Занятие 10

Метод главных компонент и другие методы исследования многомерных связей

Многомерные методы анализа: краткое повторение

1. Данные: много количественных переменных (зависимых)
2. На основании матрицы корреляций или ковариаций получаем новые переменные (не коррелирующие друг с другом) = компоненты – линейные комбинации исходных (R-mode analysis)
3. Первая переменная описывает максимум изменчивости между объектами = имеет наибольшее собственное значение (eigenvalue)
4. Для каждой компоненты есть список коэффициентов (eigenvector) и список корреляций (loadings); величина тех и других показывает, какие переменные вносят вклад в компоненты
5. можно рассчитать значения компонент для объектов
6. На основании матрицы дистанций между объектами тоже можно получить новые независимые переменные (Q-mode analysis)

Мы уже рассмотрели задачу сравнения ГРУПП по нескольким зависимым переменным (MANOVA и дискриминантный анализ).

НОВЫЕ ЗАДАЧИ:

1. Уменьшить число исходных переменных с минимальными потерями информации (чтобы использовать их в дальнейшем анализе);
2. Обнаружить скрытые закономерности в данных, которые не выявляются при анализе отдельных переменных (путём помещения объектов в пространство новых переменных, scaling).

PCA

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.

PCA

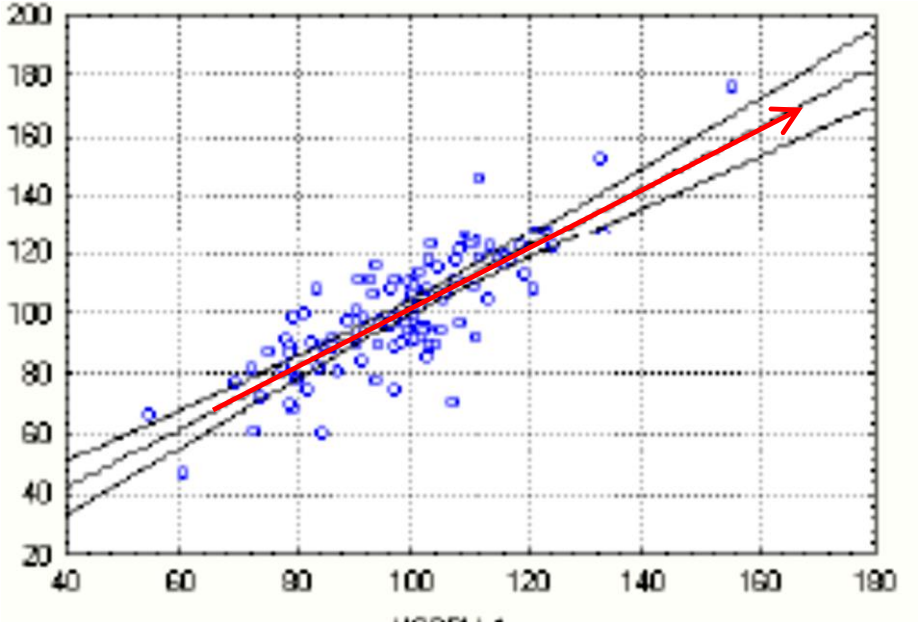
Мы хотим

Найти те реальные факторы, которые определяют изменчивость (объясняют действие) большого количества измеренных нами реальных переменных.
Подразумевается, что таких факторов гораздо меньше, чем исходных переменных.

Скажем, я люблю торт «Прага», эскимо, шоколадные трюфели, пирожное «Картошка».
Не люблю голубцы, солянку, щи, борщ.
Всего 8 переменных.
Но на самом деле мне нравится вкус шоколада и не нравится вкус капусты, и эти 2 фактора и определяют мои вкусовые предпочтения.

PCA

Подразумевается, что наши реально измеренные переменные на самом деле являются линейными комбинациями этих подлежащих факторов.
Наша задача – выявить эти факторы и именно их использовать в дальнейшем анализе.



2 коррелирующие переменные превращаются в одну компоненту

Анализ главных компонент (principal component analysis, PCA)

n объектов
p переменных
PCA трансформирует переменные в новые **не коррелирующие** друг с другом **главные компоненты** (их ***k,*** от 1 до *p*) = факторы (principal components = factors).

Это линейные комбинации исходных переменных, аналогичные дискриминантным функциям, только деления на группы теперь нет.

Для каждого объекта рассчитываются значения новых компонент, которые можно использовать в другом анализе.

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

Этапы анализа

Этап 0. Подготовка данных к анализу.

✓Проверка распределений переменных на соответствие нормальному (строгое соответствие не обязательно);

✓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; иначе невозможны будут операции с матрицами).

PCA

Этапы анализа

Этап 1. Получение компонент

Для каждой компоненты (= фактора) получаем **eigenvalue** и **eigenvector.**

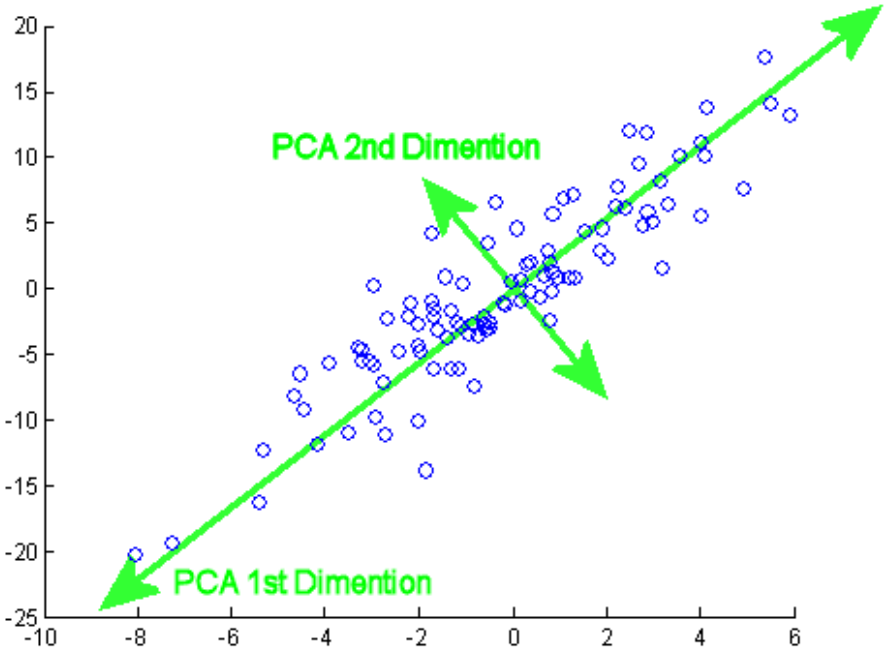
Eigenvalues: их получают из матрицы корреляций, их сумма = числу переменных, т.е., дисперсия каждой переменной в среднем принимается за 1. Первая компонента объясняет как можно больше общей дисперсии и имеет максимальное eigenvalue.
Разумно оставлять только те компоненты, для которых eigenvalues > 1, т.е., объясняющие больше общей дисперсии, чем средняя переменная; число компонент будет меньше числа исходных переменных.
Напоминание: они независимы между сбой, т.е., ортогональны.

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

Компоненты можно получать и из матрицы ковариаций; это осмысленно, если нам принципиально нужно оставить различия в абсолютных значениях переменных.

Этапы анализа

Этап 1. Получение компонент



Компоненты специально получаются ортогональными, т.е., они не коррелируют друг с другом.
Компонент получается **меньше, чем переменных**, в идеале – 1-3.

Чтобы компонент получилось мало, и их было легко интерпретировать, между исходными переменными должны быть **корреляции.**
Если корреляции слабые, уменьшить число переменных не получится!

Этапы анализа

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

1. Смотрим на **eigenvectors**: чем дальше коэффициент от нуля, тем больше вклад переменной в данную компоненту.
2. Смотрим на factor **loadings** (корреляции Пирсона для компоненты с каждой из исходных переменных): чем дальше от нуля, 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.

Этапы анализа

Этап 3. Вращение компонент (rotation)

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

Поворачиваем компоненты так, чтобы уменьшилось число средних корреляций (loadings), и каждая переменная стала коррелировать не более чем с одной компонентой. Varimax rotation – самый распространённый и удобный метод.

Вращение сохраняет компоненты ортогональными.

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.)

PCA

Ещё раз 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.
- ✓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».

PCA

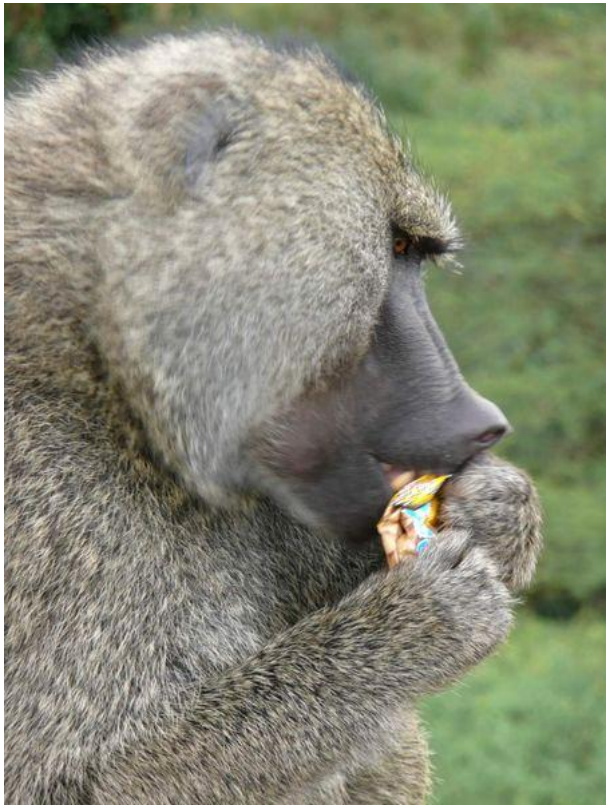
Вращение компонент (факторов)

Полученные компоненты 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 результаты этого метода трудно интерпретировать.

Анализ остатков – residuals, residual correlation – 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.

PCA

Изучаем пищевые предпочтения павианов.
Для каждой особи мы оценили количество еды 10 разных типов, съеденной за неделю. Павианы едят разную еду, поэтому типов пищи – 10.
особей в анализе – 100.
Но реальных факторов, определяющих эти предпочтения, наверняка меньше.



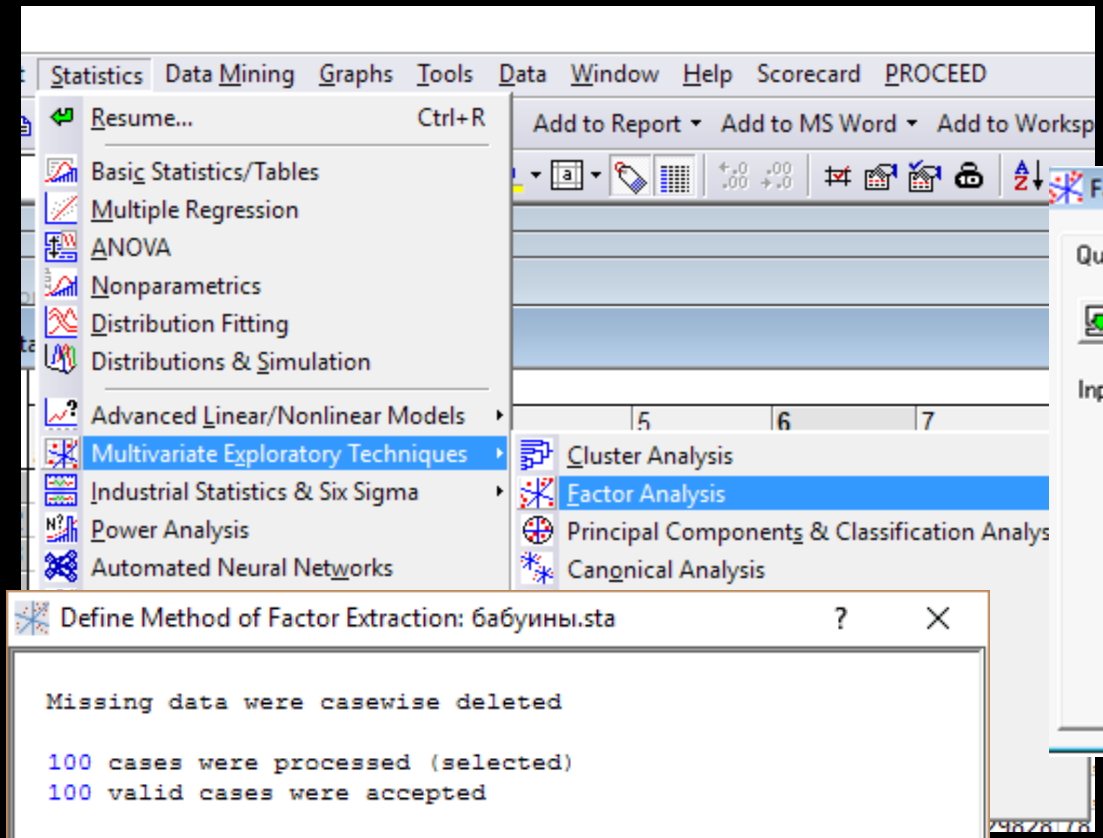
Мы хотим: 1) уменьшить число переменных для дальнейшего анализа; 2) определить, сколько (и каких) факторов определяют пищевые предпочтения павианов.

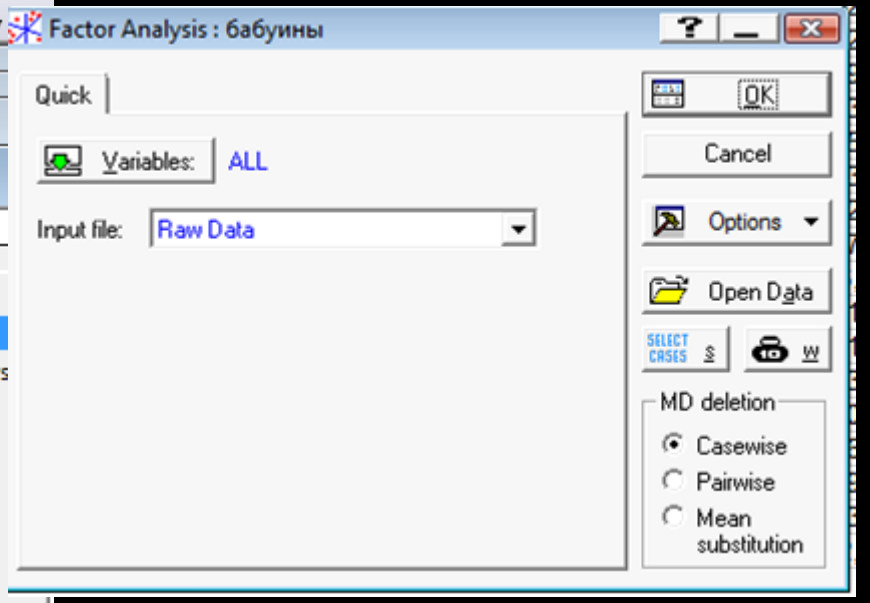
PCA

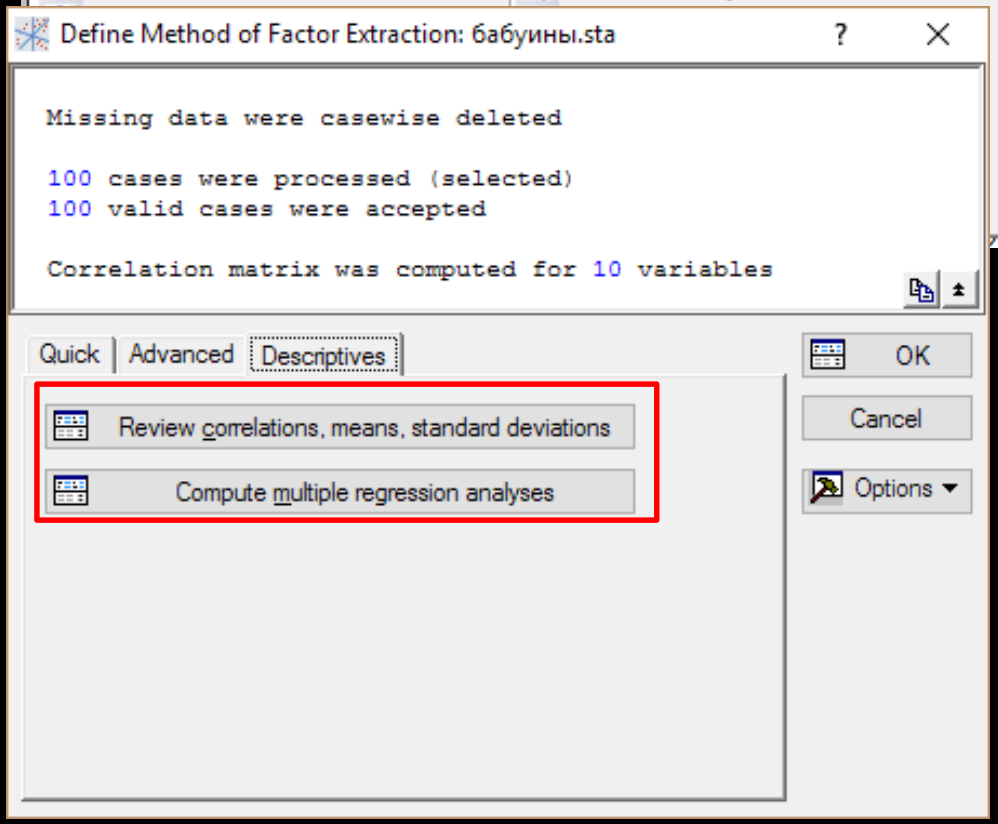
апельсины,
бананы,
яблоки,
помидоры,
огурцы,
мясо,
курица,
рыба,
насекомые,
черви.

10 переменных измерено для каждого животного

Principal component analysis

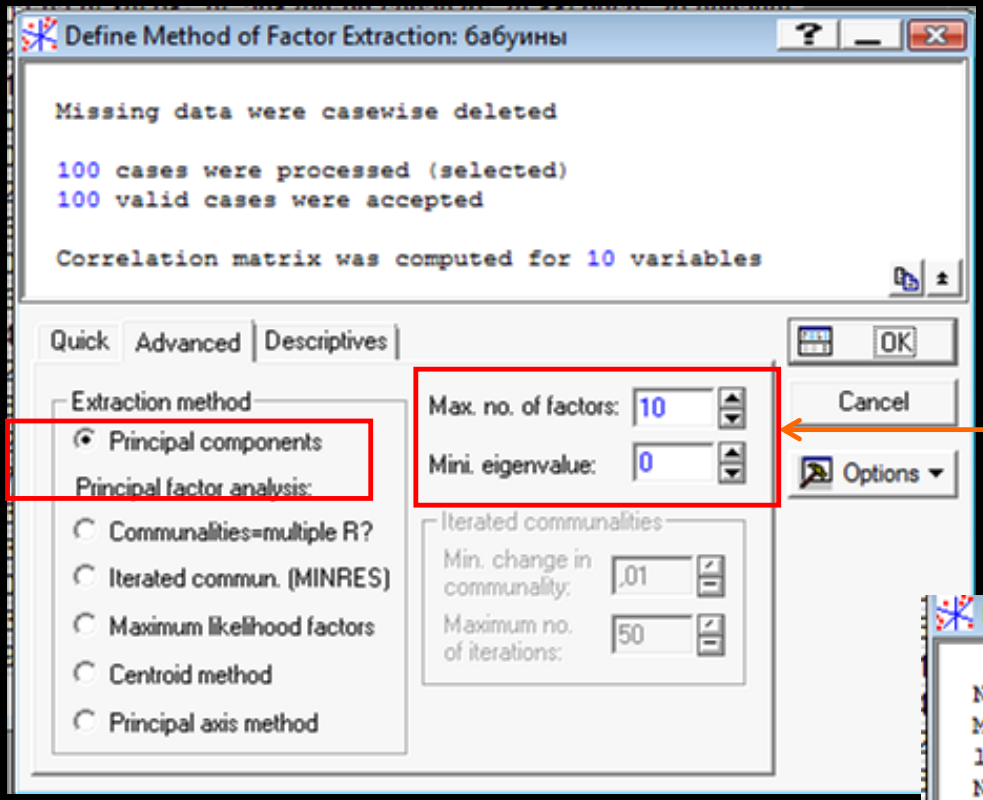




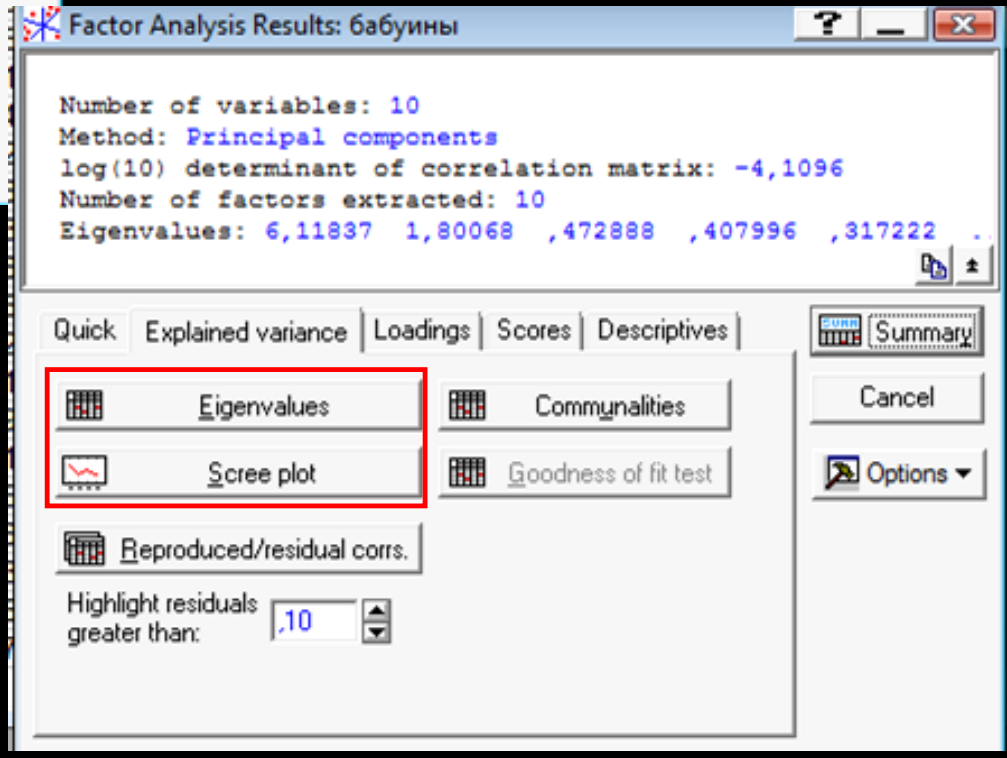


Подготовка: проверка на соответствие нормальному распределению; поиск аутлаеров; поиск чрезмерно сильных корреляций переменных (r>0.95)

PCA



Можно задать min кол-во дисперсии, которое должен объяснять фактор, чтобы его включили в анализ. Обычно min = 1, что соответствует средней дисперсии одной переменной, т.к. матрица корреляций стандартизирует переменные, и сумма дисперсий в них = сумме единиц на главной диагонали = числу переменных



Сперва оценим качество новых компонент: eigenvalues

PCA

values (бабуины)

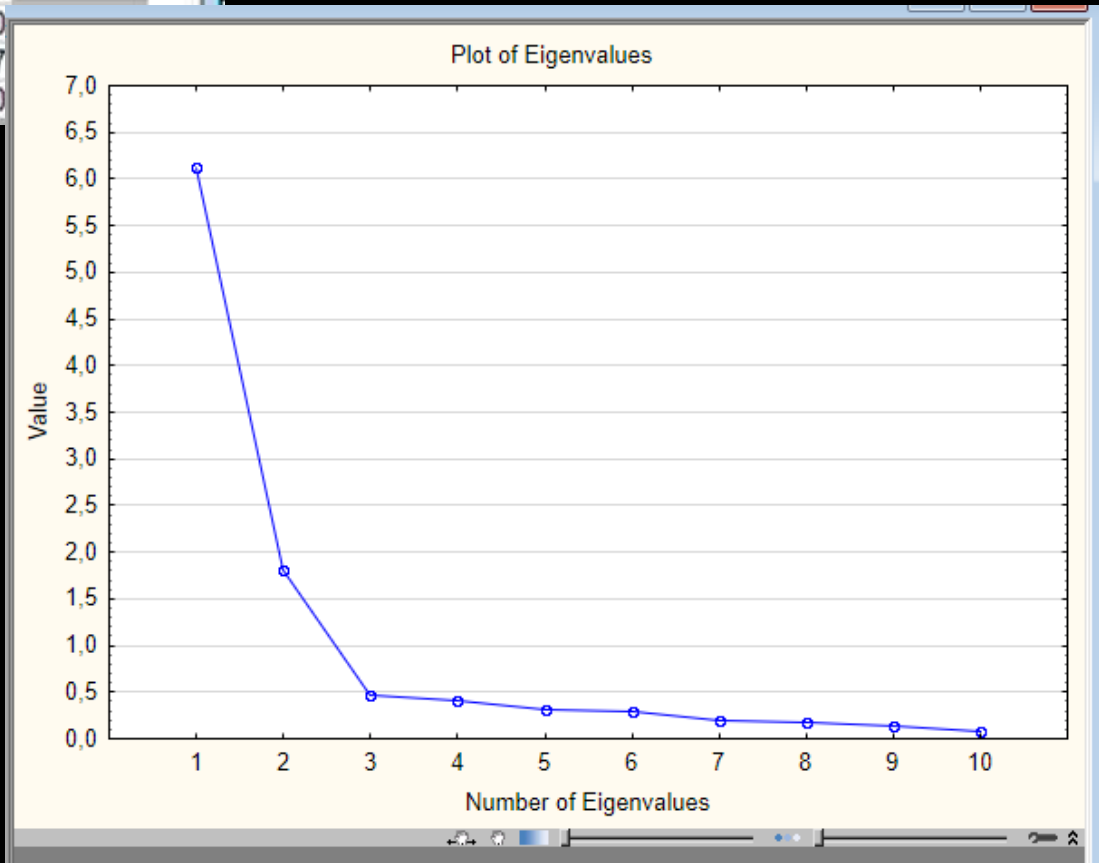
Eigenvalues (бабуины)
Extraction: Principal components

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

В публикацию

Здесь имеет смысл оставлять только 2 компоненты (объясняют 79% всей изменчивости) – их и оставляем.

Scree plot: показывает какую долю всей изменчивости объясняет каждая компонента. Рекомендуют оставлять компоненты выше излома, включая саму точку излома.



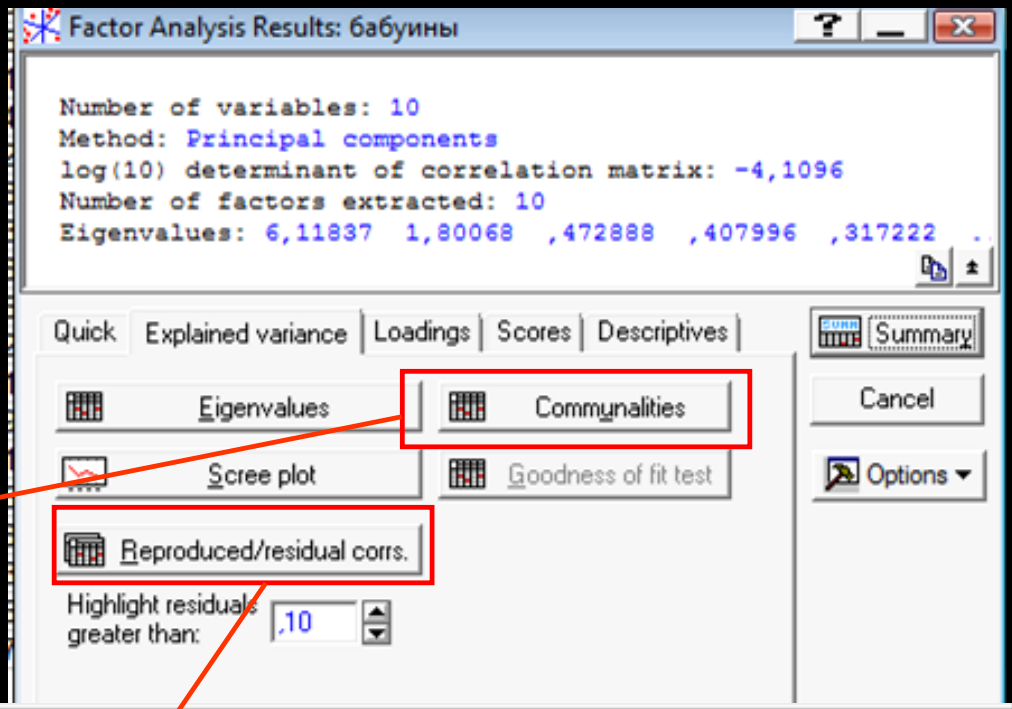
PCA

Оценка качества модели из 2-х компонент

Communalities (бабуины)
Extraction: Principal components
Rotation: Unrotated

Variable	From 1 Factor	From 2 Factors	Multiple R-Square
апельсины, г	0,425888	0,690307	0,559765
бананы, г	0,573012	0,817809	0,734614
яблоки, г	0,556077	0,764634	0,653962
помидоры, г	0,886667	0,887143	0,865921
огурцы, г	0,766702	0,769369	0,739078
мясо, г	0,331847	0,697844	0,541367
курица, г	0,450630	0,832506	0,738620
рыба, г	0,411564	0,740954	0,584202
насекомые, г	0,905382	0,905565	0,884012
червяки, г	0,810599	0,812918	0,779456

Communalities: для каждой переменной показывают, какая доля её изменчивости объясняется нашими факторами



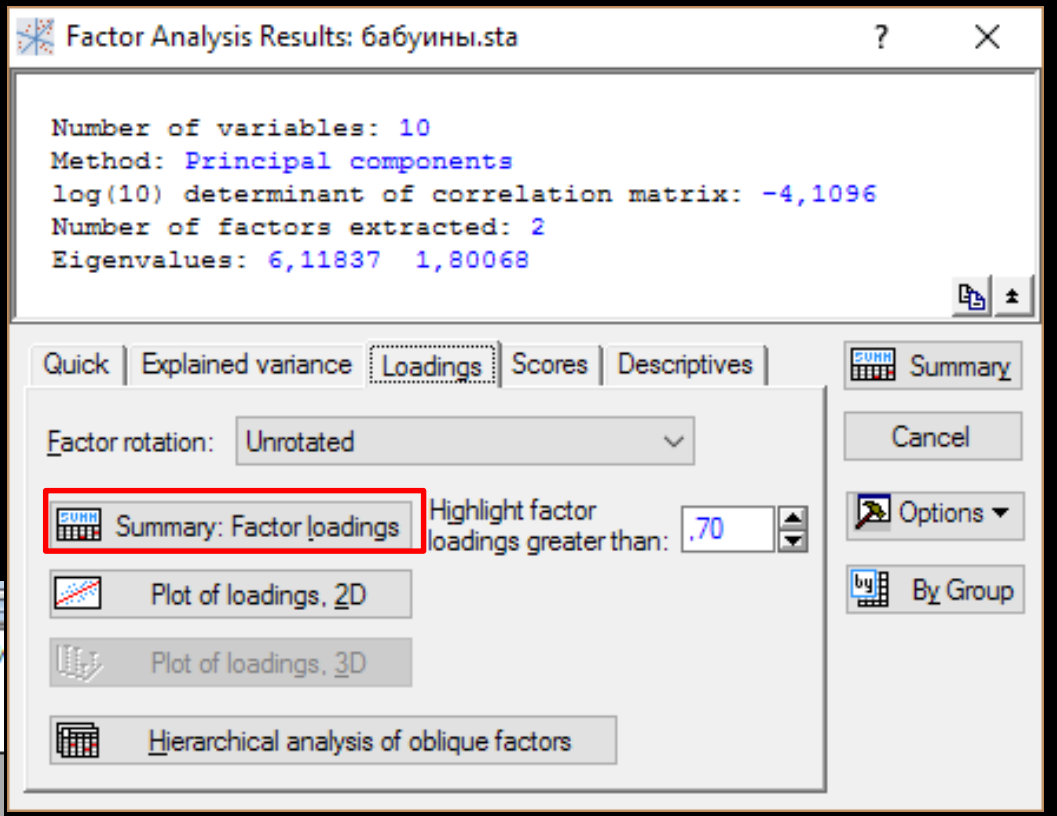
Residual Correlations (бабуины)
Extraction: Principal components
(Marked residuals are > ,100000)

Variable	апельсины, г	бананы, г	яблоки, г	помидоры, г	огурцы, г	мясо, г
апельсины, г	0,31	-0,10	-0,07	-0,01	-0,08	0,08
бананы, г	-0,10	0,18	-0,06	-0,01	0,01	0,01
яблоки, г	-0,07	-0,06	0,24	-0,06	-0,05	0,01
помидоры, г	-0,01	-0,01	-0,06	0,11	-0,02	-0,02
огурцы, г	-0,08	0,01	-0,05	-0,02	0,23	0,03
мясо, г	0,08	0,01	0,01	-0,02	0,03	0,30
курица, г	0,02	-0,02	0,02	-0,01	-0,06	-0,10
рыба, г	0,01	0,03	0,04	-0,03	-0,05	-0,13
насекомые, г	-0,02	-0,02	-0,02	0,01	-0,02	-0,04
червяки, г	-0,06	-0,02	-0,02	-0,00	-0,04	-0,06

Остаточные корреляции между переменными, не объясненные моделью; чем они больше, тем хуже модель.

PCA

Интерпретация компонент (факторов): посмотрим, какие переменные дают в них наибольший вклад.



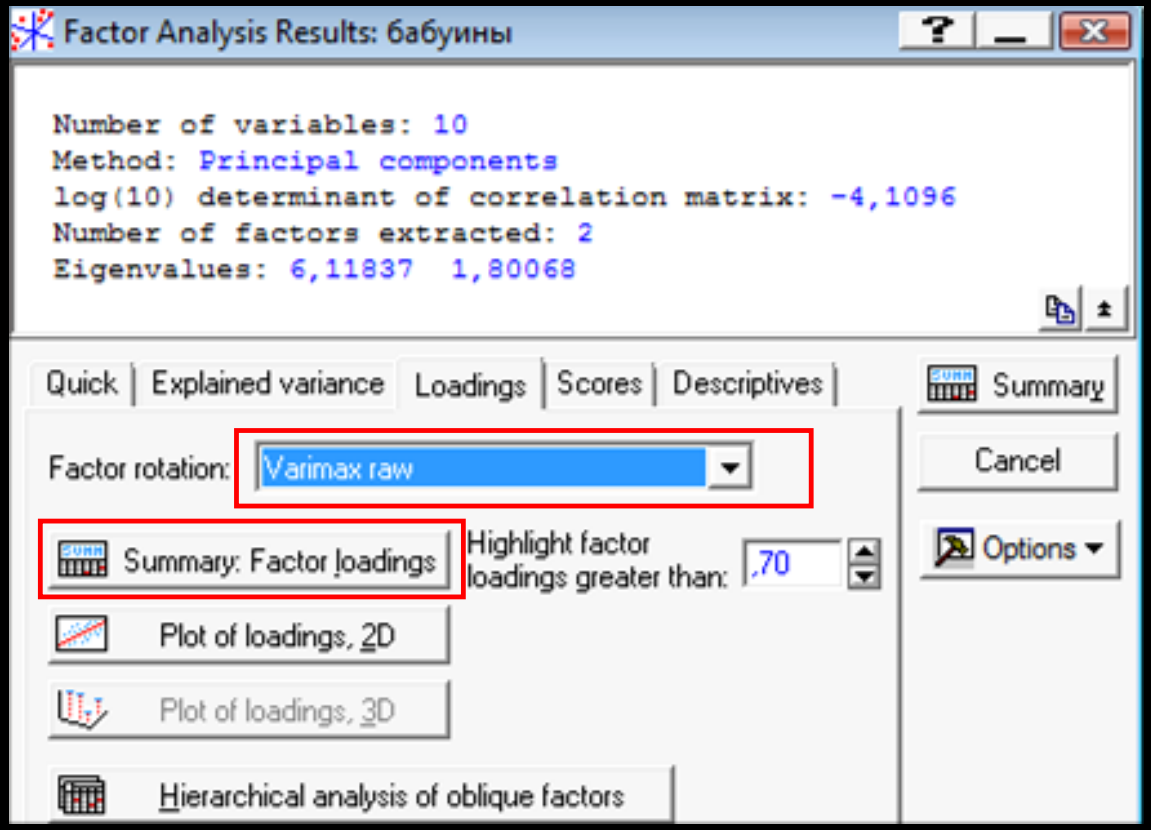
Factor Loadings (Unrotated) (бабуи
Extraction: Principal components
(Marked loadings are >,700000)

Variable	Factor 1	Factor 2
апельсины, г	-0,652601	0,514217
бананы, г	-0,756976	0,494770
яблоки, г	-0,745706	0,456680
помидоры, г	-0,941630	-0,021835
огурцы, г	-0,875615	0,051643
мясо, г	-0,576062	-0,604977
курица, г	-0,671289	-0,617962
рыба, г	-0,641532	-0,573925
насекомые, г	-0,951516	0,013513
черви, г	-0,900333	0,048154
Expl.Var	6,118369	1,800682
Prp.Totl	0,611837	0,180068

Это loadings – корреляции компонент с исходными переменными

Вращение компонент

Проведём вращение, чтобы улучшить их структуру (метод вращения необходимо указать в публикации).



Varimax raw – самый распространённый способ поворота компонент

PCA

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

В публикацию

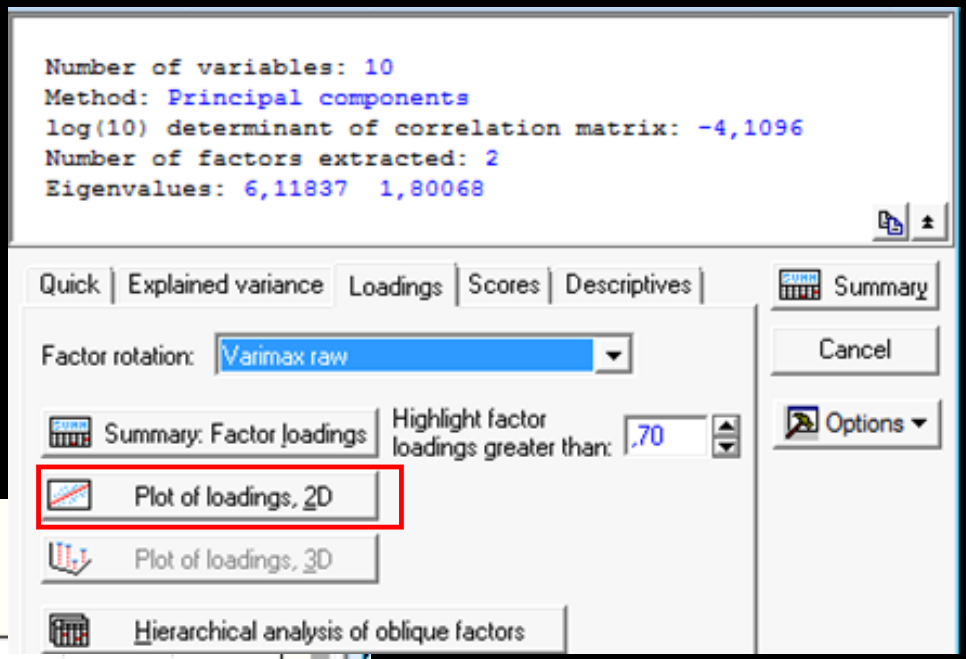
Фактор 1 в основном связан с растительной пищей, фактор 2 – с животной.

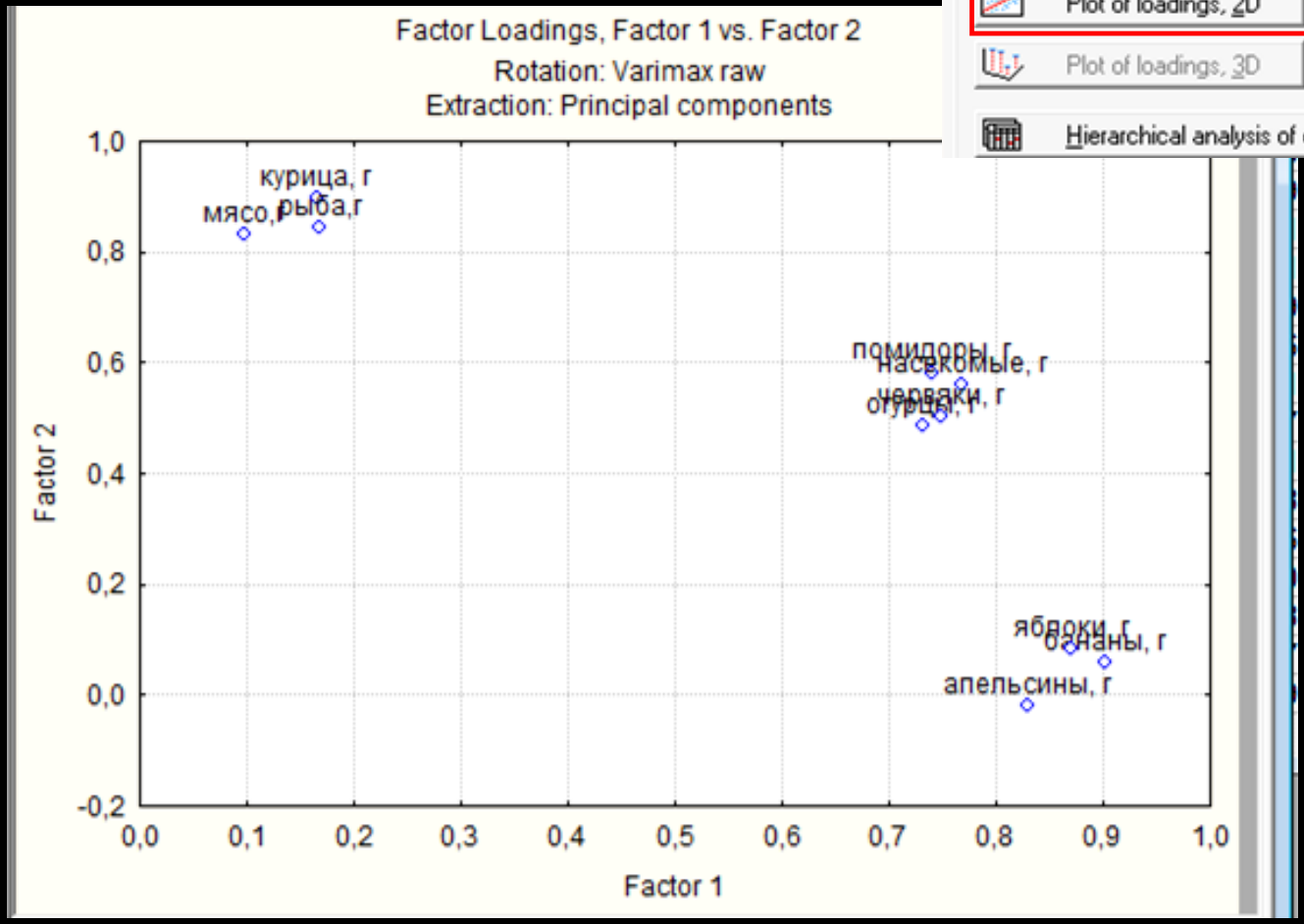
Итак, пищевые предпочтения павианов составлены из двух основных факторов – отношением к животной и растительной пище.

Если вращение компонент не улучшило интерпретируемость компонент, лучше оставить исходные компоненты

PCA

Посмотрим, как исходные переменные расположились в пространстве новых факторов

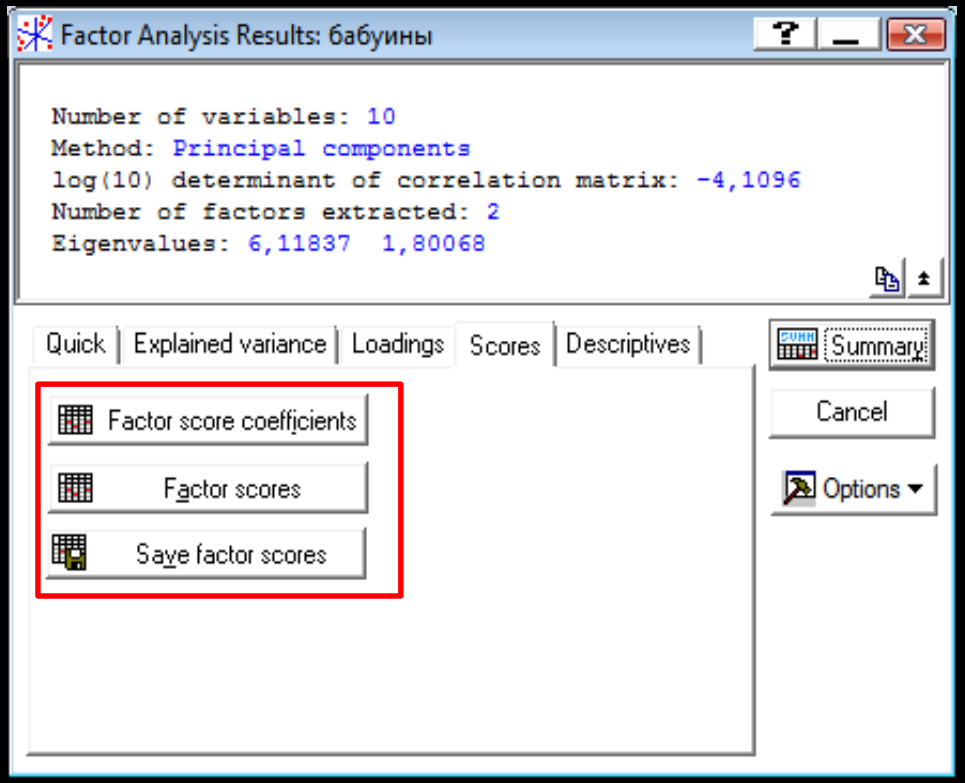


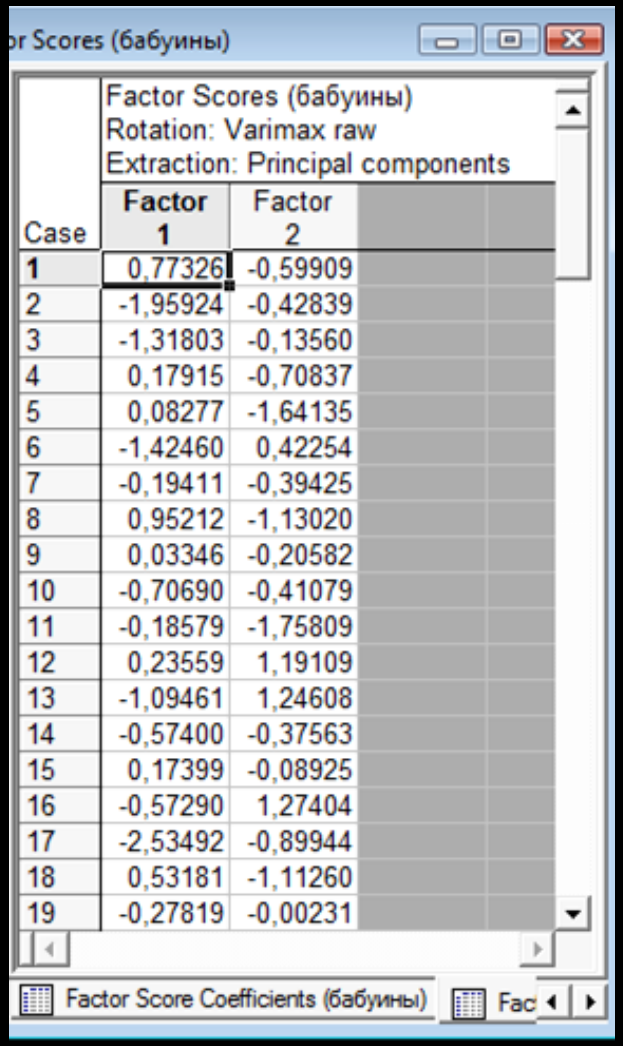


PCA

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

Коэффициенты показывают вклад переменных в компоненту.





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)

Они не коррелируют между собой (удобно в анализе), их легко интерпретировать

Для публикаций

В методы: использовали метод главных компонент; применяли к ним вращение Varimax raw. Иногда весть PCA идёт в раздел методов, когда с его помощью получают новые переменные.

“To reduce the dimensionality in the data set and to simplify the analysis, we applied principal components analysis (PCA) on variables …
All variables were standardized, and two principal components were extracted after Varimax raw rotation.”

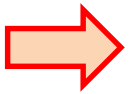
В результаты: собственные значения выделенных компонент; доля изменчивости, которую они объясняют; расшифровка, какие исходные переменные они отражают (не забыть указать увеличивается или уменьшается переменная с ростом компоненты).

“…The two components explained 86% of the variance in the spatiotemporal distribution of males and females. PC1 reflected an increase in the number of adult males within 100 m…”

PCA

Замечание:

На протяжении всей процедуры PCA не было протестировано НИ ОДНОЙ статистической гипотезы.



✓ Из обычного PCA нельзя сделать **никаких качественных выводов** (что-то равно чему-то, А зависит от Б, и проч.);

✓ исследование структуры многомерных данных в PCA – **описательная** процедура;

✓ часто PCA - **предварительный этап** анализа; полученные компоненты идут как переменные в другие статистические процедуры;

✓ **случайные ошибки**, аутлаеры могут радикально изменить структуру и значения компонент, их интерпретацию, и никто этого не заметит.

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.

Расширенный вариант PCA в программе

Данные – структура сообщества животных Африки в разных условиях

Можно выбрать матрицу ковариаций вместо матрицы корреляций (если важно сохранить в модели различия в дисперсии переменных)

PCA

Select variables for analysis, supplementary, active case, and group

1	2	3	4
8 - Veg_liana	1 - TransectID	1 - TransectID	1 - TransectID
9 - Veg_DBH	2 - Distance	2 - Distance	2 - Distance
10 - Veg_Canopy	3 - HuntCat	3 - HuntCat	3 - HuntCat
11 - Veg_Understory	4 - NumHouseholds	4 - NumHouseholds	4 - NumHouseholds
12 - RA_Apes	5 - LandUse	5 - LandUse	5 - LandUse
13 - RA_Birds	6 - Veg_Rich	6 - Veg_Rich	6 - Veg_Rich
14 - RA_Elephant	7 - Veg_Stems	7 - Veg_Stems	7 - Veg_Stems
15 - RA_Monkeys	8 - Veg_liana	8 - Veg_liana	8 - Veg_liana
16 - RA_Rodent	9 - Veg_DBH	9 - Veg_DBH	9 - Veg_DBH
17 - RA_Ungulate	10 - Veg_Canopy	10 - Veg_Canopy	10 - Veg_Canopy
18 - Rich_AllSpecies	11 - Veg_Understory	11 - Veg_Understory	11 - Veg_Understory
Spread / Zoom	Spread / Zoom	Spread / Zoom	Spread / Zoom
Variables for analysis: 12-17	Supplementary variables: 6-10	Active cases variable:	Grouping variable: 5

OK | 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 c

Principal Components and Classification Analysis Results: sprcies di...

```
No. of active vars:6              No. of supplementary vars:5
No. of active cases: 24           No. of supplementary cases: 0

Eigenvalues: 3,05143  1,00845  ,906915  ,706142  ,327046  ...
```

Number of factors : 3 | Quality of representation : 82,8 %

OK | Cancel | Options | By Group

Quick | Variables | Cases | Descriptives

- Factor coordinates of variables
- Factor & variable correlations
- Plot var. factor coordinates, 2D
- Communalities (Cosine ?)
- Contributions of variables
- Eigenvalues
- Scree plot
- Eigenvectors

Options for plot of factor coord.

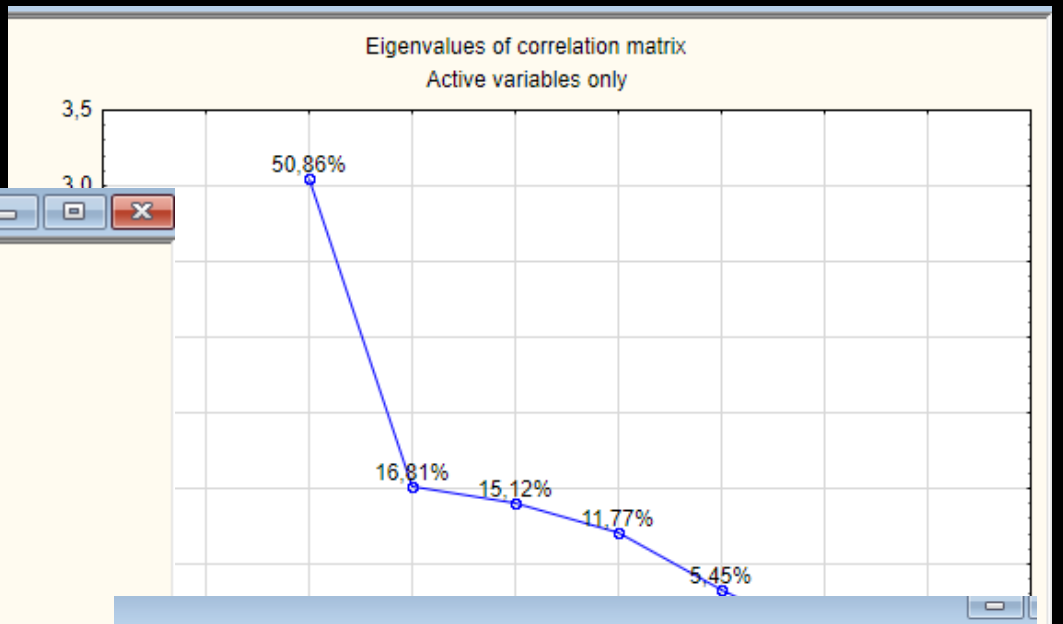
- [x] Vectors (points to origin)
- [x] Unit circle
- (•) Variable names
- ( ) Variable numbers
- ( ) No Names/Numbers

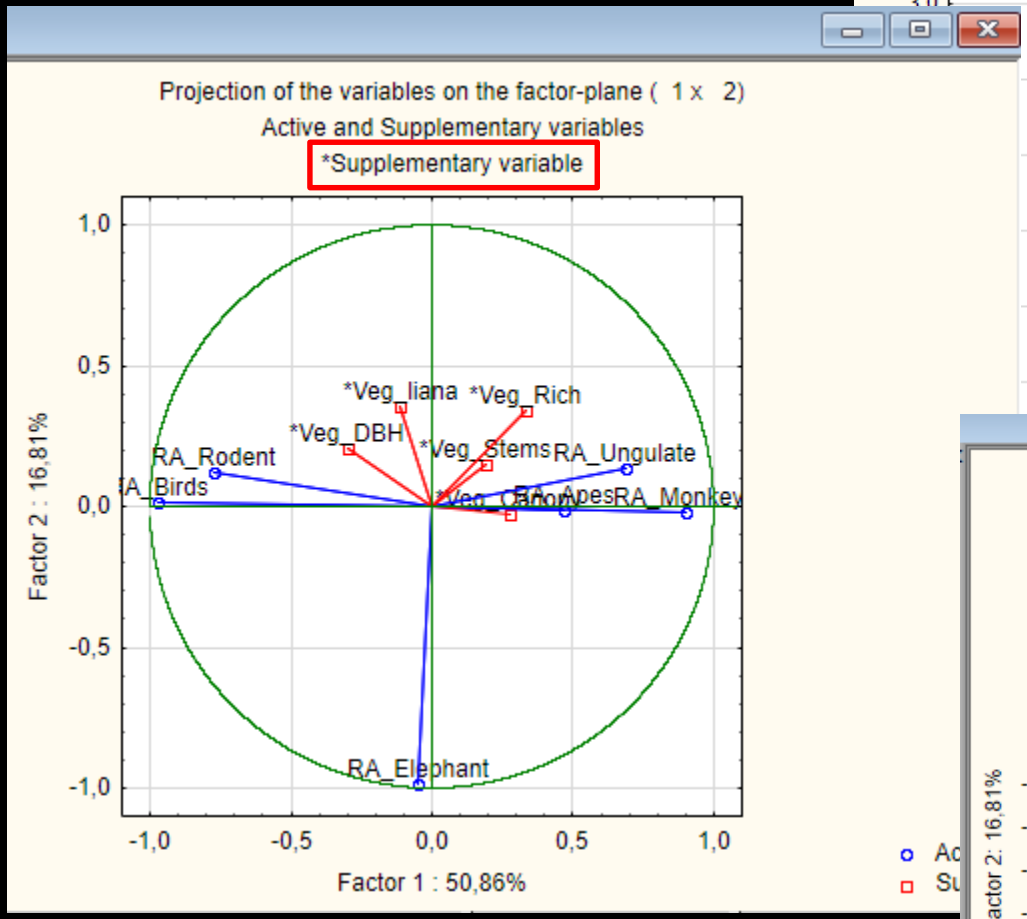
Есть возможность сразу задать набор дополнительных переменных и группирующие переменные

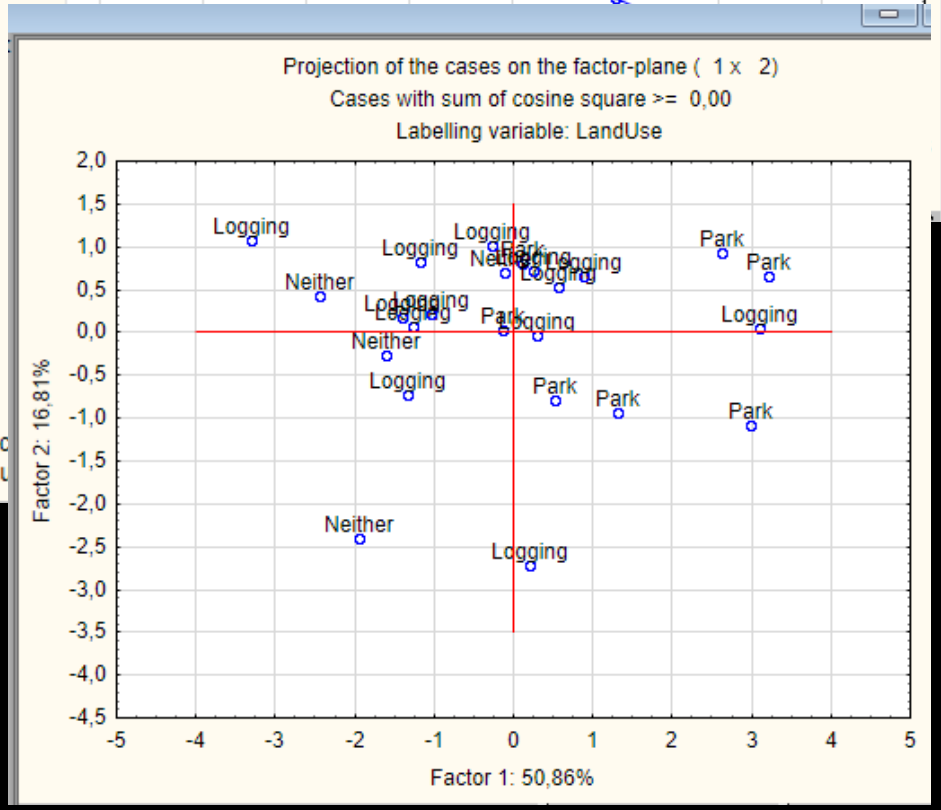
Сразу видно, как меняется качество модели с уменьшением числа компонент; отличные картинки, с дополнительными и группирующими переменными.

PCA

Более информативное представление компонент







Возможность представить объекты в пространстве новых компонент с ярлыками групп из группирующей переменной

Связь PCA с MANOVA и регрессионным анализом.

1. Если мы хотим сравнить группы объектов, и зависимых переменных много, можно провести MANOVA (преимущество: можно сразу анализировать действие и нескольких группирущих переменных на несколько зависимых одним движением руки), а можно – сначала PCA, а потом – ANOVA (преимущество: возможность исследовать структуру зависимых переменных и выделить несколько компонент).

2. Если мы хотим провести множественный регрессионный анализ, а между предикторами присутствуют серьёзные корреляции, можно сначала сделать PCA для независимых переменных (можно даже без сокращения их числа), а потом – регрессионный анализ, и проблема скоррелированности предикторов исчезнет.

Correspondence analysis

- ✓Метод анализа сложных таблиц сопряжённости.
- ✓Анализ взаимосвязи 2-х категориальных переменных.
- ✓Если у каждой переменной по 2 категории (2x2 таблица) – точный критерий Фишера;
- ✓>2 категорий – критерий Хи-квадрат, но его результат для >4 категорий трудно интерпретировать!
- ✓Для таких таблиц *n* x *p*, где *n* и *p* – кол-во категорий в ДВУХ качественных переменных, придуман СА.
- ✓ ЦЕЛЬ – представить связь между двумя переменными, уменьшив число категорий в одной из переменных и потеряв как можно меньше информации;

Примеры:

- ✓сравнение видового состава в нескольких регионах;
- ✓сравнение соотношений разных социальных контактов у разных половозрастных групп;
- ✓соотношение разных фенотипов в разных популяциях…

Correspondence analysis

✓эти новые категории рассматриваются как новые «переменные» (столбцы, категории переменной A); строки таблицы (категории переменной B) рассматриваются как «объекты».

✓ основа – хи-квадрат статистика; строится матрица, её элементы содержат разности наблюдаемых и ожидаемых частот.

Переменная A: приверженность к курению;
Переменная B: должность.

	Smoking Category				
Staff Group	(1) None	(2) Light	(3) Medium	(4) Heavy	Row Totals
(1) Senior Managers	4	2	3	2	11
(2) Junior Managers	4	3	7	4	18
(3) Senior Employees	25	10	12	4	51
(4) Junior Employees	18	24	33	13	88
(5) Secretaries	10	6	7	2	25
Column Totals	61	45	62	25	193

Correspondence analysis

✓В результате преобразования матриц получается таблица с меньшим числом столбцов (=«переменных» = категорий переменной A);

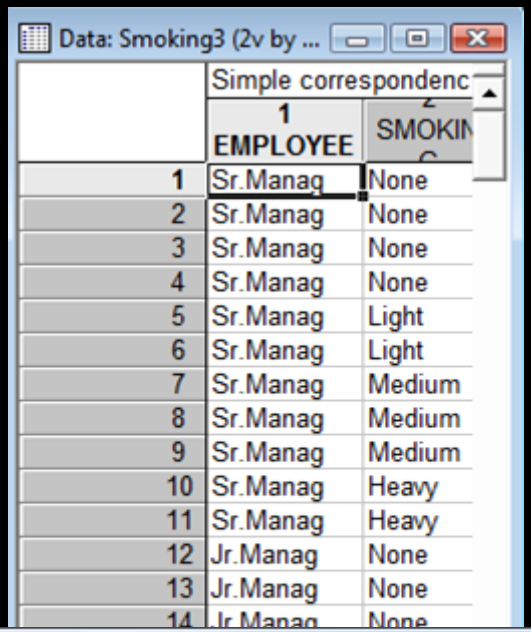
✓Сумма элементов главной диагонали - **Inertia**=Chi-square/Total N (вместо суммы дисперсии);

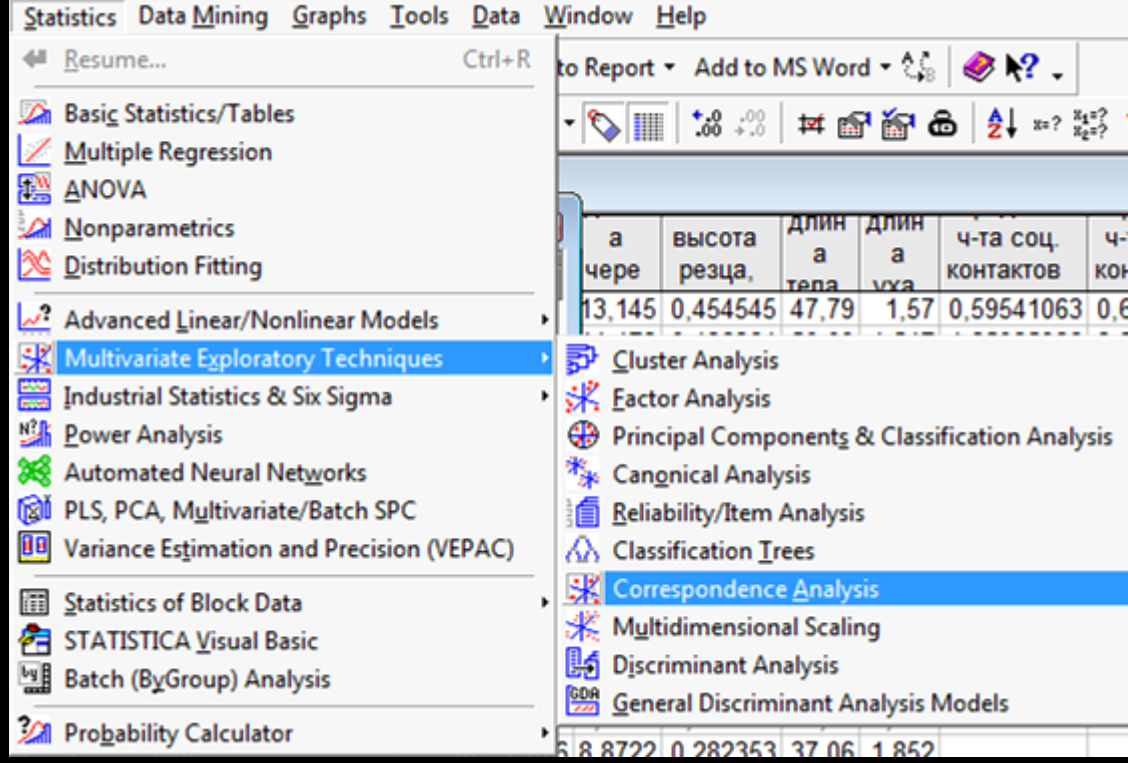
✓для каждого столбца получается eigenvalue – доля общей Inertia, которая на него приходится

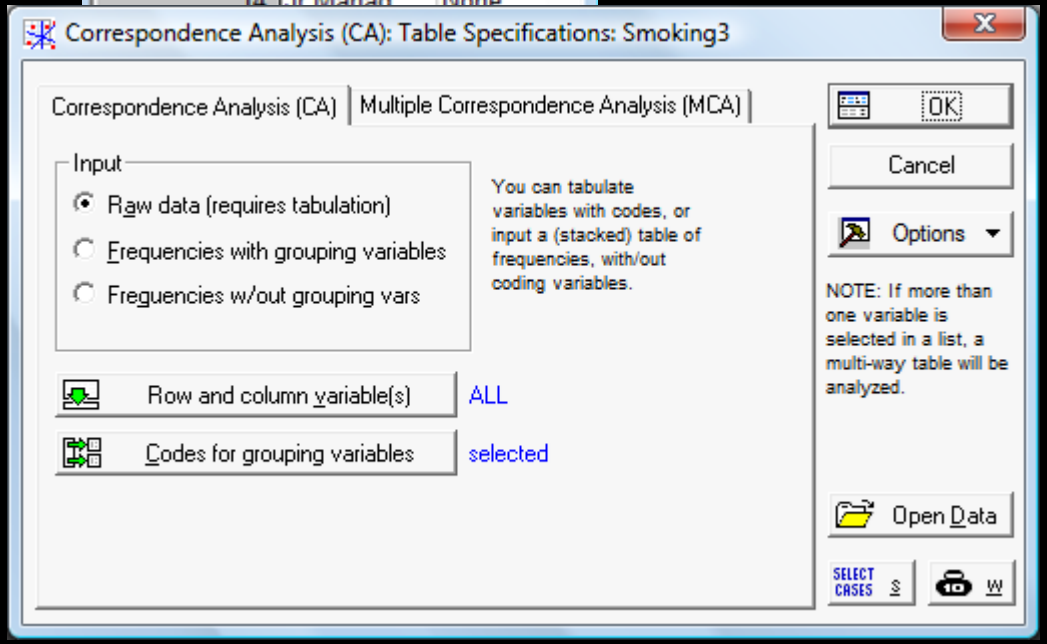
✓В результат строится картинка, где «объекты» (строки = категории переменной B) располагаются в пространстве новых категорий переменной A.

✓Чисто описательная процедура, хороша для построения картиной и обсуждения данных.

Correspondence analysis







Главная информация – дистанции между рядами («объектами») в пространстве новых компонент;
Интерпретация компонент – по тому, какие группы они лучше всего разделяют.

Взаиморасположение объектов в пространстве новых переменных

В публикацию

Correspondence Analysis Results: Smoking3

```
Variables and number of categories:
      Row variables: EMPLOYEE(5)
   Column variables: SMOKING(4)
Eigenvalues: ,0748 ,0100 ,0004
Total chi-square=16,4416  df=12 p=,1719
```

Quick | Advanced | Options | Review | Supplementary points

Summary

Cancel

Options

Row and column coordinates

Print/report Summary Box

Eigenvalues | Plot

Unstandardized matrices

Plots of coordinates

Row, 1D | 2D | 3D

Column, 1D | 2D | 3D

Row & col., 1D | 2D | 3D

Plot selected dimensions only

Truncate labels to 2 chars

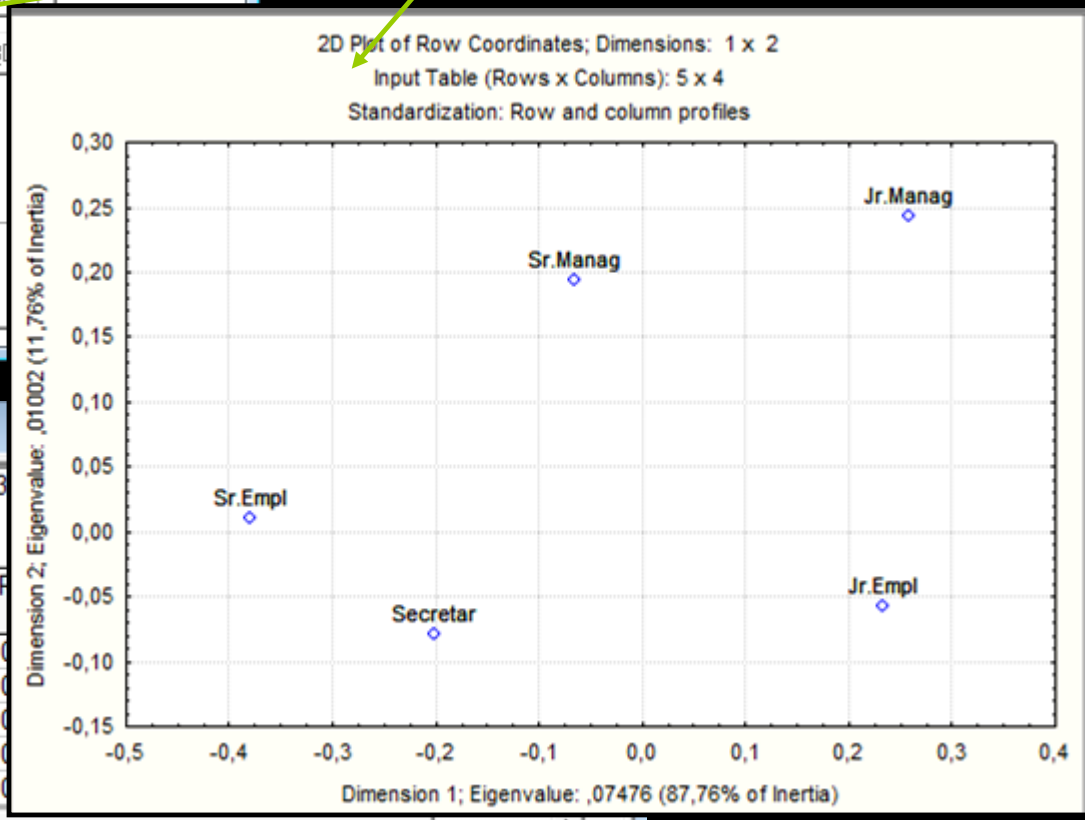
Use identical X/Y/(Z) scales

oordinates and Contributions to Inertia (Smoking3)

Row Coordinates and Contributions to Inertia (Smoking3)
Input Table (Rows x Columns): 5 x 4
Standardization: Row and column profiles

Row Name	Row Number	Coordin. Dim.1	Coordin. Dim.2	Mass	Quality	
Sr.Manag	1	-0,065768	0,193737	0,056995	0,892568	
Jr.Manag	2	0,258958	0,243305	0,093264	0,991082	
Sr.Empl	3	-0,380595	0,010660	0,264249	0,999817	
Jr.Empl	4	0,232952	-0,057744	0,455959	0,999810	
Secretar	5	-0,201089	-0,078911	0,129534	0,998603	

Row Coordinates and Contributions to Inertia (Smoking3) | Column Coordinates and Contributions to Inertia (Smoking3) | Plot of E



Canonical analysis

Многомерный анализ, позволяет проанализировать взаимосвязь между **двумя наборами** переменных: (несколько количественных независимых – несколько количественных зависимых).

Основа для анализа – **матрицы корреляций** внутри каждого из наборов переменных и корреляций всех пар переменных из разных наборов.

Например, можно проверить взаимосвязи между:

✓Набором морфологических промеров и физиологическими измерениями;

✓ набором внешних факторов и набором симптомов;

✓ видовым составом и экологическими переменными

✓ …

Canonical analysis

Концепция очень близка PCA.

✓в рез-те операций над этими матрицами получаются canonical **roots** – пары компонент (по одной из каждого набора) такие, что в первой паре корреляция максимальна, во второй – меньше и т.д.

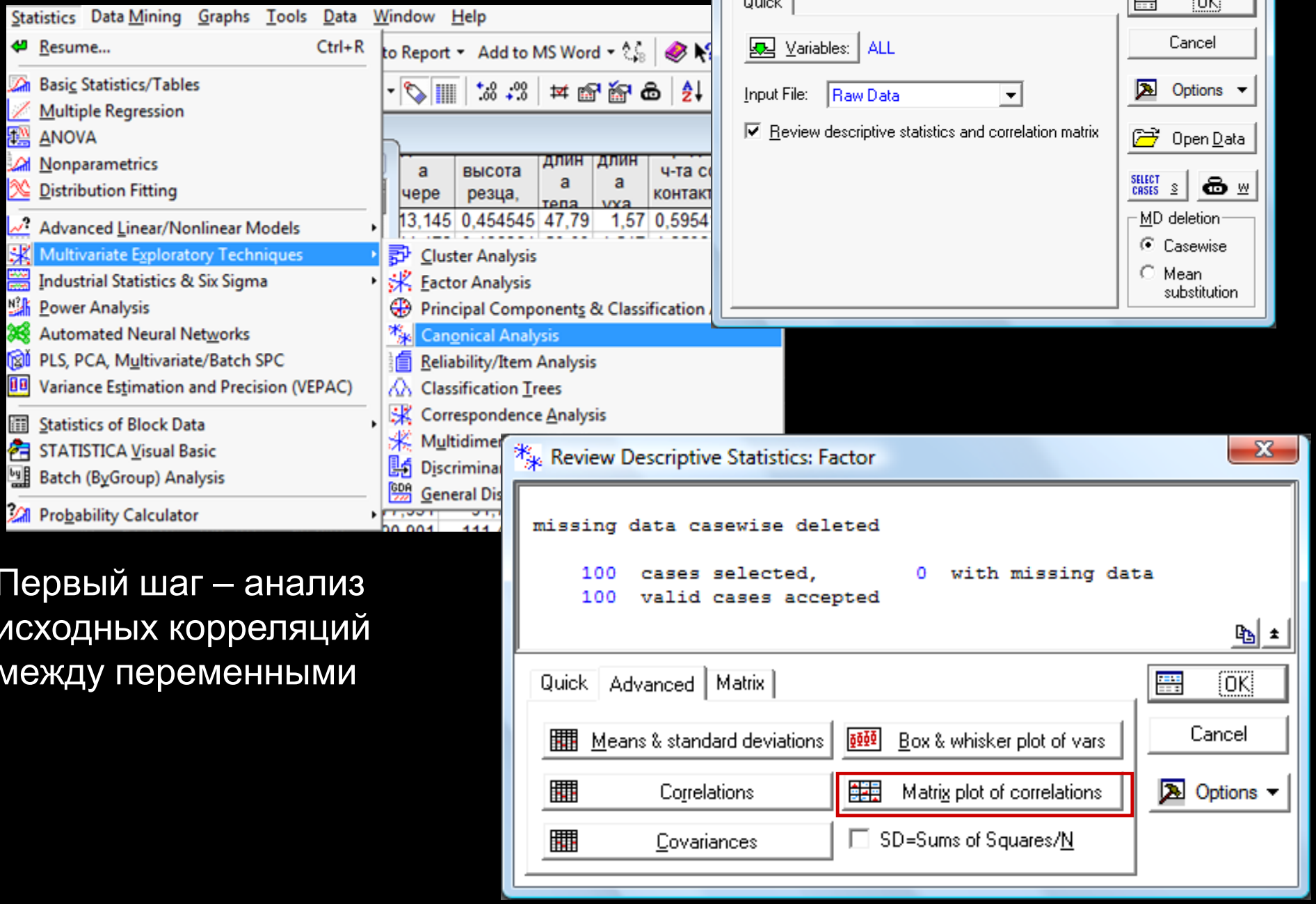
✓Компоненты, составляющие **roots -** линейные комбинации исходных переменных.

✓Eigenvalues – доля изменчивости, объяснённая корреляцией соответствующих компонент.

✓Интерпретация канонических корней: **canonical weights** = eigenvector, показывают вклад каждой переменной в данный корень; **factor structure** = factor loadings, показывают корреляции корней с исходными переменными.

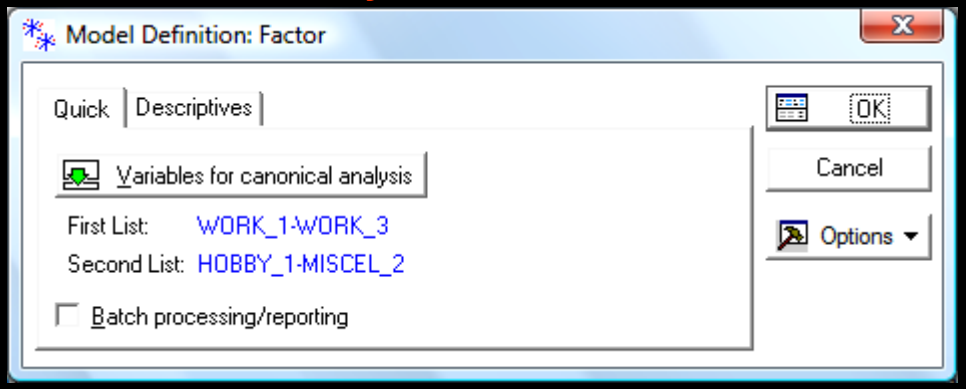
В примере – корреляция между разными показателями удовлетворения работой и удовлетворения хобби и домашними делами.

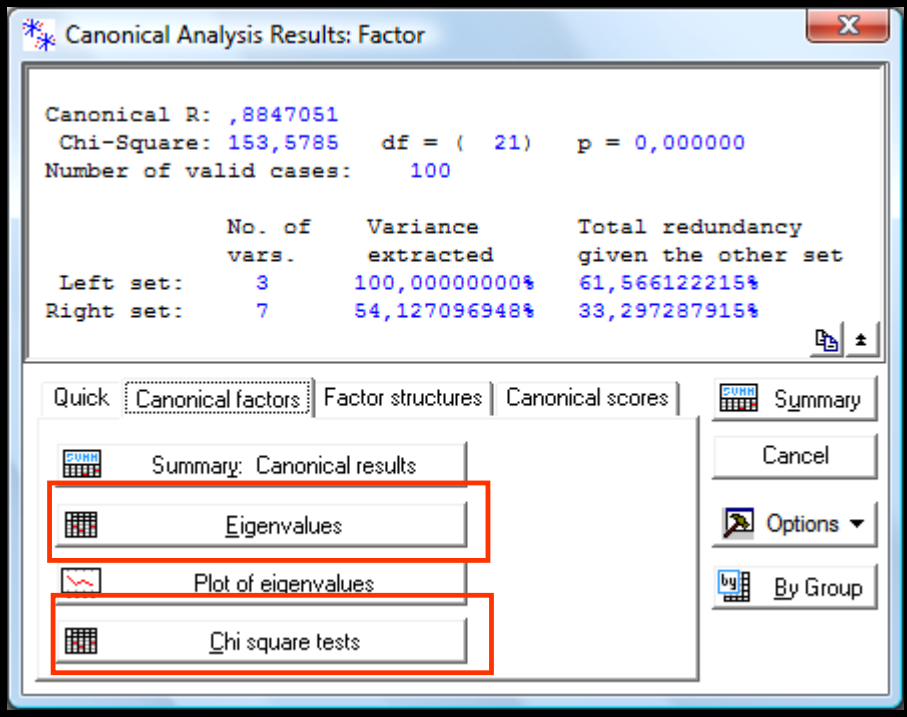
Canonical analysis



Первый шаг – анализ исходных корреляций между переменными

Canonical analysis





Root	Eigenvalues (Factor.sta)		
	Root 1	Root 2	Root 3
Value	0,782703	0,073484	0,038957

Eigenvalues – показывают, какую долю общей изменчивости в наборах объясняет данная корреляция компонент (корень).

Root Removed	Chi-Square Tests with Successive Roots Removed (Factor.sta)					
	Canonicl R	Canonicl R-sqr.	Chi-sqr.	df	p	Lambda Prime
0	0,884705	0,782703	153,5785	21	0,000000	0,193486
1	0,271080	0,073484	10,8516	12	0,541689	0,890422
2	0,197374	0,038957	3,7153	5	0,591097	0,961043

Значимость корреляций между наборами переменных для каждого корня; измеряется сначала для всех корней, потом после удаления 1-го корня и т.п.

Canonical analysis

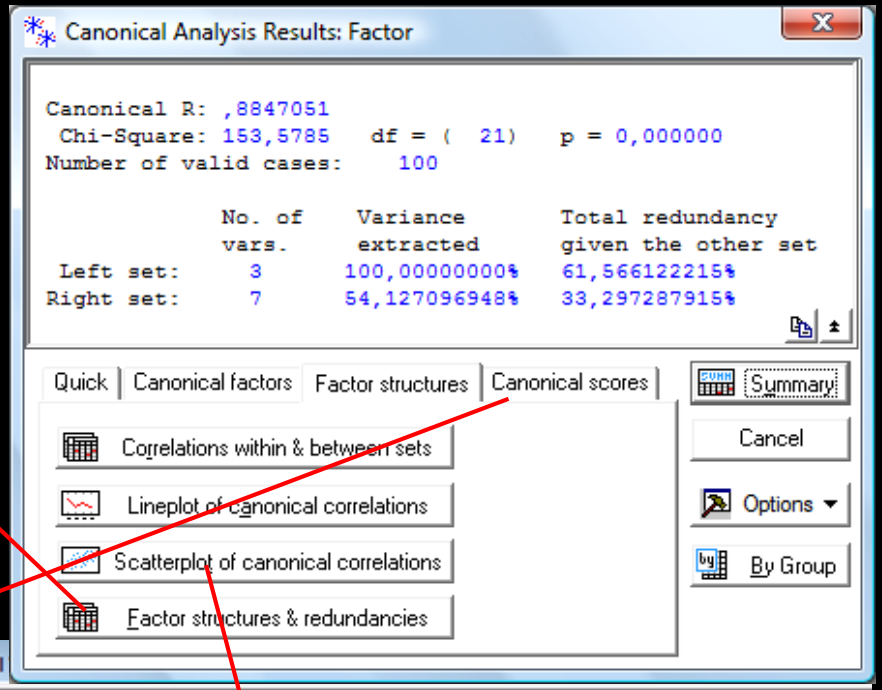
Variable	Factor Structure, left set (Factor.sta) Root 1	Root 2	Root 3
WORK_1	0,796461	-0,575348	0,186075
WORK_2	0,952643	0,100725	-0,286926
WORK_3	0,875390	0,217192	0,431880

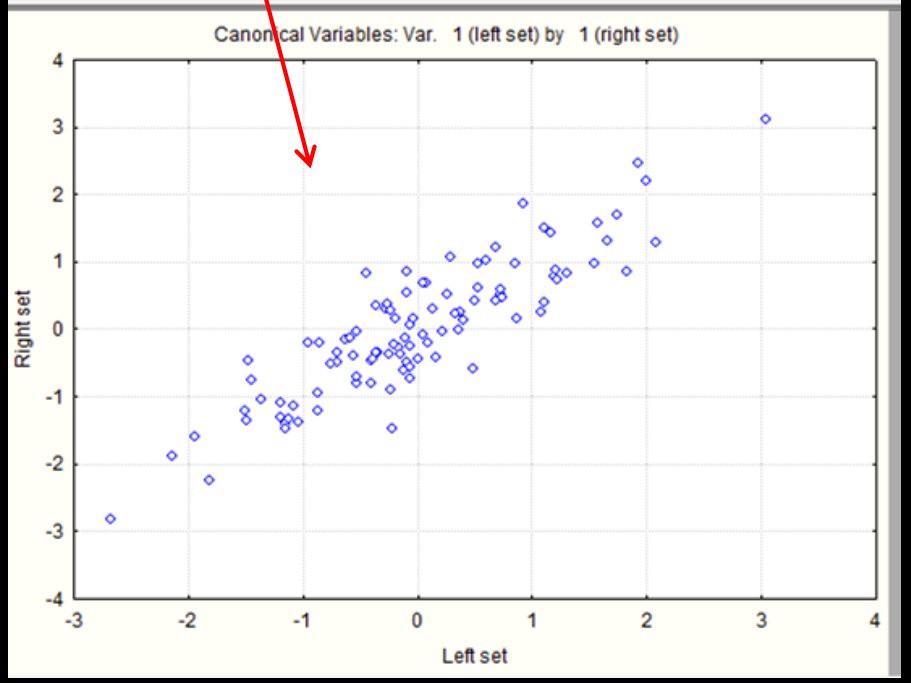
Корреляции корней с исходными переменными (loadings)

Variable	Canonical Weights, left set (Factor.sta) Root 1	Root 2	Root 3
WORK_1	0,217021	-1,35890	0,24350
WORK_2	0,592485	0,37139	-1,38770
WORK_3	0,300124	0,83222	1,28861

Вклад каждой переменной в соответствующий корень (eigenvectors)

Redundancy – доля изменчивости, которая объясняется данной корреляцией





Требования к выборкам для канонического анализа

1. Внутри переменных должно быть многомерное *нормальное распределение* (оценка – на основе гистограмм);
2. Связь переменных должна быть *линейной*;
3. Размер выборки не должен быть меньше 50, оптимальный – от 20 х число переменных.
4. Важно исключить аутлаеры
5. коэффициентов корреляции, близких *единице*, быть не должно.

Для публикаций

В методы: использовали канонический анализ; переменные соответствовали условиям нормального распределения.

В результаты:

✓собственные значения полученных корней;

✓значимость корреляций для каждого корня (Chi-square test): Canonical R, Chi-square, df, p;

✓доля изменчивости, которую объясняет модель в каждом наборе (variance extracted);

✓Интерпретация корней: какие исходные переменные отражают (loadings = factor structure для каждого набора).