Лекция 9б. Факторен анализ с SPSS

Доц. д-р С. Гочева-Илиева
ПУ"П.Хилендарски", , snow@uni-plovdiv.bg

1. Въведение във ФА

ФА е статистическа техника за преобразуване на множество от m корелиращи променливи в по-малък брой множество от k некорелиращи променливи (фактори), които описват възможно по-голяма част от изменчивостта на началните данни.

ФА може да се прилага за класификации, за генериране на хипотези относно причинно-следствени връзки или да подготви данните за следващи статистически обработки.

ФА се използва широко в социологията, икономиката, психологията, физиката, инженерните науки и др. области.

Идеята на ФА е представяне на оригиналните наблюдавани променливи като линейни комбинации от факторни (изкуствени променливи) + някаква грешка, във вида

$$\mathbf{X} = \mathbf{FL} + \mathbf{E} \qquad (1)$$

където $\mathbf{X} \in \mathbb{R}^{n \times m}$ е матрица на изходните взаимно корелиращи данни, $\mathbf{F} \in \mathbb{R}^{n \times k}$ е матрицата на факторните стойности (Factor scores), изчислени за всяко наблюдение, $\mathbf{L} \in \mathbb{R}^{k \times m}$ е матрица на факторните тегла (factor loadings), $\mathbf{E} \in \mathbb{R}^{n \times m}$ е матрица на грешките. В (1) броят на факторите $\boldsymbol{k}$ се определя от изследователя.

Като цяло. ФА се базира на редуцираната корелационна матрица на изходните данни. След получаване на матрицата **F** (извличане на факторните променливи – начално решение), се извършва т.н. въртене на факторите. Според метода на въртене полученото завъртяно решение

представлява графично ортогонални или наклонени направления в общото количество от данни.

Ще отбележим, че основната цел на ФА е да може получените фактори да бъдат подходящо интерпретирани и да съответстват на смисъла на изследваните данни, като ги групират по подходящ начин. Математически ФА не е строго формализирана процедура и резултатите от него са приложими именно според това дали резултатите имат реална интерпретация.

1. Общи изисквания за ФА

- Данните трябва да имат случаен характер. Ако се налага, те се подлагат на процедура за рандомизация чрез случайни извадки.
- Препоръчва се да разполагаме поне с извадка с размерност n=50 и повече.
- Променливите трябва да са количествени или интервални по тип. Категорийни данни (напр. пол или религиозна принадлежност), не са подходящи за ФА.
- Желателно е данните да имат нормално разпределение.
- Наблюденията трябва да са независими.
- ФА изисква да се проверят условия за адекватност, напр. KMO, Bartlett's test и др.

2. SPSS статистики за ФА

ФА изисква да се изчисляват:

- Корелационната матрица на променливите и техните нива на значимост (Significance levels, Sig.)
- Детерминантата на корелационната матрица
- Евентуално: обратна и репродуцирана корелационна матрица, огледална матрица (inverse, reproduced correlation matrix, anti-image)
- KMO (Kaiser-Meyer-Olkin) тест за адекватност и Bartlett's test of sphericity
- Незавъртяно начално решение (unrotated solution, communalities, and eigenvalues – собствени стойности)
- Завъртяно решение с факторните тегла (rotated solution with factor loadings)
- Факторни променливи (factor scores)

3. Алгоритъм за провеждане на основните процедури на ФА

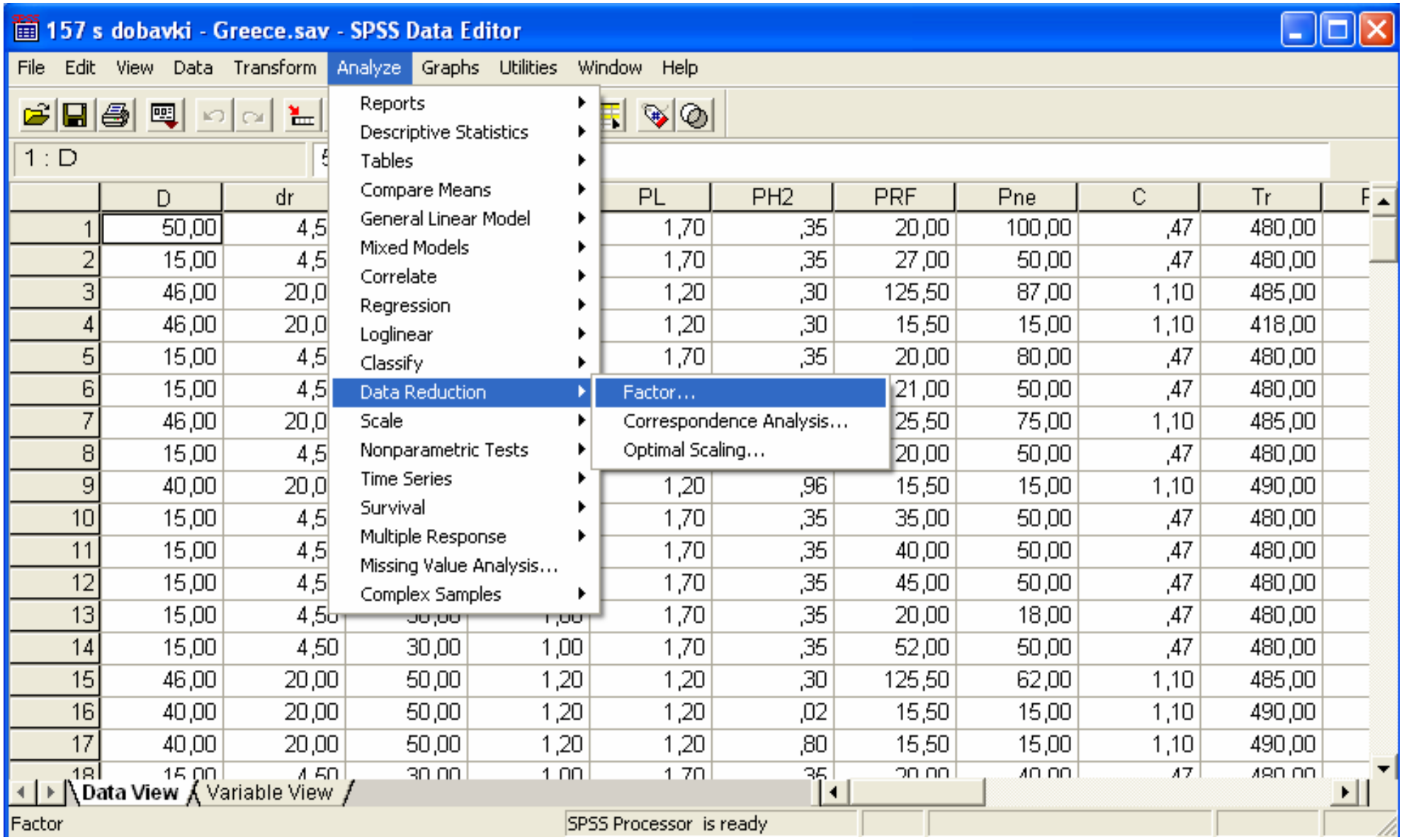
- Изчисляване на корелационната матрица
- Проверка на тестовете за адекватност на ФА
- Прилагане на Метод на главните елементи (PCA - Principal Component Analysis) (PCA) и (или) друга техника за изчисляване на натрупаните вариации и обща вариация на данните по фактори
- Избор на броя фактори
- Извличане на факторите
- Въртене на факторите и получаване на факторните тегла
- Изчисляване и съхраняване на факторните променливи за следващи анализи

4. Провеждане на ФА с SPSS

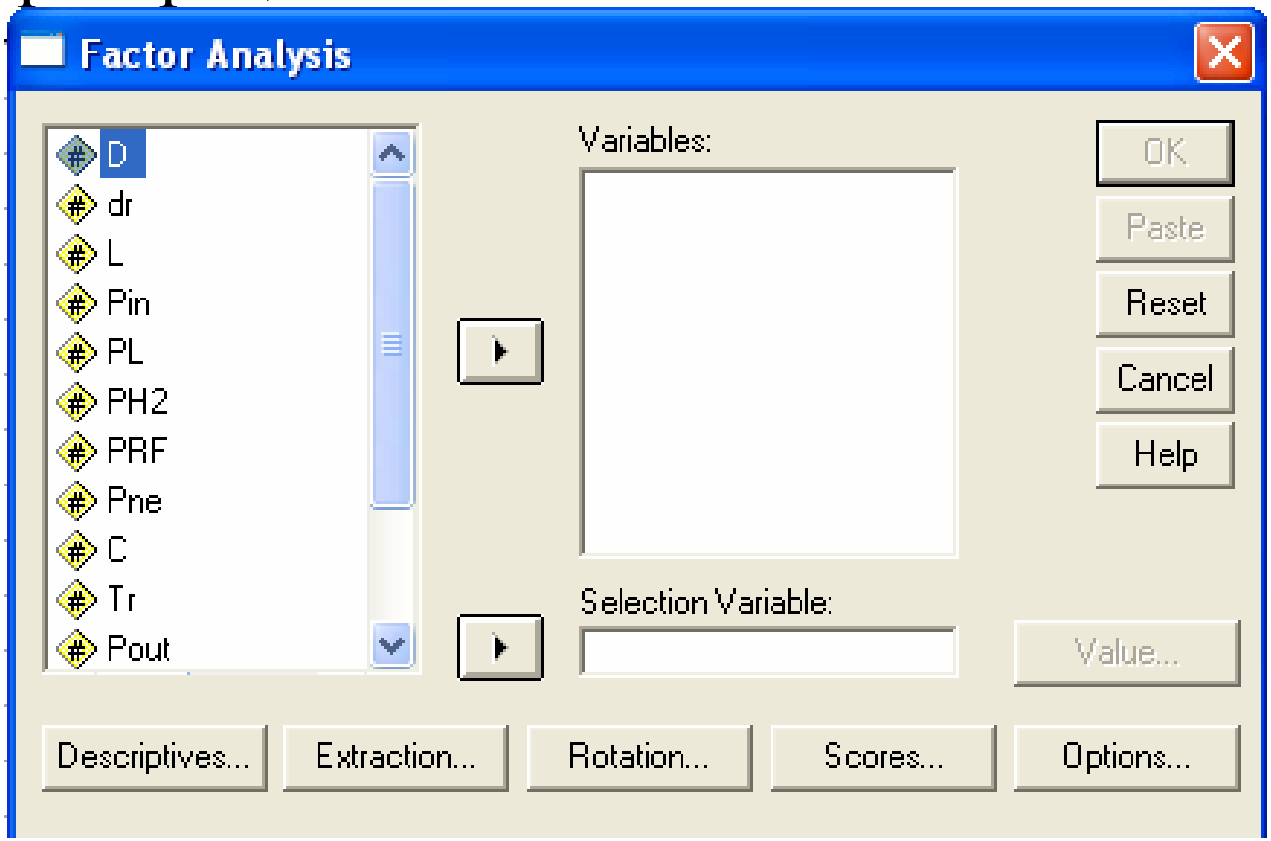
4.1) Въвеждаме данните. Те автоматично се преобразуват в стандартизирани (обезмерени) z-променливи (със средна =0 и стандартно отклонение = 1).

4.2) От главното меню на SPSS избираме:

Analyze/Data reduction/Factor

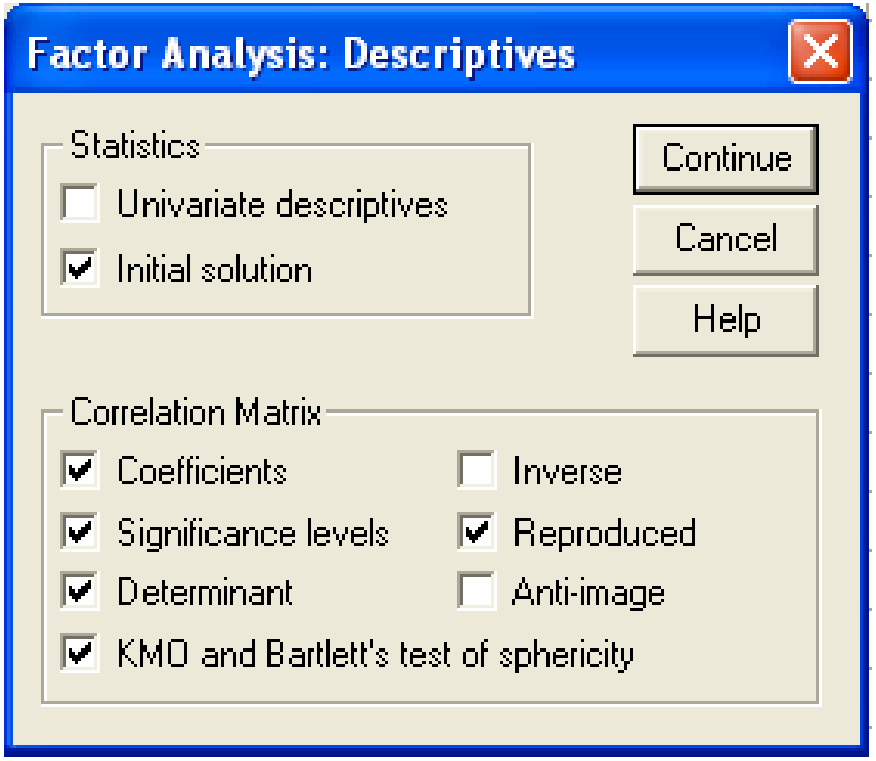


Появява се прозорецът на ФА:



4.3) Преместваме желаните променливи отляво отдясно в областта Variables. В началния стадий може да включим всички променливи, вкл. зависимите, за да проверим силата на зависимост между тях. При следващи анализи зависимите променливи и т.н. единични променливи се пропускат.

4.4) Избираме опции от менюто Descriptives... + Continue.



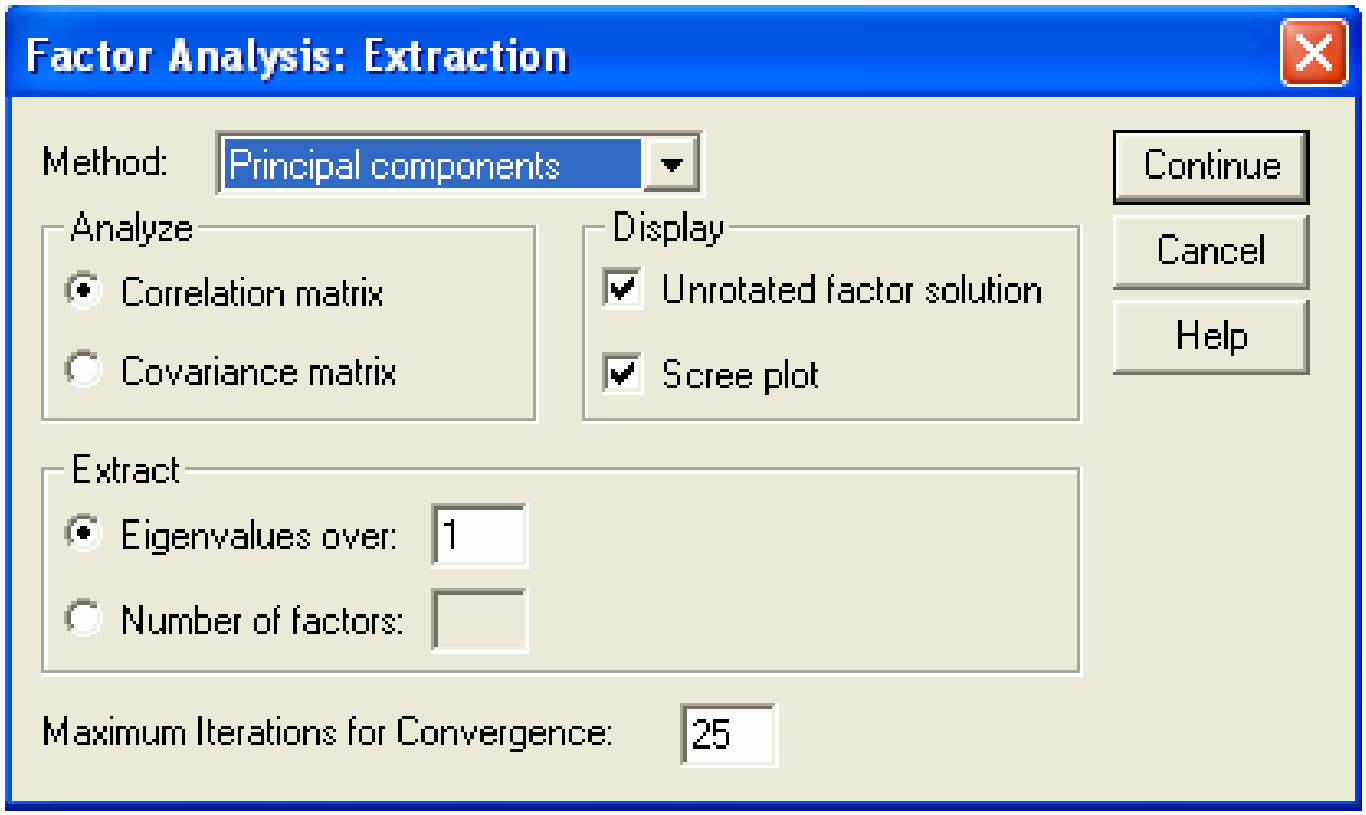
Целта ни е да определим кои променливи корелират силно една с друга и техните корелации са статистически значими (т.е. корелационните коефициенти да са поне >0,3 и нагоре, със съответен коефициент на значимост Sig. <0,05). Ако има

променлива, която корелира силно със зависимите променливи, но не корелира с останалите независими променливи, тя се нарича единична променлива. Всички единични променливи се отстраняват по-нататък от ФА и се добавят при друг тип анализи (регресионен, диспресионен и др.). I

ФА е приемлив, когато едновременно:

- KMO тестът за адекватност на извадката е >0,5
- Bartlett's test за сферичност на облака данни има Sig. <0,05. Изчисляване на корелационната матрица

4.5) От прозореца Factor Analysis отваряме Extraction... **.**
Избираме метода на извличане на факторите и кои резултати да покажем. Натискаме Continue.

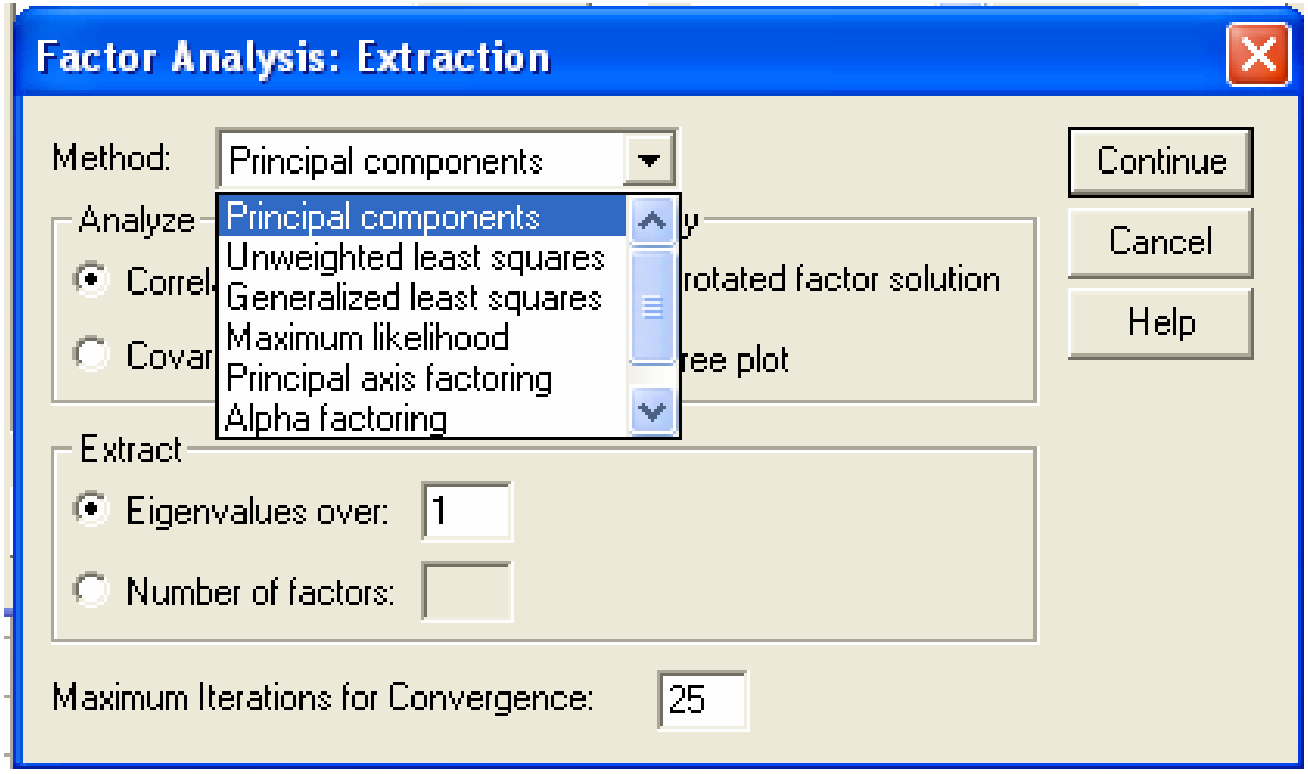


SPSS предлага 7 различни метода за извличане на фактори:

- Principal component analysis (по подразбиране)- метод на главните елементи
- Unweighted least squares (обикновен метод на най-малките квадрати)
- Generalized least squares (обобщен МНМК)
- Maximum likelihood
- Principal axis factoring
- Alpha factoring
- Image factoring

Стандартно се работи с Principal components. Това е метод, който преобразува всичките *n* променливи в *n* нови променливи, според тяхното относително тегло в общата вариация на облака данни. Следователно, сумата от техните вариации по този метод е винаги равна на 1.

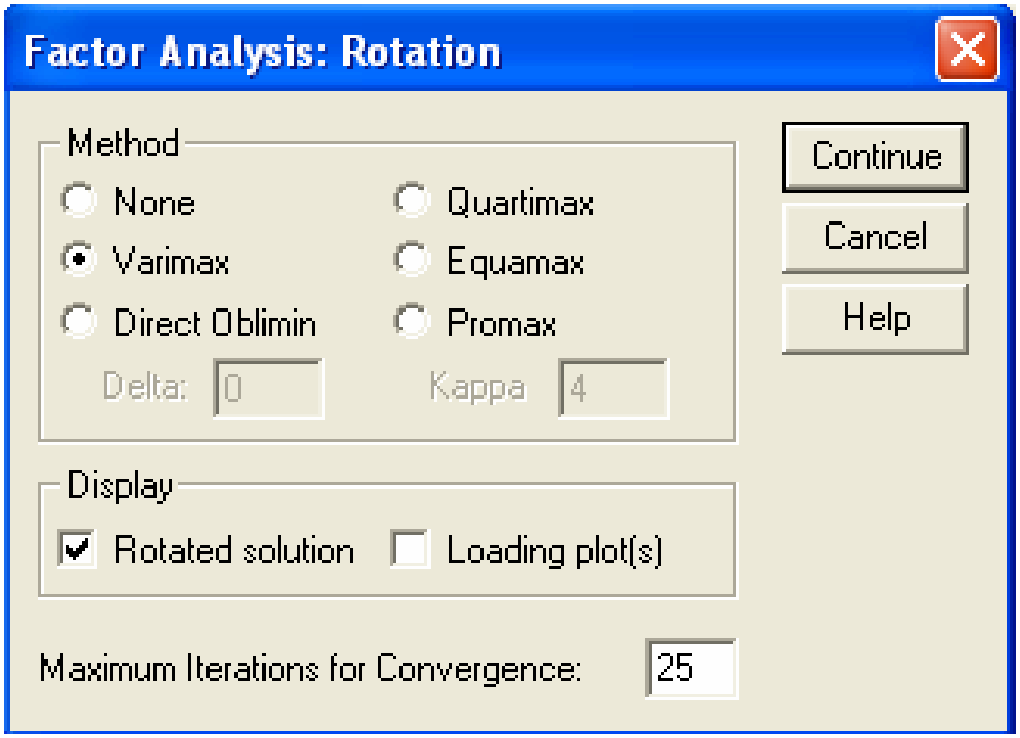
Методът изчислява абсолютните стойности на т.н. собствени стойности на корелационната матрица. За ФА обикновено се избират тези фактори, които имат с.ст. >= 1, но ако моделът не е достатъчно добър, се избират и с по-малки с.ст.



4.6) Метод на въртене – от бутона 

и Continue.

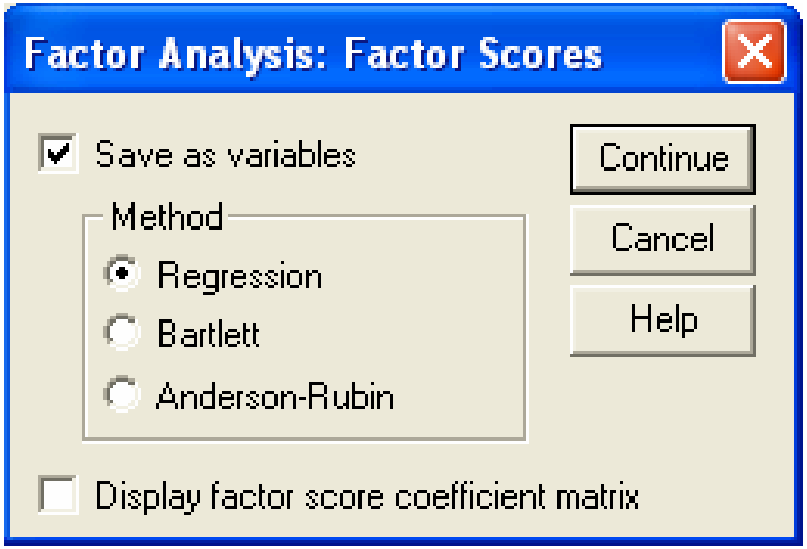
В SPSS има 6 метода на въртене. Най- стандартният е Varimax метод. Математически въртенето е получаване на нов базис с взаимноортогонални или наклонени променливи.



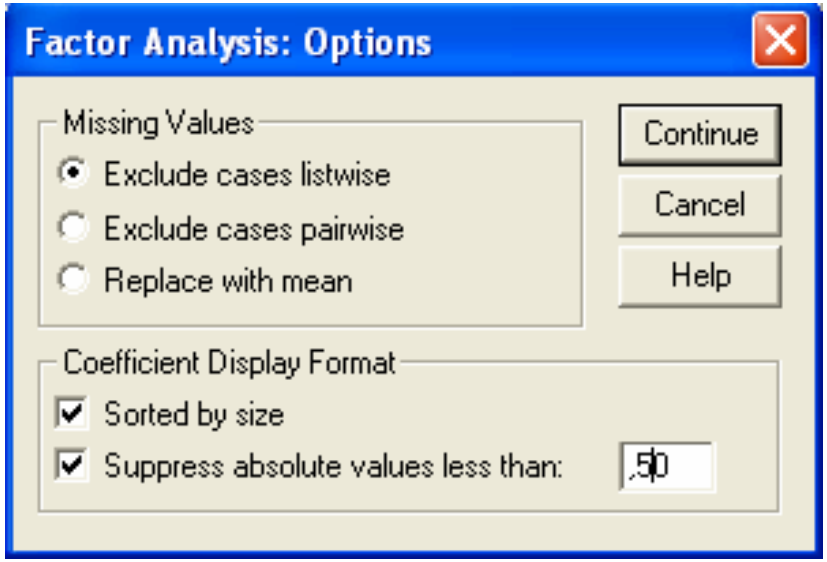
4.7) Изчисляване и запомняне на факторните стойности (променливи) - 

.

Тази процедура се използва, когато са успешни предишните дотук.



4.8) В прозореца Factor analysis/Options можем да изберем метод за обработване на липсващи данни и формата на компонентите на ФА.



4.9) Накрая в основния прозорец Factor analysis избираме OK, за да стартираме анализа

Резултатите се появяват в отделен прозорец и могат да се съхранят при желание като отделен файл. Факторните променливи се добавят към изходните данни.

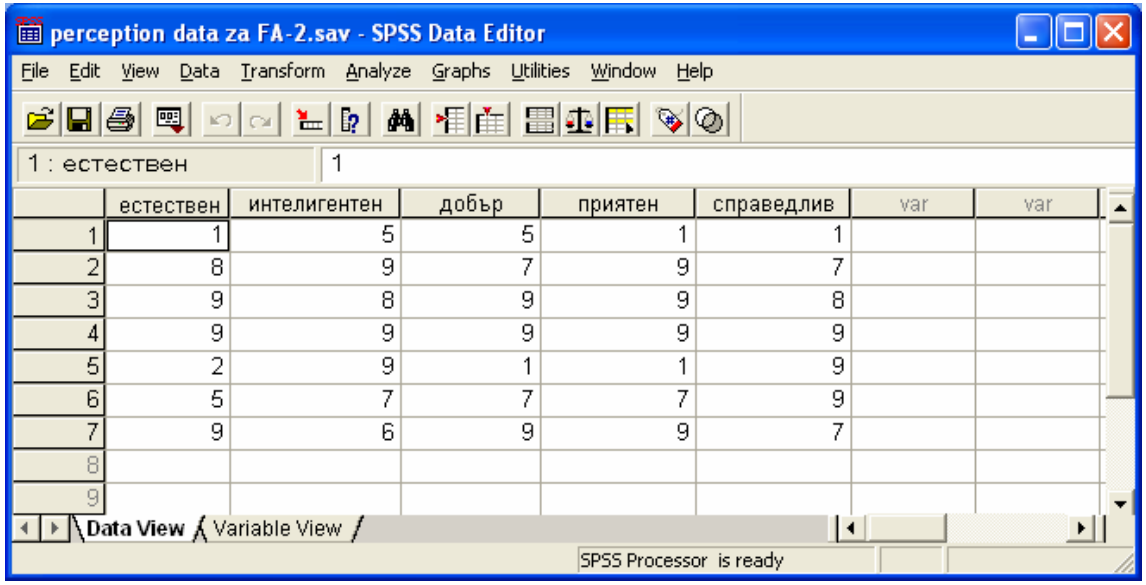
5. Пример 1 за ФА

Данни: Момиче на 12 години отговаря по 9 точкова възходяща скала (от 1 до 9) за възприятията си от 7 свои познати. Класирането е по следните 5 описания: “естествен”, “интелигентен”, “добър”, “приятен” и “справедлив”. Да се направи групиране на данните с ФА.

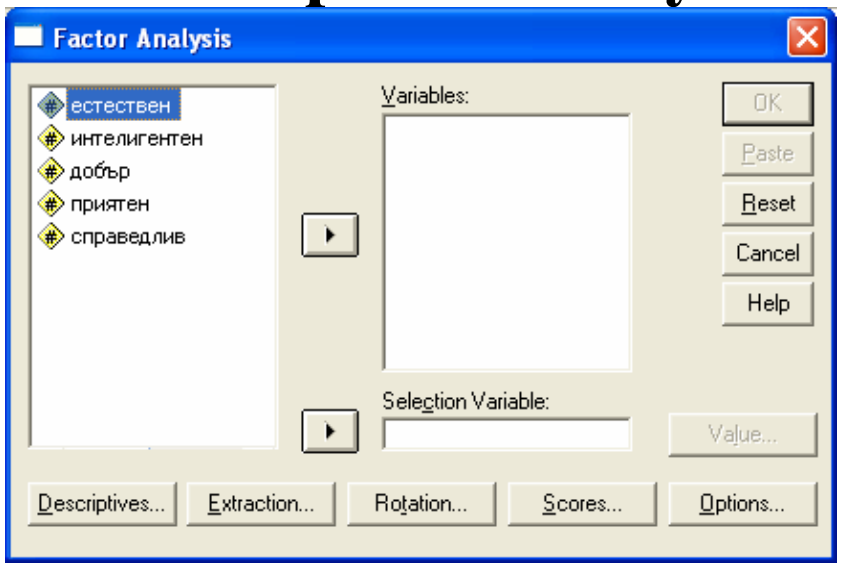
Таблица:

	естествен	**интелигентен**	**добър**	**приятен**	**справедлив**
съученичка 1	**1**	**5**	**5**	**1**	**1**
сестра	**8**	**9**	**7**	**9**	**7**
съученичка 2	**9**	**8**	**9**	**9**	**8**
баща	**9**	**9**	**9**	**9**	**9**
учител	**2**	**9**	**1**	**1**	**9**
съученик	**5**	**7**	**7**	**7**	**9**
съученичка 3	**9**	**6**	**9**	**9**	**7**

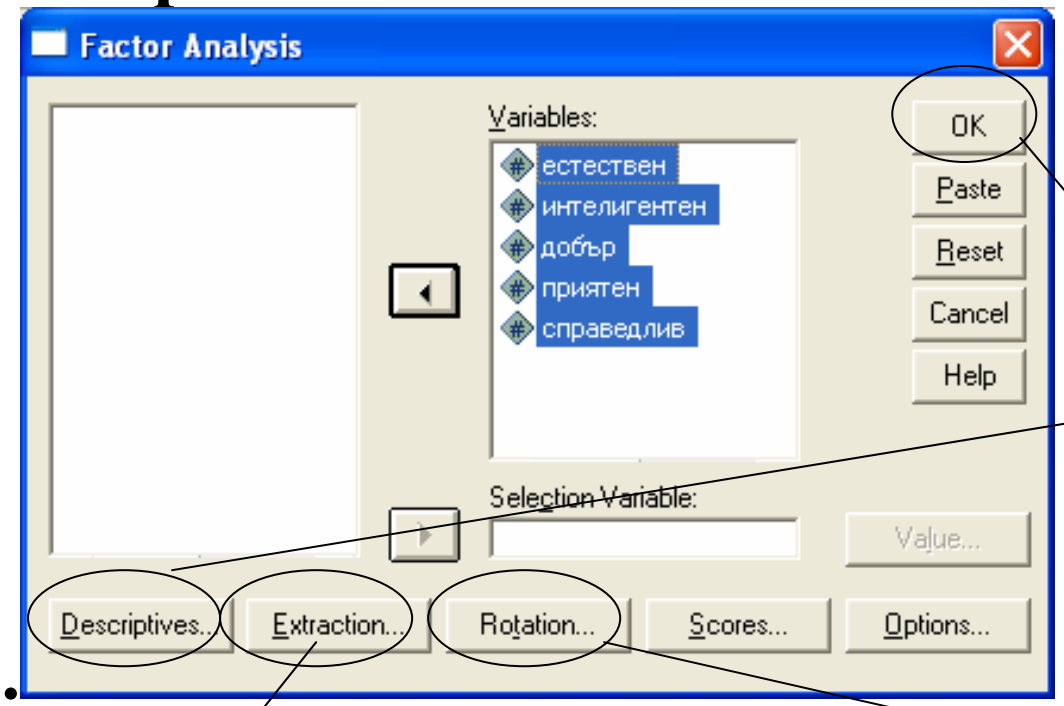
5.1. Въвеждаме данните:



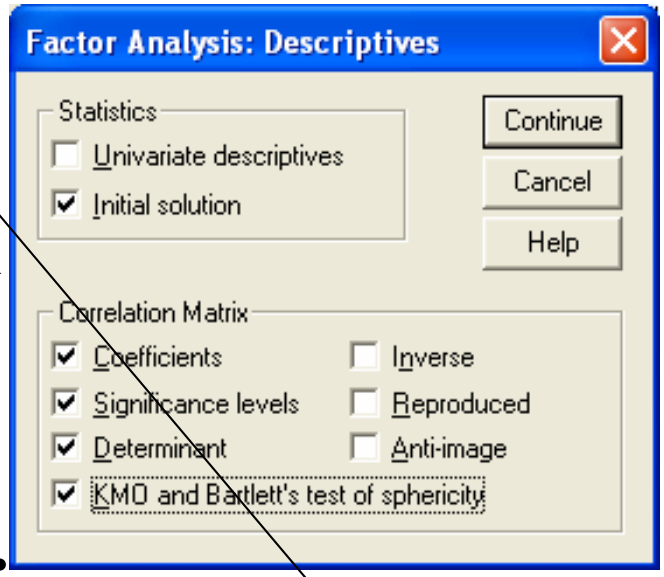
5.2. Избираме: Analyze/Data reduction/Factor



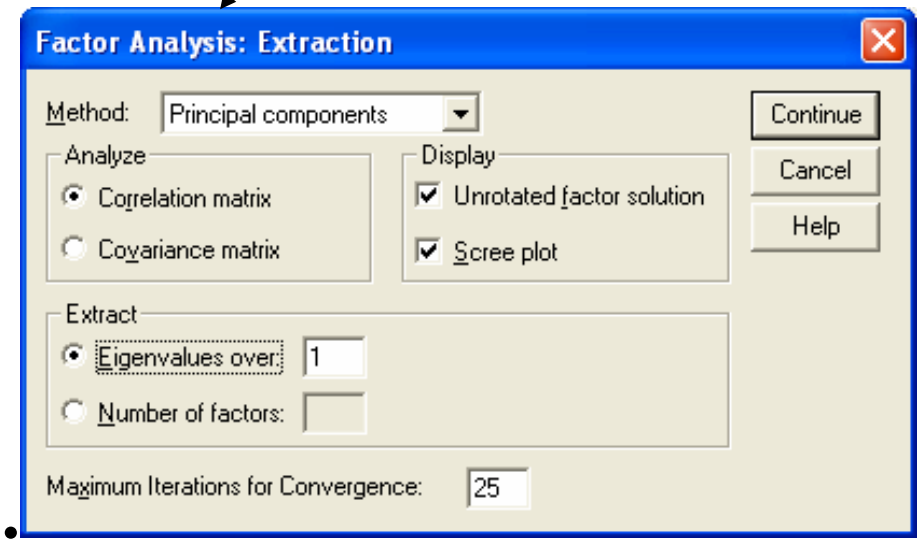
5.3. Пренасяме в полето Variables и отваряме подменютата:



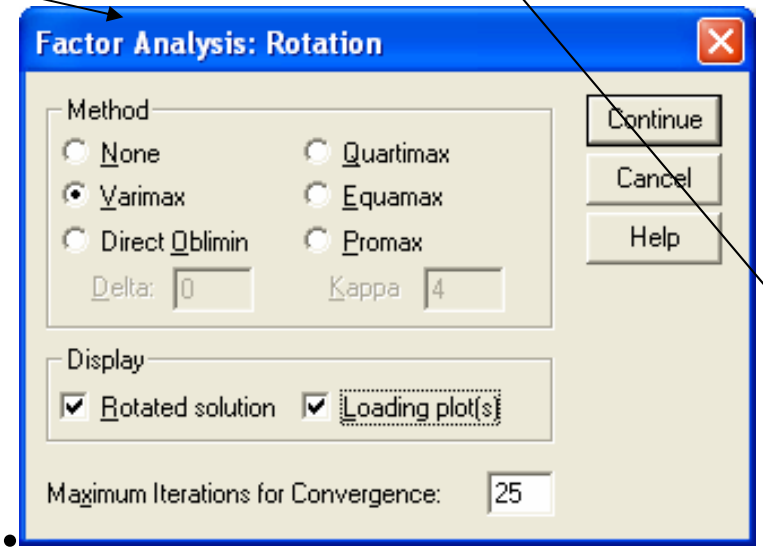
1.



2.



3.



4.

5.

5.4. Резултати:
а) корелационна матрица

Correlation Matrix[a]

		естествен	интелигентен	добър	приятен	справедлив
Correlation	естествен	1,000	,338	,854	,969	,484
	интелигентен	,338	1,000	-,101	,280	,737
	добър	,854	-,101	1,000	,886	,125
	приятен	,969	,280	,886	1,000	,472
	справедлив	,484	,737	,125	,472	1,000
Sig. (1-tailed)	естествен		,229	,007	,000	,136
	интелигентен	,229		,415	,271	,029
	добър	,007	,415		,004	,394
	приятен	,000	,271	,004		,142
	справедлив	,136	,029	,394	,142	

a. Determinant = ,001

Матрицата е симетрична. Детерминантата е 0.001, не е 0 и формално ФА може да се проведе. Гледаме само корелационните коефициенти > 0,5. Съответните им нива на значимост от долната половина на таблицата в случая имат нива на значимост Sig. <0,05. Това показва, че тези корелационни зависимости са статистически значими и трябва да участват в анализа. Останалите са незначими. В частност за тази извадка най-голям е корелационният коефициент между "приятен" и "естествен" (0.969) и той е значим.

б) тестове за адекватност на ФА:

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,689
Bartlett's Test of Sphericity	Approx. Chi-Square	23,320
	df	10
	Sig.	,010

Тук KMO=0,689 >0,5, следователно данните са подходящи за ФА и моделът е адекватен. Бартлет тестът има ниво на значимост Sig.=0.010<0.05, което означава, че облакът от данни е сферичен.

в) компоненти от МГЕ (PCA)

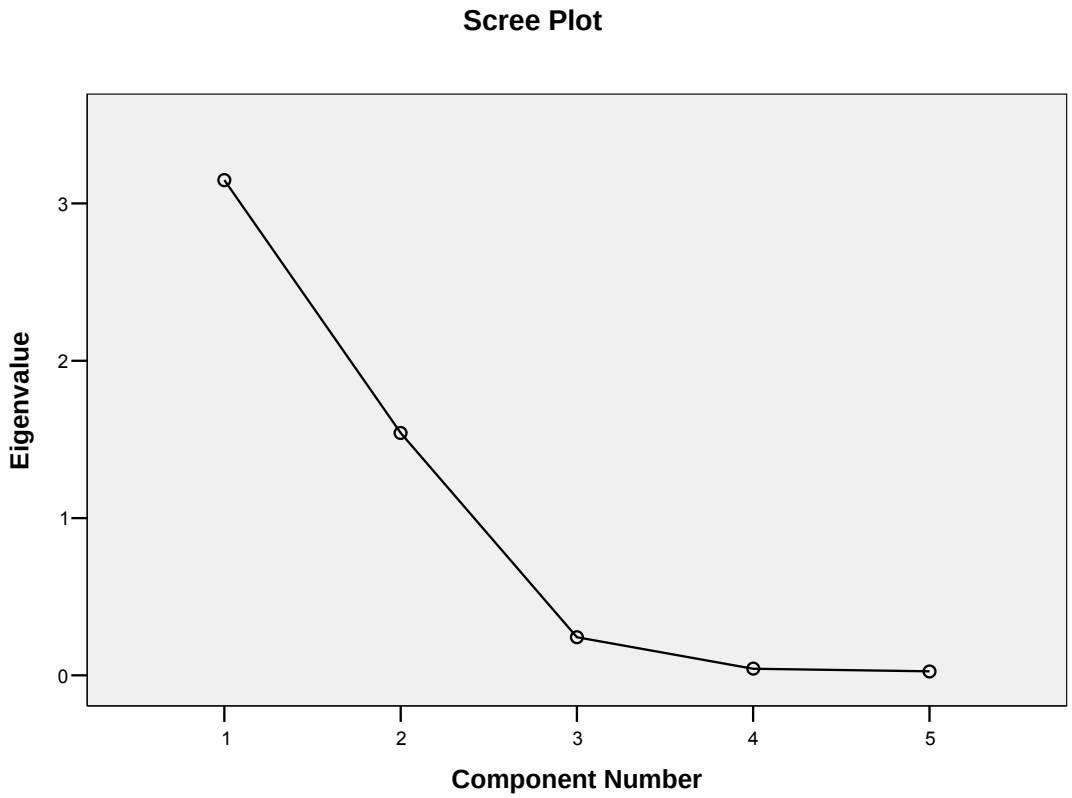
Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,148	62,966	62,966	3,148	62,966	62,966
2	1,541	30,814	93,780	1,541	30,814	93,780
3	,243	4,860	98,639			
4	,043	,851	99,491			
5	,025	,509	100,000			

Extraction Method: Principal Component Analysis.

Два фактора ще обяснят общо 93,780% от цялата извадка. Третата компонента ще осигури до 98,639% общо. Затова в първи вариант избираме 2 фактора.

г) графика на с.ст.



Тази графика нагледно показва колко фактори е добре да вземем. Спира се там, където вече става полегата, т.е. два или три фактора стигат.

д) незавъртяно, начално решение за 2 фактора

Component Matrix[a]

	Component	
	1	2
естествен	,972	-,155
интелигентен	,461	,827
добър	,803	-,577
приятен	,970	-,204
справедлив	,637	,677

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Няма ясно разделяне на променливите по фактори (компоненти) . Напр. променливата “справедлив” им тегла 0.637 в първи фактор и 0.677 във втори. Така тя не може ясно да се групира в никой от двата фактора. Продължаваме анализа с въртене.

е) завъртяно решение с Варимакс метод

Rotated Component Matrix[a]

	Component	
	1	2
естествен	,931	,321
интелигентен	,017	,947
добър	,980	-,131
приятен	,952	,276
справедлив	,244	,897

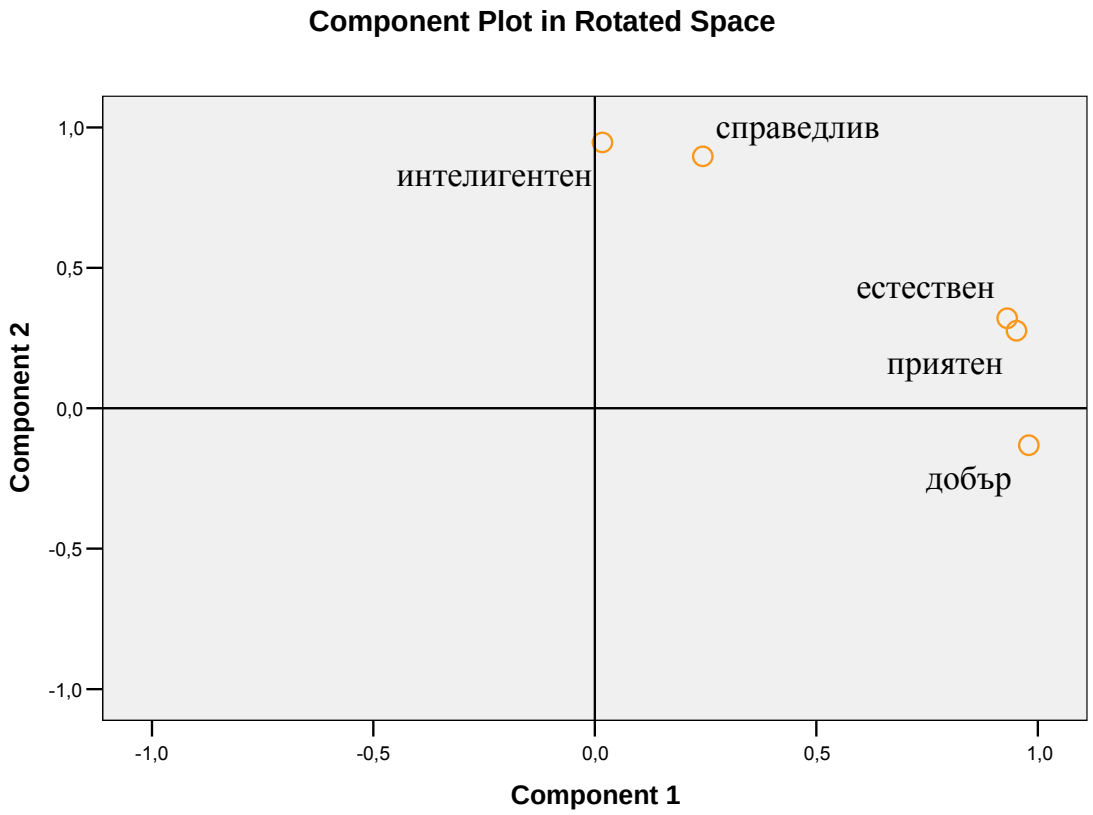
Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

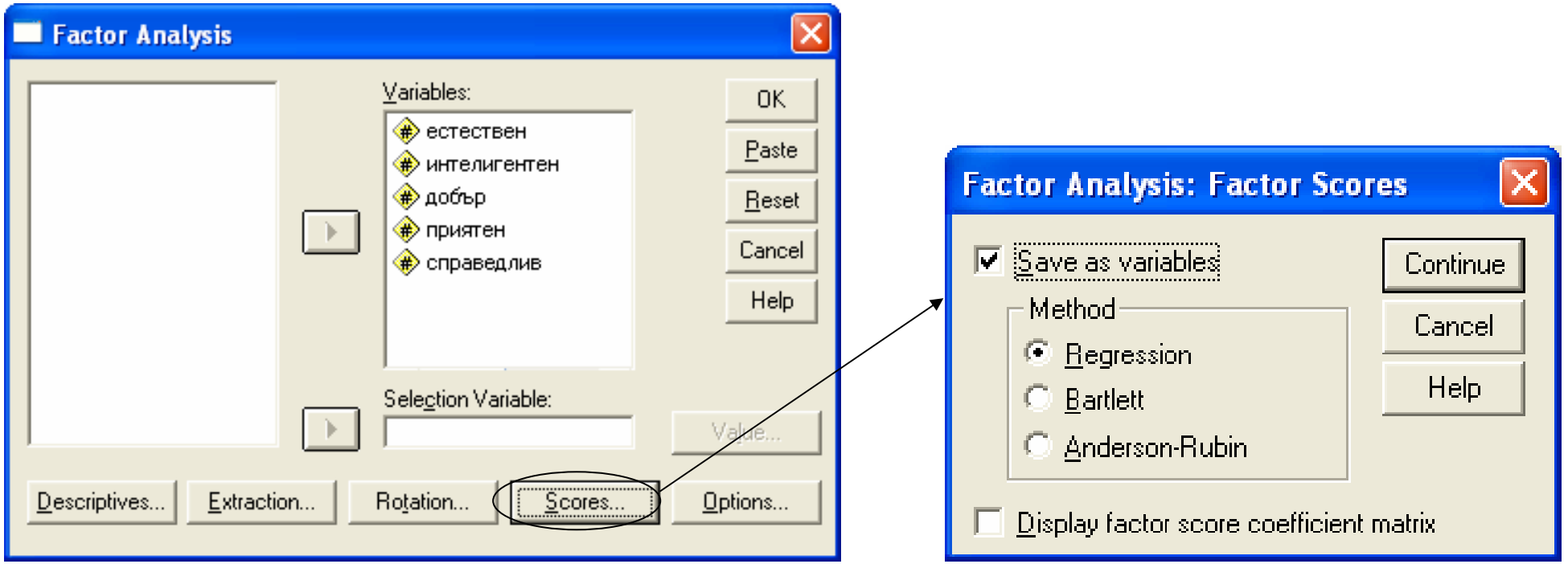
В първата компонента (фактор) ще се групират само променливите с тегла над 0,5 – това са 1, 3 и 4- ред – “естествен”, “добър” и “приятен”. Този първи фактор ще наречем “приятелски”.

Вторият фактор е съставен от групиране на 2 и 5 ред – “интелигентен” и “справедлив”. Можем да го наречем “възвишен”.

ж) графика на завъртяното решение в новото двумерно пространство от 2 фактора:

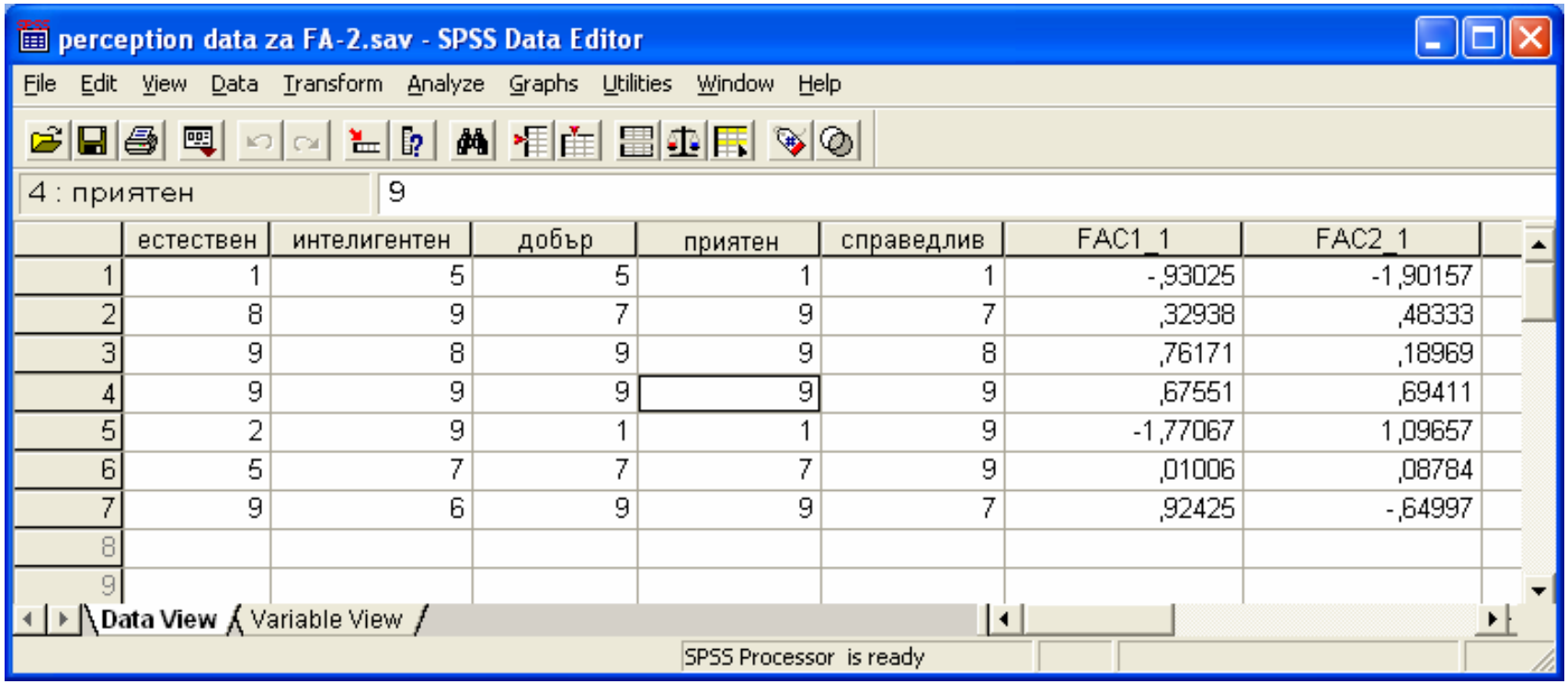


5.4. Запомняне на факторните променливи:
От менюто Factor Analysis избираме бутона Scores…



и потвърждаваме с Continue и OK.

Получаваме 2 нови променливи към първоначалните 5:


perception data za FA-2.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

4 : приятен 9

	естествен	интелигентен	добър	приятен	справедлив	FAC1_1	FAC2_1
1	1	5	5	1	1	-,93025	-1,90157
2	8	9	7	9	7	,32938	,48333
3	9	8	9	9	8	,76171	,18969
4	9	9	9	9	9	,67551	,69411
5	2	9	1	1	9	-1,77067	1,09657
6	5	7	7	7	9	,01006	,08784
7	9	6	9	9	7	,92425	-,64997
8							
9							

Data View Variable View

SPSS Processor is ready

Те описват нашите данни 93,78%, така че могат да се използват вместо петте променливи. При това са почти ортогонални и не корелират помежду си. Проверете последното твърдение.

Задача. Направете ФА за следните данни. Проверете адекватност на модела.

Данни: За 30 търговски марки японски вина “Сейшу” се изследва връзката между променливите:

y1 – вкус, y2 – мирис,

и променливите:

x1 – pH, ***x2*** – киселинност 1, ***x3*** – киселинност 3,

x4 – съдържание на оризова ракия, ***x5*** – горена захар,

x6 – общо количество захар, ***x7*** – алкохол, ***x8*** – формил-азот

Табл. 2. Сейшу измервания

y1	*y2*	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*
1,0	**,8**	**4,05**	**1,68**	**,85**	**3,0**	**3,97**	**5,00**	**16,90**	**122,0**
,1	**,2**	**3,81**	**1,39**	**,30**	**,6**	**3,62**	**4,52**	**15,80**	**62,0**
,5	**,0**	**4,20**	**1,63**	**,92**	**-2,3**	**3,48**	**4,46**	**15,80**	**139,0**
,7	**,7**	**4,35**	**1,43**	**,97**	**-1,6**	**3,45**	**3,98**	**15,40**	**150,0**
-,1	**-1,1**	**4,35**	**1,53**	**,87**	**-2,0**	**3,67**	**4,22**	**15,40**	**138,0**
,4	**,5**	**4,05**	**1,84**	**,95**	**-2,5**	**3,61**	**5,00**	**16,78**	**123,0**
,2	**-,3**	**4,20**	**1,61**	**1,09**	**-1,7**	**3,25**	**4,15**	**15,81**	**172,0**
,3	**-,1**	**4,32**	**1,43**	**,93**	**-5,0**	**4,16**	**5,45**	**16,78**	**144,0**

,7	,4	4,21	1,74	,95	-1,5	3,40	4,25	16,62	153,0
,5	-,1	4,17	1,72	,92	-1,2	3,62	4,31	16,70	121,0
-,1	,1	4,45	1,78	1,19	-2,0	3,09	3,92	16,50	176,0
,5	-,5	4,45	1,48	,86	-2,0	3,32	4,09	15,40	128,0
,5	,8	4,25	1,53	,83	-3,0	3,48	4,54	15,55	126,0
,6	,2	4,25	1,49	,86	2,0	3,13	3,45	15,60	128,0
,0	-,5	4,05	1,48	,30	,0	3,67	4,52	15,38	99,0
-,2	-,2	4,22	1,64	,90	-2,2	3,59	4,49	16,37	122,8
,0	-,2	4,10	1,55	,85	1,8	3,02	3,62	15,31	114,0
,2	,2	4,28	1,52	,75	-4,8	3,64	4,93	15,77	125,0
-,1	-,2	4,32	1,54	,83	-2,0	3,17	4,62	16,60	119,0
,6	,1	4,12	1,68	,84	-2,1	3,72	4,83	16,93	111,0
,8	,5	4,30	1,50	,92	-1,5	2,98	3,92	15,10	68,0
,5	,2	4,55	1,50	1,14	,9	2,60	3,45	15,70	197,0
,4	,7	4,15	1,62	,78	-7,0	4,11	5,55	15,50	106,0
,6	-,3	4,15	1,32	,31	,8	3,56	4,42	15,40	49,5
-,7	-,3	4,25	1,77	1,12	,5	2,84	4,15	16,65	164,0
-,2	,0	3,95	1,36	,25	1,0	3,67	4,52	15,98	29,5
,3	-,1	4,35	1,42	,96	-2,5	3,40	4,12	15,30	131,0
,1	,4	4,15	1,17	1,06	-4,5	3,89	5,00	16,79	168,2
,4	,5	4,16	1,61	,91	-2,1	3,93	4,35	15,70	118,0
-,6	-,3	3,85	1,32	,30	,7	3,61	4,29	15,71	48,0