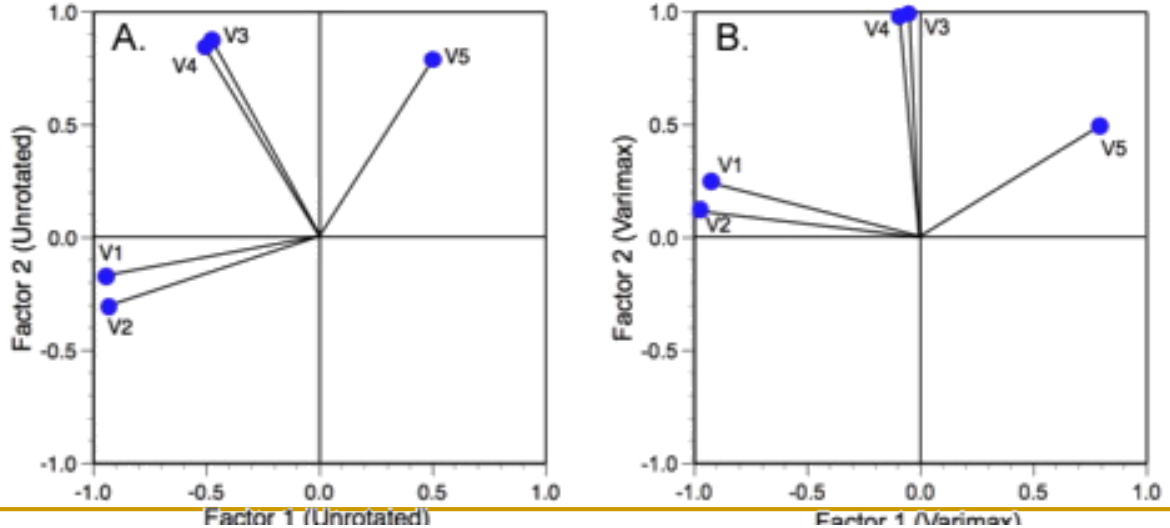
Этап 3. Вращение компонент (rotation)

Итак, мы отбросили часть компонент. Теперь можно еще немного повернуть оставшиеся компоненты.

Цель вращения – облегчить интерпретацию компонент: чтобы коэффициенты и loadings были близки либо 0, либо 1.

Поворачиваем их так, чтобы уменьшилось число средних корреляций (loadings), и каждая переменная коррелировала бы с одной компонентой. Varimax rotation – самый распространённый и удобный метод.

Вращение сохраняет компоненты ортогональными.



Вращение компонент (факторов)

Обычно используют **ортогональное** вращение – факторы остаются перпендикулярными друг другу. Самый стандартный метод - varimax.
Не ортогональное вращение – oblique rotation, допускает корреляцию компонент. Оси координат становятся не перпендикулярными, и возникают трудности с визуализацией и интерпретацией результатов.

Анализ остатков – residuals, residual correlation – позволяет оценить, насколько много информации было потеряно 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.

Этап 4. Интерпретация новых компонент

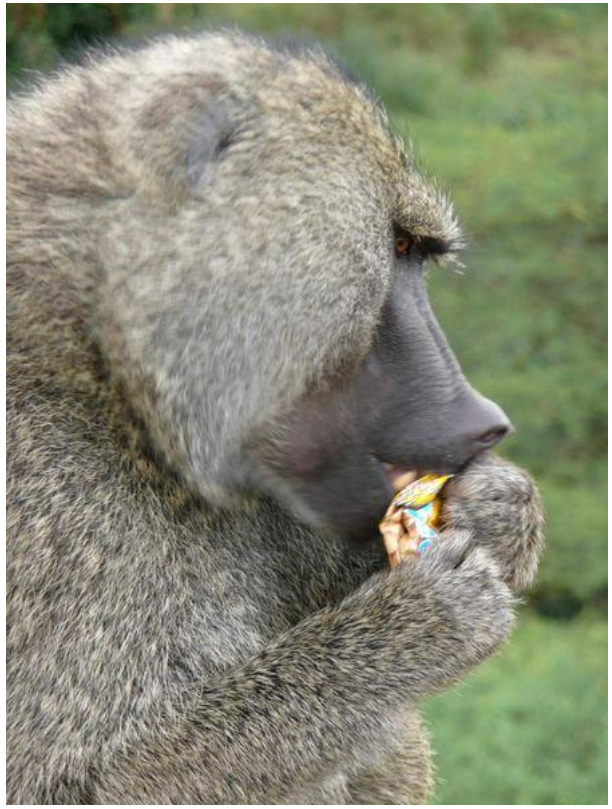
Получение **factor loadings** после вращения (проверка, стало ли лучше). Рассмотрение корреляций новых, «повёрнутых» компонент с исходными переменными, понимание их биологического смысла.

Этап 5. Получение значений новых переменных для каждого объекта (для дальнейшего анализа.)

Итог о компонентах (факторах):

- ✓В многомерном пространстве первая компонента располагается вдоль наибольшей дисперсии облака объектов.
- ✓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.
- ✓Оставляем столько компонент, сколько обеспечит интерпретируемость результатов. Нет смысла оставлять компоненты, с которыми не коррелирует ни одна исходная переменная. Правило «eigenvalue =1».

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.

Однако реальных факторов, определяющих эти предпочтения, наверняка меньше.

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

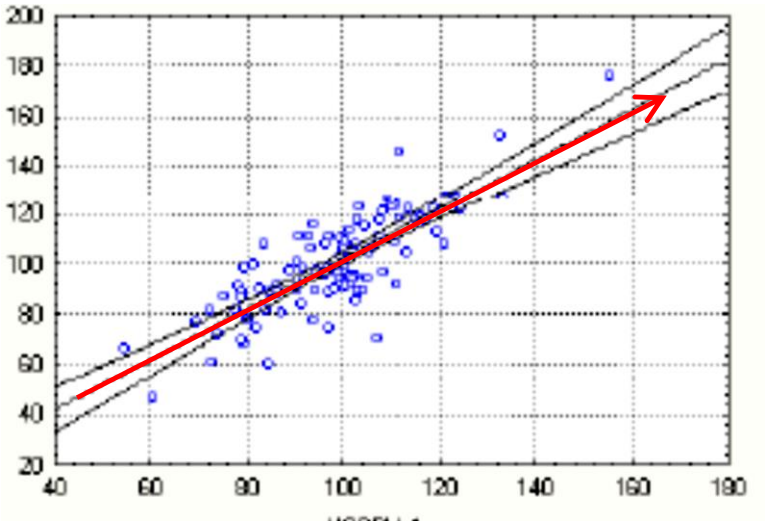
Итак,

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.

Подразумевается, что наши реально измеренные переменные являются <u>линейными комбинациями</u> этих подлежащих факторов.



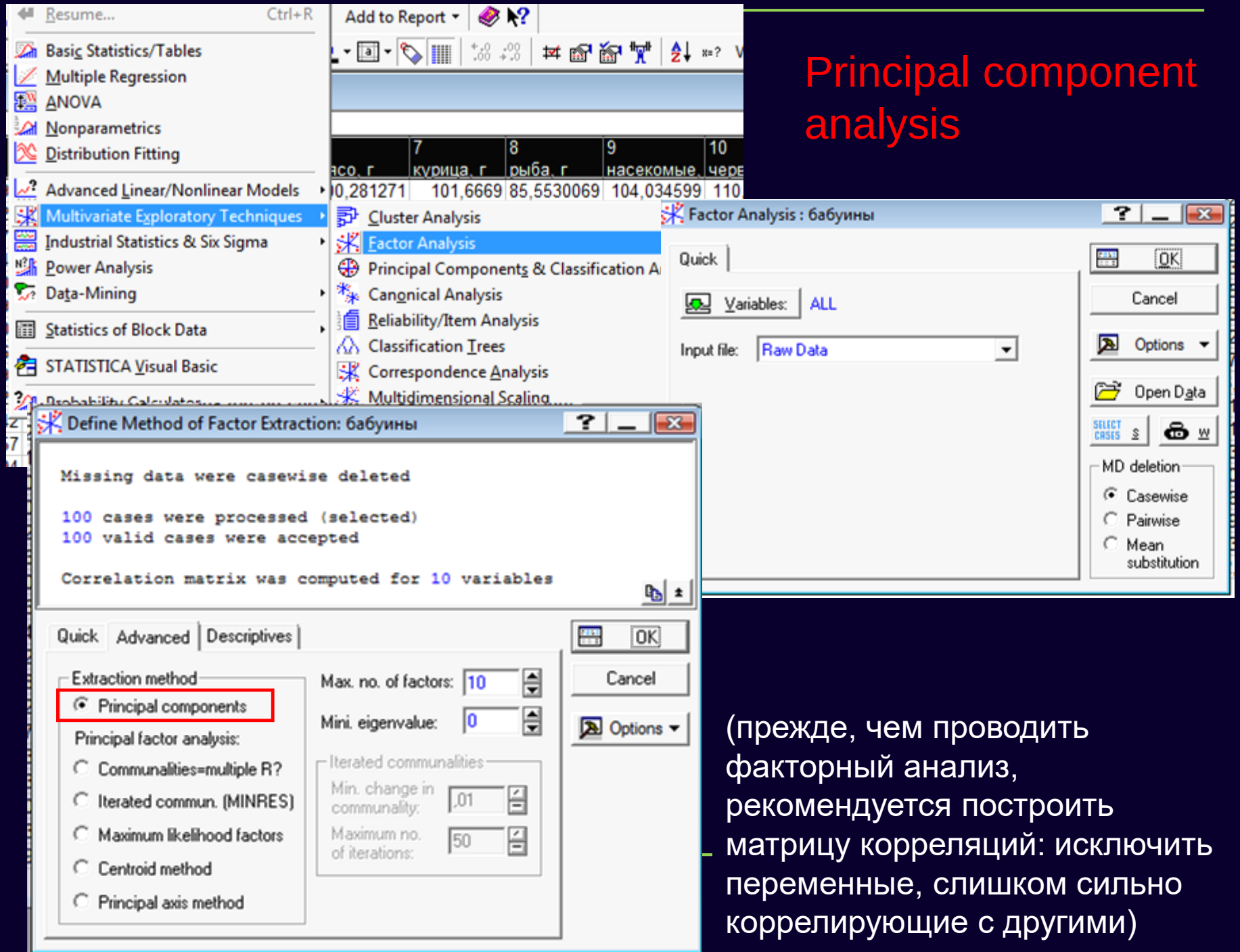
Примерно так будет проходить новая ось OX – первая компонента.

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

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

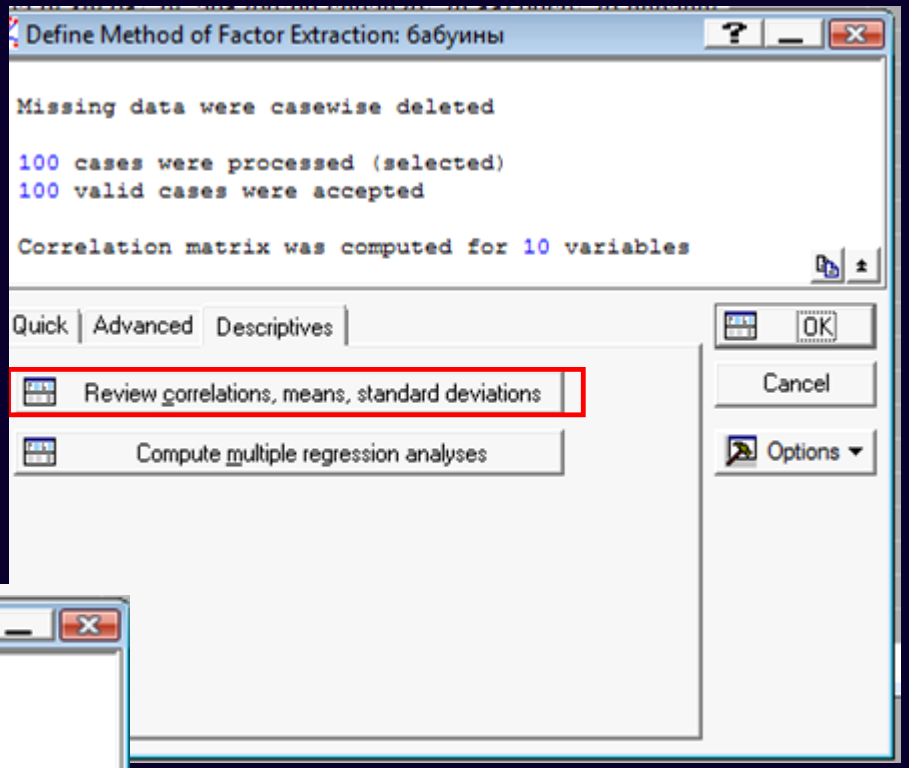
Мы измеряем привлекательность пищи каждого типа, для каждого зверя. Сколько факторов скрывается за разными предпочтениями павианов в еде?

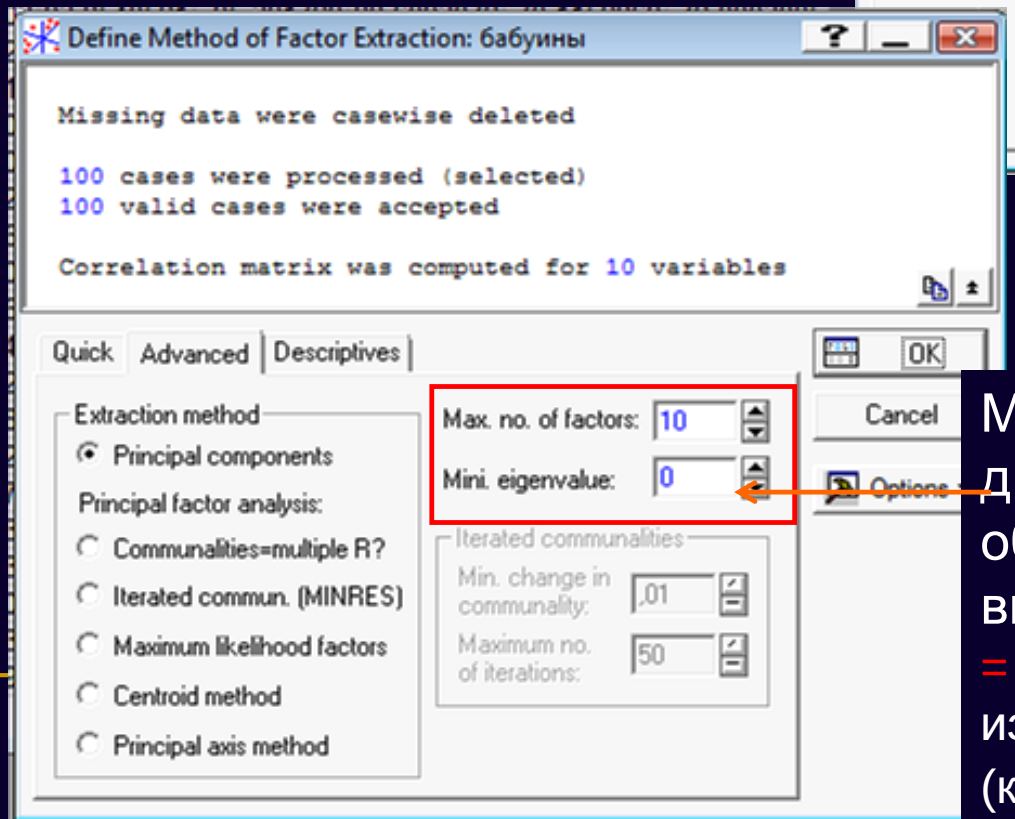
Principal component analysis



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

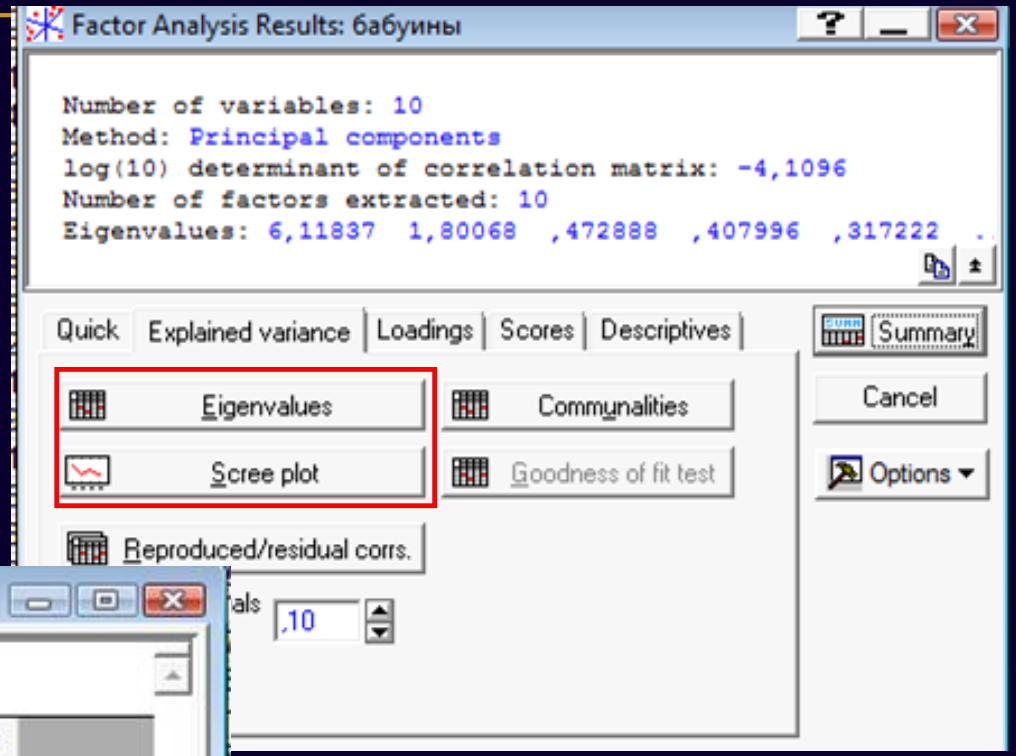
Просмотрим матрицу корреляций:
Не должно быть слишком сильно коррелирующих друг с другом переменных (иначе матрица не может быть транспонирована: *matrix ill-conditioning*)

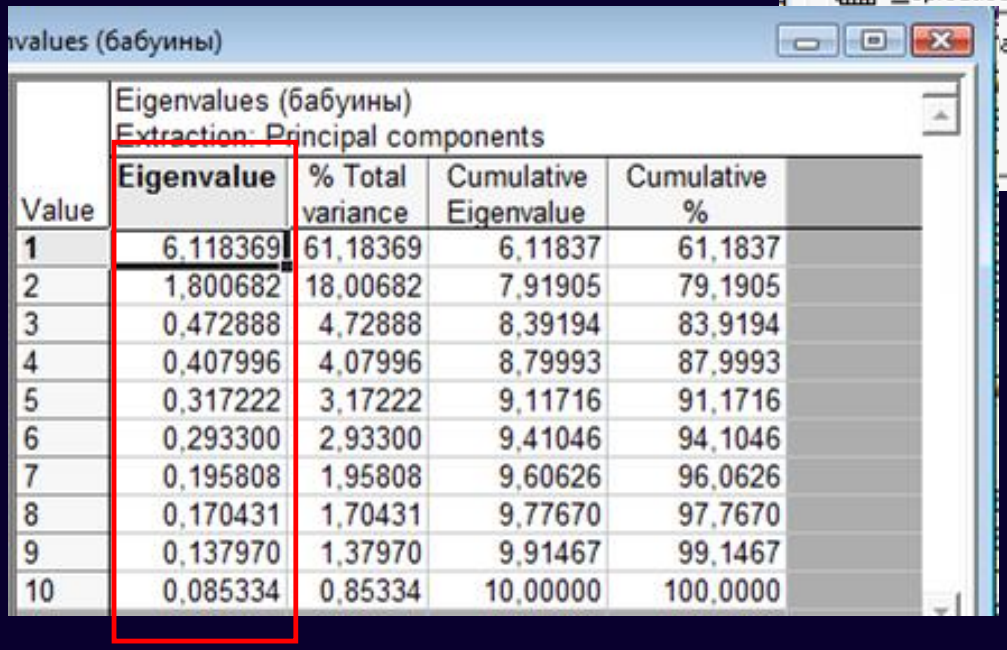




Можно задать min количество дисперсии, которое должен объяснять фактор, чтобы его включили в анализ (обычно min = 1, что соответствует случайной изменчивости одной переменной (критерий Кайзера))

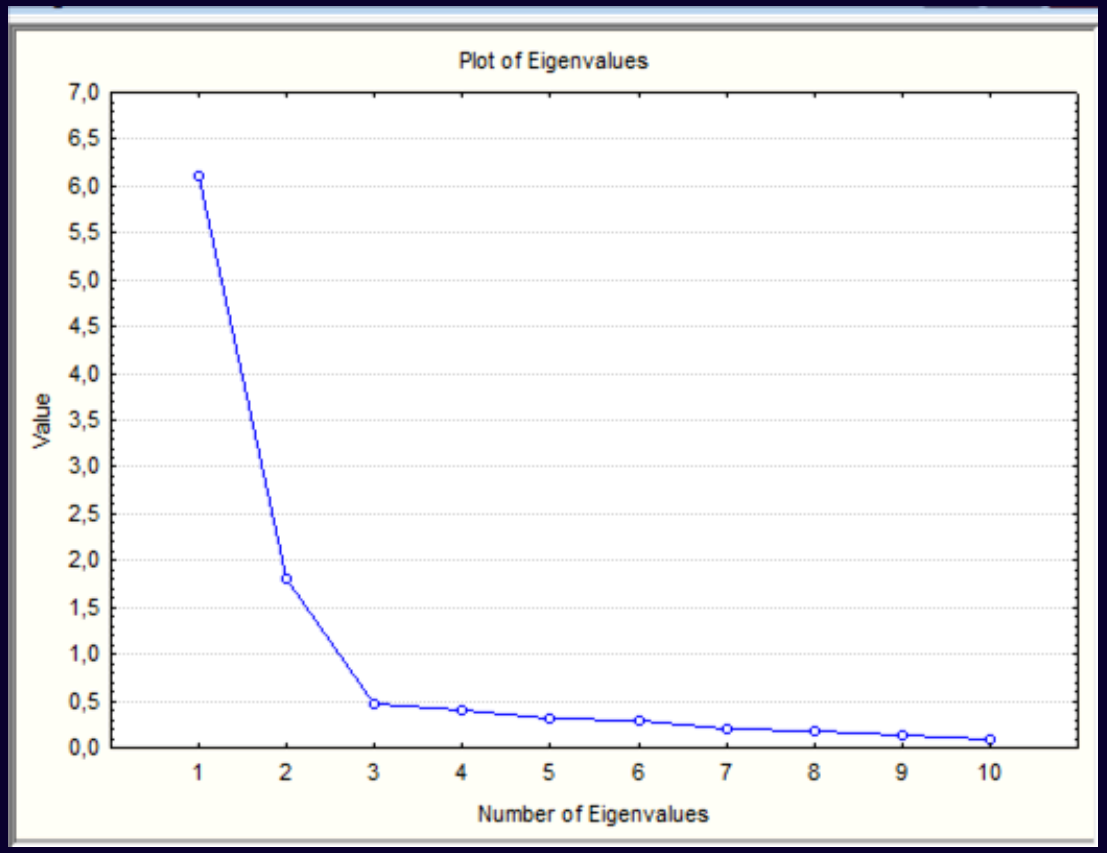
Собственные значения (eigenvalues)– определяют, какую долю общей дисперсии объясняет данный фактор.





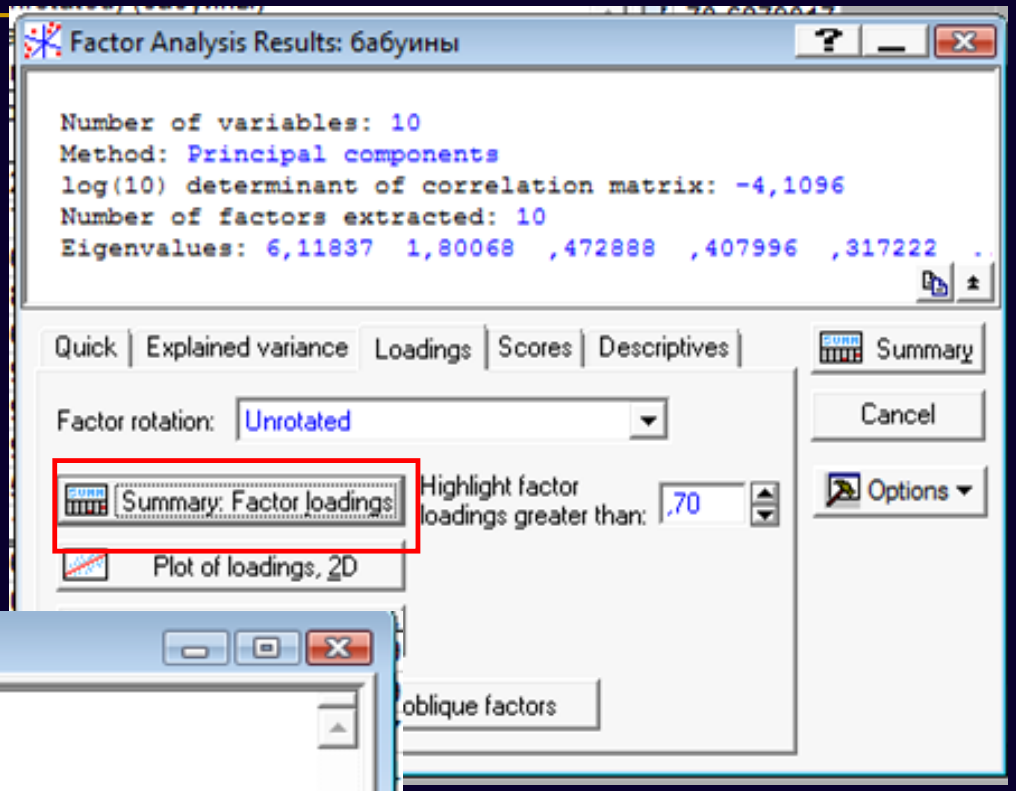
Eigenvalues (бабуины)
Extraction: Principal components

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

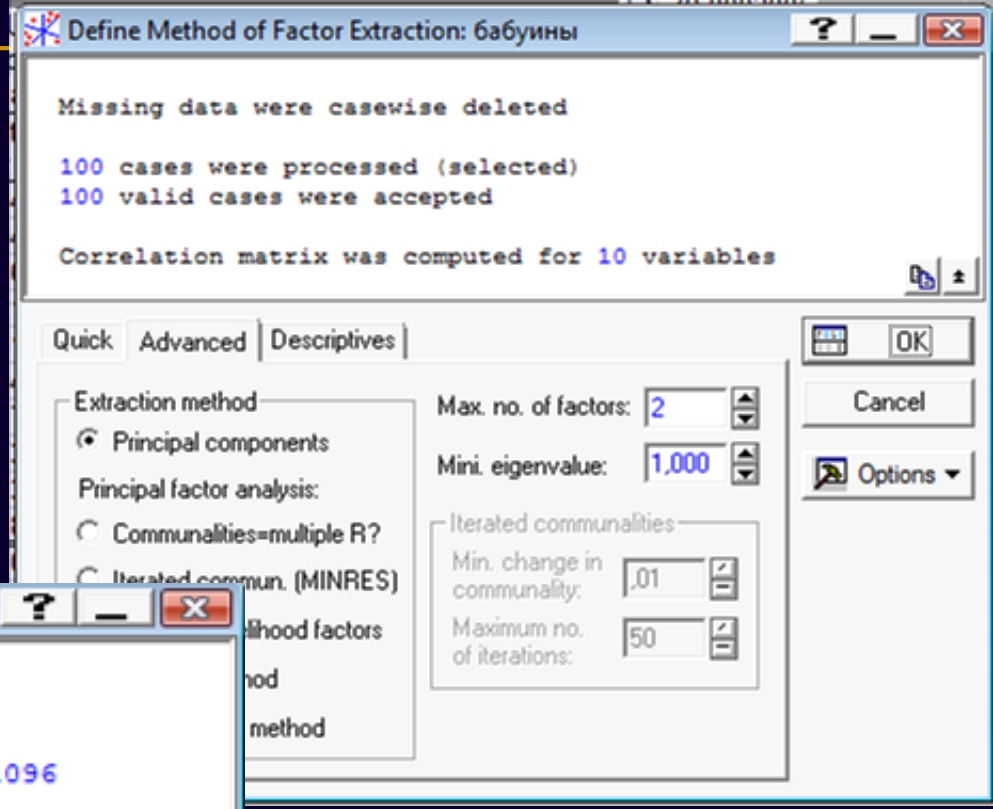


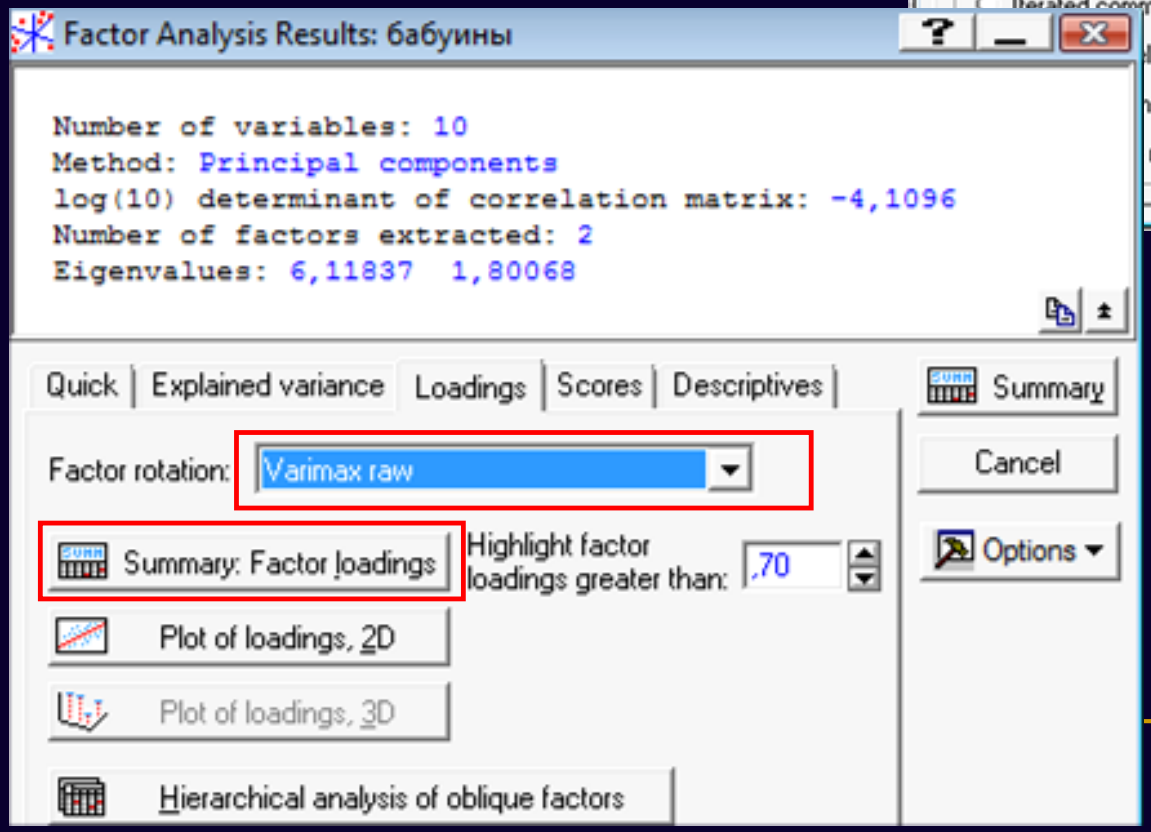
tor Loadings (Unrotated) (бабуины)

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:

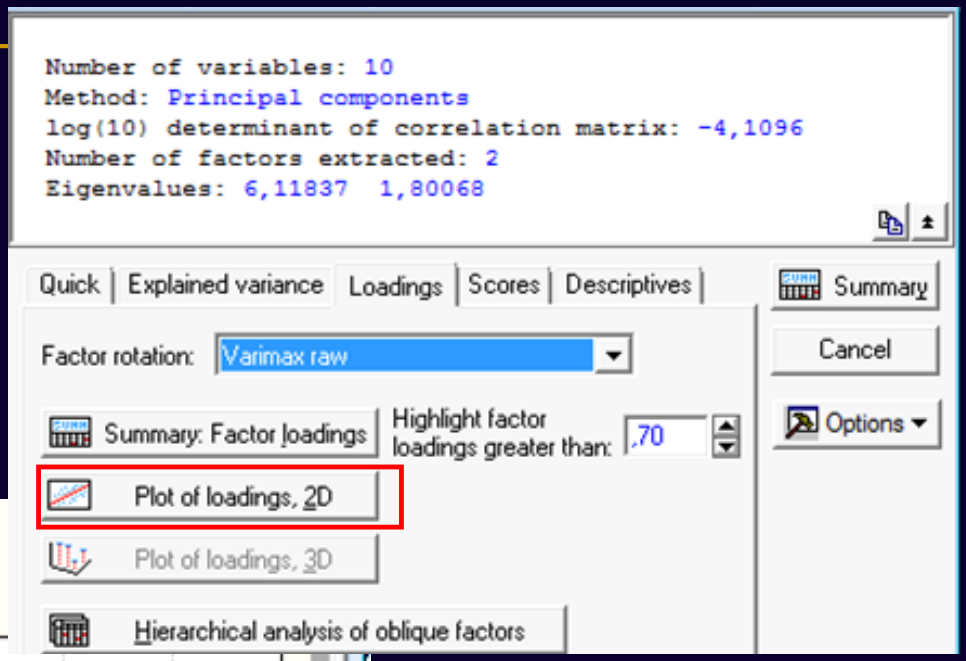
r Loadings (Varimax raw) (бабуины)

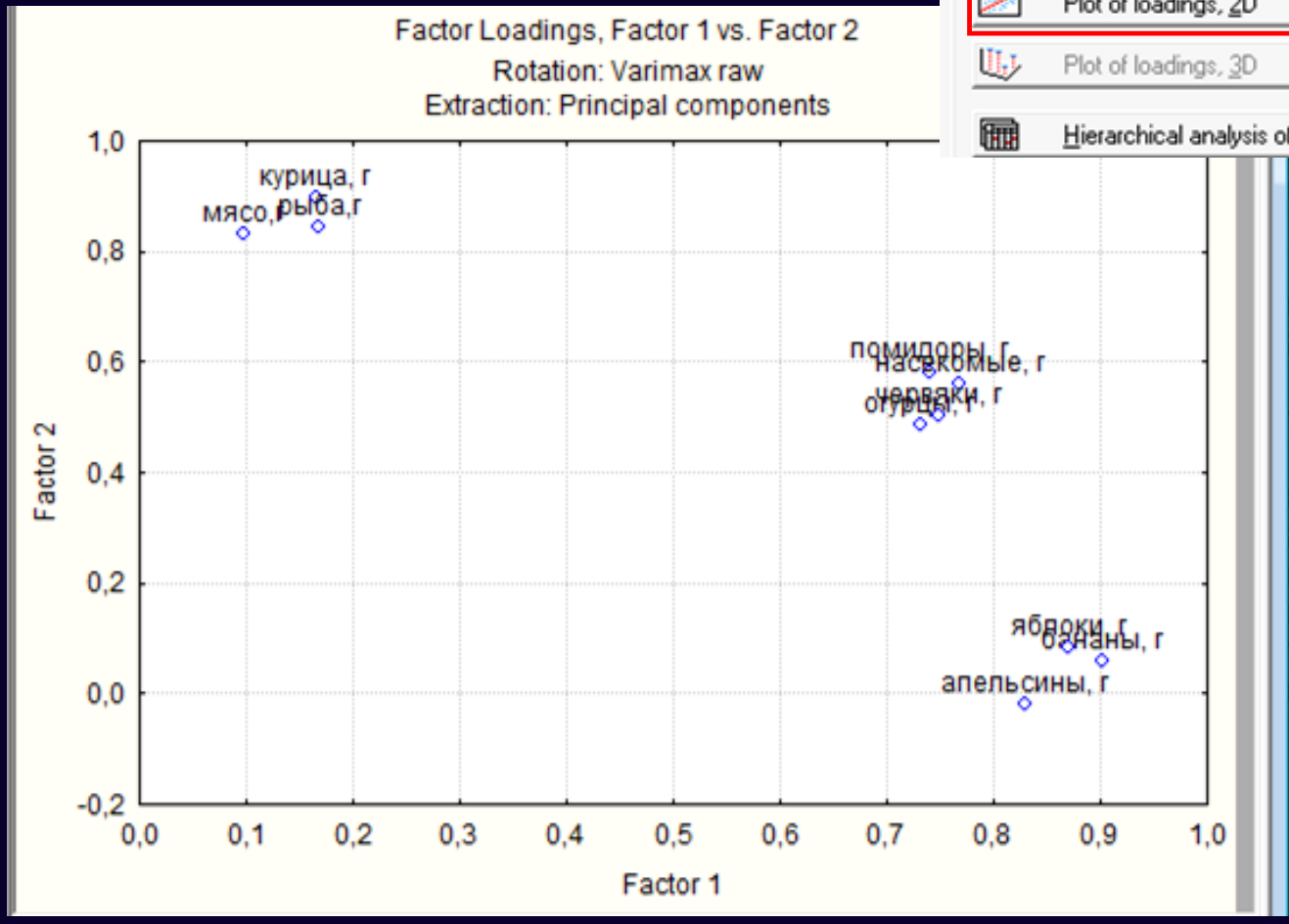
Variable	Factor Loadings (Varimax raw) (ба Extraction: Principal components (Marked loadings are > ,700000)	
	Factor 1	Factor 2
апельсины, г	0,830623	-0,019320
бананы, г	0,902408	0,058905
яблоки, г	0,870524	0,082595
помидоры, г	0,739857	0,582885
огурцы, г	0,731191	0,484489
мясо,г	0,097371	0,829676
курица, 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.

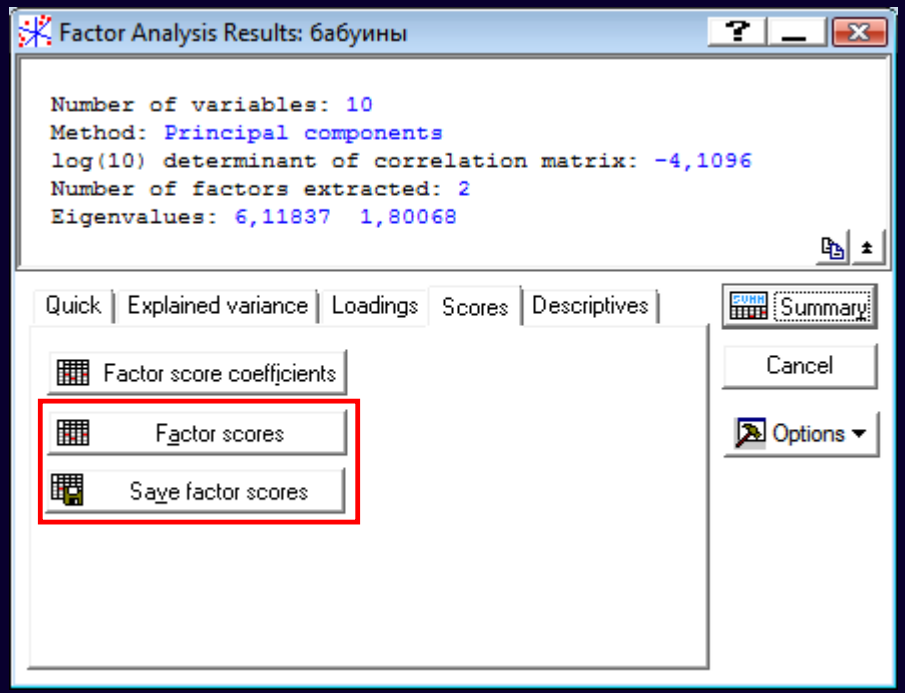
Итак, пищевые предпочтения павианов составлены из двух основных факторов – отношением к животной и растительной пище.

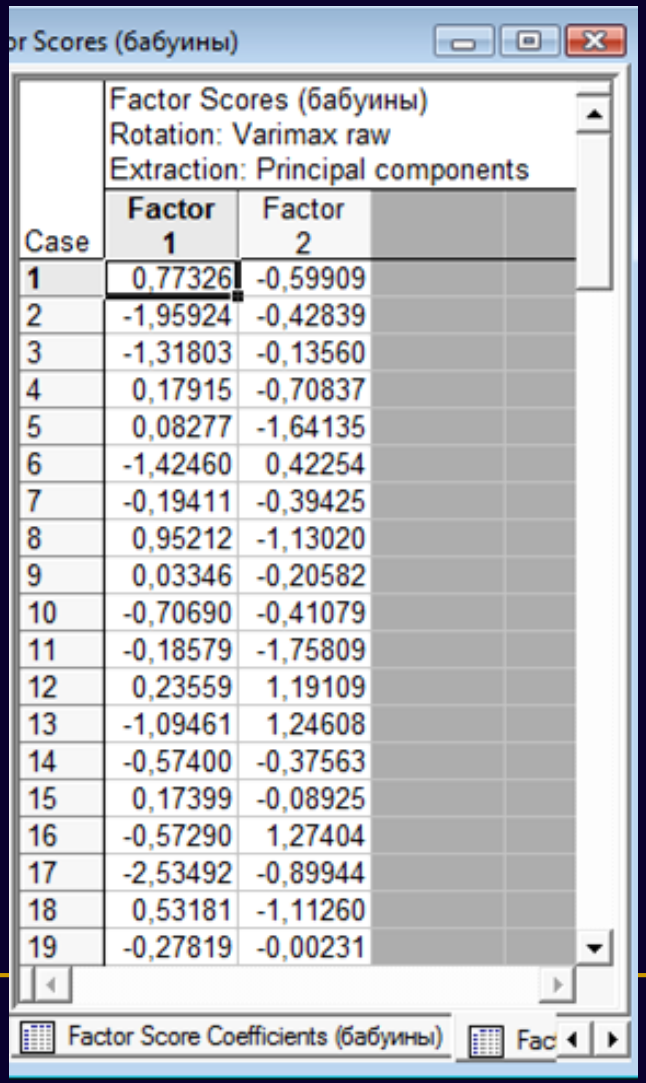
Посмотрим, как исходные переменные расположились в пространстве новых факторов





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





Factor Scores (бабуины)
Rotation: Varimax raw
Extraction: Principal components

Case	Factor 1	Factor 2
1	0,77326	-0,59909
2	-1,95924	-0,42839
3	-1,31803	-0,13560
4	0,17915	-0,70837
5	0,08277	-1,64135
6	-1,42460	0,42254
7	-0,19411	-0,39425
8	0,95212	-1,13020
9	0,03346	-0,20582
10	-0,70690	-0,41079
11	-0,18579	-1,75809
12	0,23559	1,19109
13	-1,09461	1,24608
14	-0,57400	-0,37563
15	0,17399	-0,08925
16	-0,57290	1,27404
17	-2,53492	-0,89944
18	0,53181	-1,11260
19	-0,27819	-0,00231

Factor Score Coefficients (бабуины)

Требования к выборкам для проведения PCA

1. Связь переменных должна быть ***линейной*** (это важно! Диагностика - осмотр скаттерплотов);
2. многомерное **нормальное** распределение – не критично (оценка – на основе гистограмм; если есть группы – внутри групп);
3. **Трансформация** данных – для линейных связей;
4. Исключение **аутлаеров!** (особенно если они меняют интерпретацию компонент);
5. Рекомендуется, чтобы **размер выборки** был >50, строго ≥25 объектов; оптимальный – ≥100. Чем больше переменных, тем больше должен быть размер выборки.
6. Между переменными должна быть ненулевая корреляция, но коэффициентов корреляции, близких единице, тоже быть не должно.

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.

Факторный анализ – более общее понятие, чем метод ГК. И МГК, и факторный анализ основаны на идее неких факторов, определяющих изменчивость измеренных переменных.

МГК	Факторный анализ
В основе большого числа переменных – **мало факторов**	В основе большого числа переменных – **мало факторов**
Эти факторы для удобства принимаются **ортогональными**	Ортогональность факторов **не подразумевается**
Их получают как **линейную** комбинацию переменных	Их получают как **линейную** комбинацию переменных
В основе – манипуляции с дисперсией**, вся общая дисперсия** распределяется между компонентами.	В основе – действия с корреляциями, используется **только дисперсия** переменной, общая и для других переменных.
Удобнее для уменьшения размерности.	**Лучше отражает реальную структуру данных.**

Другие многомерные методы, близкие анализу главных компонент

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

Другие многомерные методы, близкие анализу главных компонент

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

Спасибо за внимание!