



Тема 7

Статистические методы снижения размерности признакового пространства



Снижение размерности признакового пространства обусловлено:

- необходимость наглядного представления данных
- стремление к лаконизму исследуемых моделей:
отбор наиболее информативных показателей
- необходимость существенного сжатия хранимой статистических данных



Инструментарий снижения размерности признакового пространства:

- метод главных компонент
- методы факторного анализа
- многомерное шкалирование
- отбор существенных предикторов в регрессионном анализе
- отбор типобразующих признаков в дискриминантном анализе
- целенаправленное проецирование и отбор типобразующих признаков в кластер-анализе



Взаимосвязи

Классификация

Снижение
размерности

Факторный анализ - совокупность статистических методов, которые на основе реально существующих связей признаков (или объектов) позволяют выявить латентные обобщающие характеристики организационной структуры и механизма развития изучаемых явлений и процессов.



$$Z_j = a_{j1}F_1 + a_{j2}F_2 + \dots + a_{jn}F_n,$$

где Z_j – стандартизированные переменные

F_1, F_2, \dots, F_n – общие факторы,

$$\text{Var } F_1 \geq \text{Var } F_2 \geq \dots \geq \text{Var } F_n$$

$a_{j1}, a_{j2}, \dots, a_{jn}$ – факторные нагрузки

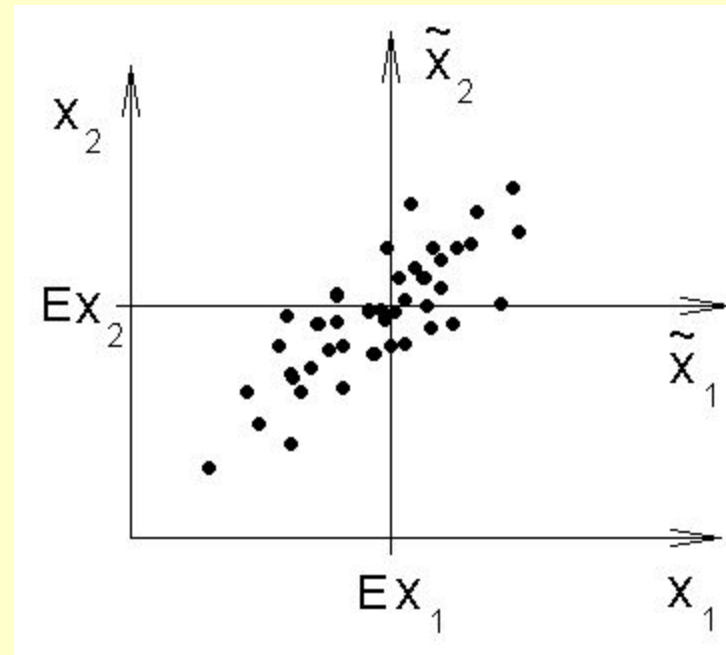
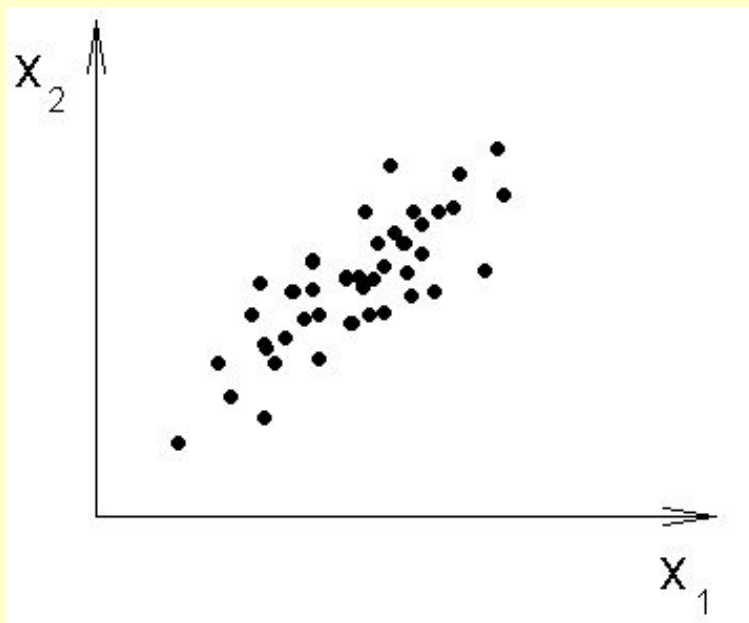


Взаимосвязи

Классификация

Снижение
размерности

Метод главных компонент



Стандартизация

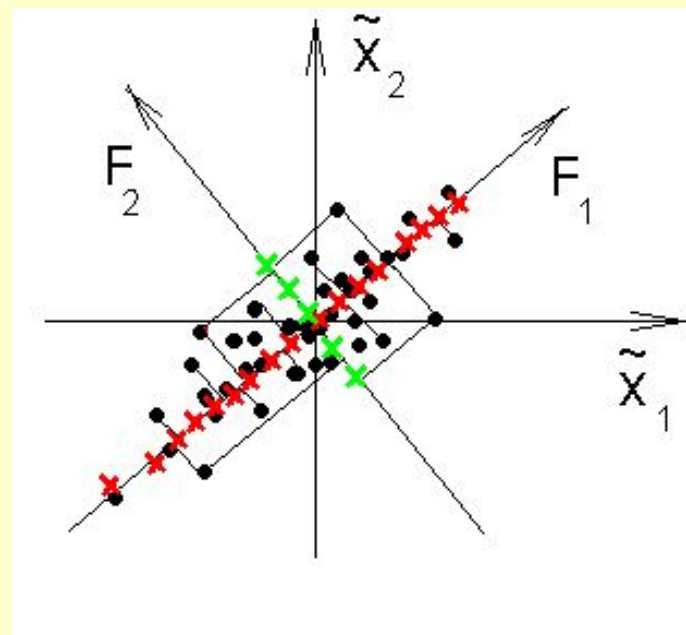
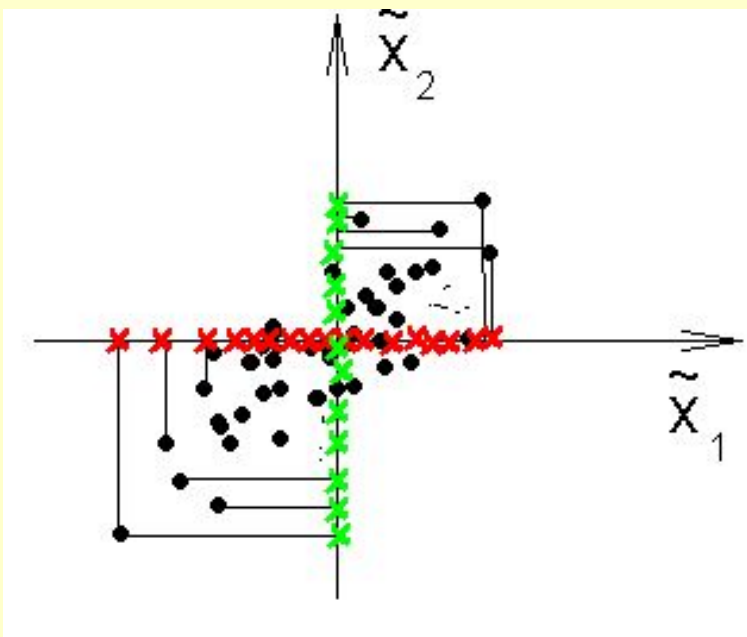


Взаимосвязи

Классификация

Снижение
размерности

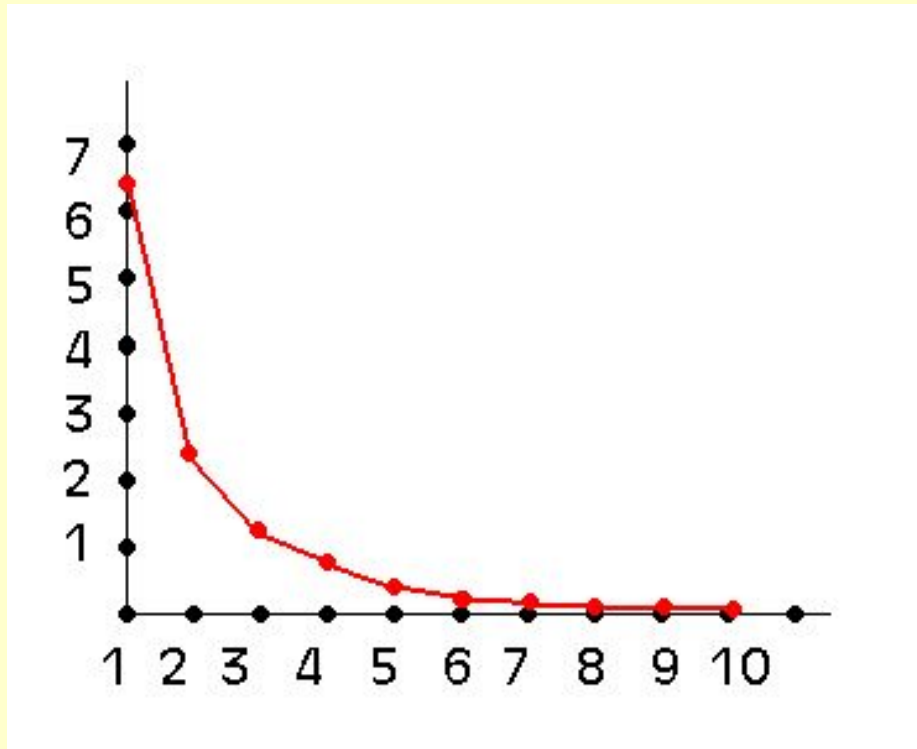
Метод главных компонент



Вращение



Объяснение общей дисперсии



$$\frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \dots + \lambda_{10}} = \frac{\text{Var } F_1 + \text{Var } F_2 + \text{Var } F_3}{\text{общая дисперсия}}$$

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Label	Values	Missing	Columns	Align	Measure
1	марка	String		None	None	8	Left	Nominal
2	цена	Numeric		None	None	8	Right	Scale
3	т_сост	Numeric		None	None	8	Right	Scale
4	возраст	Numeric		None	None	8	Right	Scale
5	пробег	Numeric		None	None	8	Right	Scale
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								

- Reports
- Descriptive Statistics
- Tables
- Compare Means
- General Linear Model
- Mixed Models
- Correlate
- Regression
- Loglinear
- Classify
- Data Reduction**
- Scale
- Nonparametric Tests
- Time Series
- Survival
- Multiple Response
- Missing Value Analysis...
- Complex Samples

- Factor...**
- Correspondence Analysis...
- Optimal Scaling...

Data View Variable View

Factor

SPSS Processor is ready



Взаимосвязи

Классификация

Снижение размерности

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										

Factor Analysis

Variables:

- марка
- цена
- т_сост
- возраст
- пробег

Selection Variable:

OK Paste Reset Cancel Help

Descriptives... Extraction... Rotation... Scores... Options...

Data View Variable View /

SPSS Processor is ready



Взаимосвязи

Классификация

Снижение размерности

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										

Factor Analysis

Variables: марка

Variables: цена, т_сост, возраст, пробег

Selection Variable:

OK, Paste, Reset, Cancel, Help, Descriptives..., Extraction..., Rotation..., Scores..., Options...

Factor Analysis: Descriptives

Statistics: Univariate descriptives, Initial solution

Correlation Matrix: Coefficients, Inverse, Significance levels, Reproduced, Determinant, Anti-image, KMO and Bartlett's test of sphericity

Continue, Cancel, Help

Data View Variable View / SPSS Processor is ready



Взаимосвязи

Классификация

Снижение размерности

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										

Factor Analysis

Variables: марка

Variables: цена, т_сост, возраст, пробег

Selection Variable:

Descriptives... Extraction... Rotation...

Factor Analysis: Extraction

Method: Principal components

Analyze: Correlation matrix, Covariance matrix

Display: Unrotated factor solution, Scree plot

Extract: Eigenvalues over: 1, Number of factors:

Maximum Iterations for Convergence: 25

Data View Variable View / SPSS Processor is ready



Взаимосвязи

Классификация

Снижение размерности

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										

Factor Analysis

Variables: марка

Variables: цена, т_сост, возраст, пробег

Selection Variable: Value...

Buttons: Descriptives..., Extraction..., Rotation..., Scores..., Options..., OK, Paste, Reset, Cancel, Help

Factor Analysis: Extraction

Method: Principal components

Analyze: Correlation matrix, Covariance matrix

Display: Unrotated factor solution, Scree plot

Extract: Eigenvalues over: 1, Number of factors:

Maximum Iterations for Convergence: 25

Buttons: Continue, Cancel, Help

Data View | Variable View /

SPSS Processor is ready



Взаимосвязи

Классификация

Снижение размерности

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										

Factor Analysis

Variables: марка

Variables: цена, т_сост, возраст, пробег

Selection Variable:

Buttons: Descriptives..., Extraction..., Rotation..., Scores..., Options..., OK, Paste, Reset, Cancel, Help

Factor Analysis: Rotation

Method: Varimax, None, Direct Oblimin, Quartimax, Equamax, Promax

Delta: 0, Kappa: 4

Display: Rotated solution, Loading plot(s)

Maximum Iterations for Convergence: 25

Buttons: Continue, Cancel, Help

Data View | Variable View /

SPSS Processor is ready



Взаимосвязи

Классификация

Снижение размерности

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										

Factor Analysis

Variables: марка

Variables: цена, т_сост, возраст, пробег

Selection Variable:

OK, Paste, Reset, Cancel, Help

Descriptives..., Extraction..., Rotation..., Scores..., Options...

Factor Analysis: Factor Scores

Save as variables

Method: Regression, Bartlett, Anderson-Rubin

Display factor score coefficient matrix

Continue, Cancel, Help

Data View Variable View /

SPSS Processor is ready



Взаимосвязи

Классификация

Снижение размерности

cars.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	марка	String	8	0		None	None	8	Left	Nominal
2	цена	Numeric	8	0		None	None	8	Right	Scale
3	т_сост	Numeric	8	0		None	None	8	Right	Scale
4	возраст	Numeric	8	0		None	None	8	Right	Scale
5	пробег	Numeric	8	0		None	None	8	Right	Scale
6										
7										

Factor Analysis

Variables: марка

Variables: цена, т_сост, возраст, пробег

Selection Variable:

OK, Paste, Reset, Cancel, Help

Descriptives..., Extraction..., Rotation..., Scores..., Options...

Factor Analysis: Options

Missing Values: Exclude cases listwise, Exclude cases pairwise, Replace with mean

Coefficient Display Format: Sorted by size, Suppress absolute values less than: .10

Continue, Cancel, Help

Data View Variable View /

SPSS Processor is ready



Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Factor Analysis

Correlation Matrix^a

	цена	т_сост	возраст	пробег	
Correlation	цена	1,000	,489	-,447	-,461
	т_сост	,489	1,000	,101	-,536
	возраст	-,447	,101	1,000	-,314
	пробег	-,461	-,536	-,314	1,000
Sig. (1-tailed)	цена		,032	,047	,042
	т_сост	,032		,361	,020
	возраст	,047	,361		,127
	пробег	,042	,020	,127	

a. Determinant = ,220

SPSS Processor is ready



Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,427
Bartlett's Test of Sphericity	Approx. Chi-Square	17,922
	df	6
	Sig.	,006

Communalities

	Initial	Extraction
цена	1,000	,898
т_сост	1,000	,711
возраст	1,000	,946
пробег	1,000	,809

Extraction Method: Principal Component Analysis.

SPSS Processor is ready



Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

← → + - [Icons]

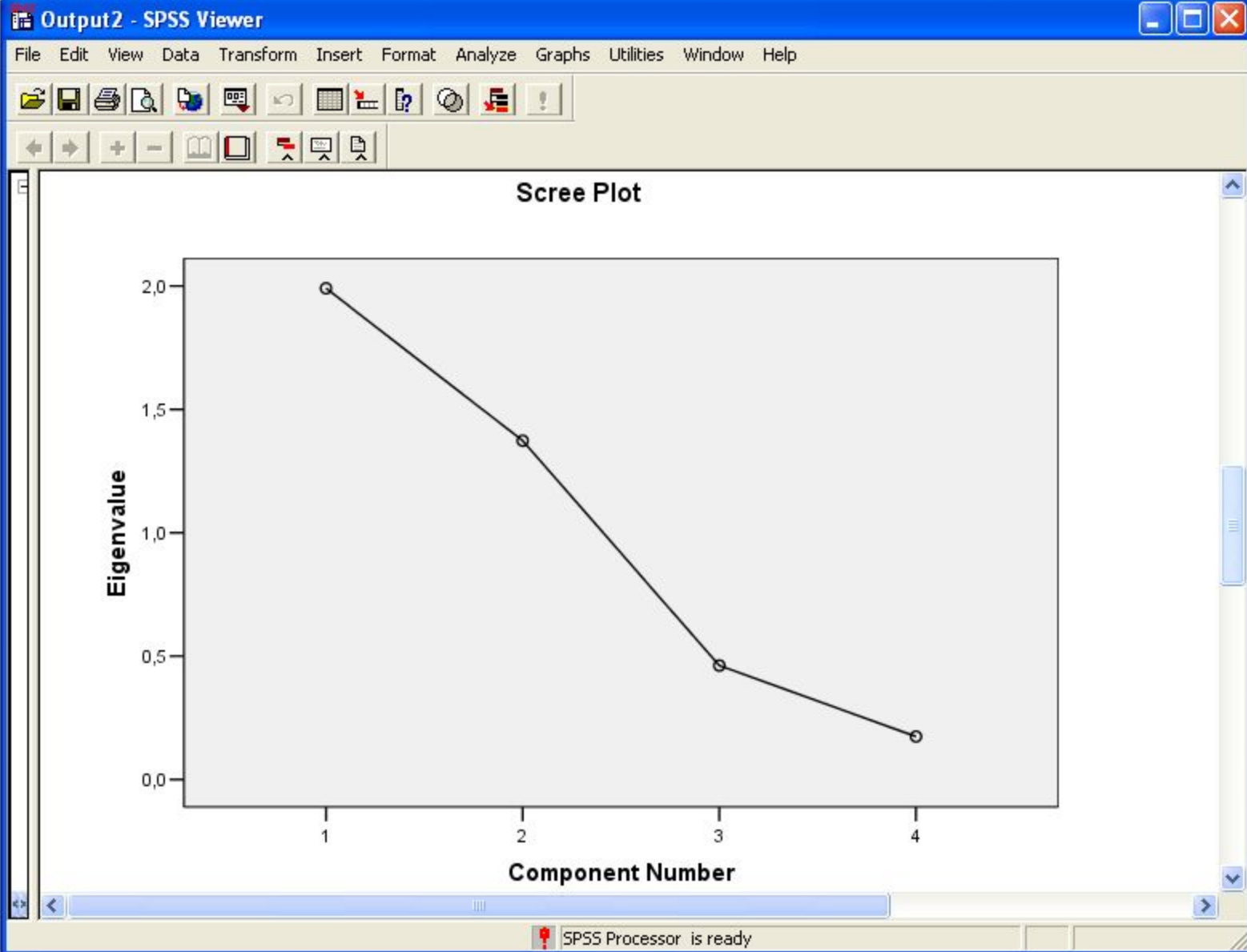
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1,991	49,772	49,772	1,991	49,772	49,772	1,962	49,053	49,053
2	1,373	34,329	84,101	1,373	34,329	84,101	1,402	35,048	84,101
3	,462	11,544	95,645						
4	,174	4,355	100,000						

Extraction Method: Principal Component Analysis.

SPSS Processor is ready

H: 172 , W: 346 pt





Output2 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help



Component Matrix^a

	Component	
	1	2
т_сост	,832	,138
пробег	-,815	-,381
цена	,797	-,513
возраст	-,017	,972

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Rotated Component Matrix^a

	Component	
	1	2
пробег	-,878	-,197
т_сост	,842	-,045
возраст	,193	,953
цена	,667	-,673

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.



Взаимосвязи

Классификация

Снижение размерности

Component Transformation Matrix

Component 1	0.978	-0.218
Component 2	0.218	0.978

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Component Score Coefficient Matrix

переменная	Component	
	1	2
Т-сост	0.978	-0.218
возраст	0.218	0.978
продол	-0.218	0.978

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Component Score Covariance Matrix

Component 1	1.000	0.000
Component 2	0.000	1.000

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.



Простые методы факторного анализа:

- **однофакторная модель Ч. Спирмена:** выделяется один латентный и один характерный факторы. Для возможно существующих других латентных факторов делается предположение об их незначительности
- **бифакторная модель Г.Хользингера:** допускается влияние на вариацию элементарных признаков не одного, а нескольких латентных факторов (обычно двух) и одного характерного фактора



□ **центроидный метод Л. Тэрстоуна:** корреляция между переменными рассматривается как пучок векторов, а латентный фактор геометрически представляется как уравновешивающий вектор, проходящий через центр этого пучка



Аппроксимирующие методы факторного анализа:

- главных факторов
- групповой
- минимальных остатков
- канонический
- распознавания образов
- максимального правдоподобия
- ...

Основные стадии факторного анализа:

- **Вычисление корреляционной матрицы для всех переменных, участвующих в анализе;** если переменная слабо коррелирована с остальными, следует подумать о ее исключении при следующем запуске программы (учитывая при этом ее общность и нагрузки).
- **Извлечение факторов;** оцениваются нагрузки факторов; выбирается либо метод главных компонент, либо один из методов факторного анализа.
- **Вращение факторов для создания упрощенной структуры;** вращение увеличивает, либо уменьшает нагрузки на каждый фактор; просмотр результатов позволяет уменьшить первоначальное число факторов.
- **Интерпретация факторов.**