

Aufgabe 1:

i) Asympt. Verteilung von $\log \hat{\lambda}_j$

- Delta Theorem

$$\sqrt{n} (X - E(X)) \overset{a}{\sim} N(0, \Sigma) \Rightarrow \sqrt{n} (g(X) - g(E(X))) \overset{a}{\sim} N(0, \nabla g(E(X))' \Sigma \nabla g(E(X)))$$

- Asympt. Verteilung

Wir definieren die Funktion $g_j(\lambda) = (0, \dots, 0, \overset{j\text{-te Stelle}}{\log \hat{\lambda}_j}, 0, \dots, 0)'$ mit

$$\nabla g_j(\lambda) = \left(\frac{\partial g_j(\lambda)}{\partial \lambda_1}, \dots, \frac{\partial g_j(\lambda)}{\partial \lambda_p} \right)' = (0, \dots, 0, \frac{1}{\lambda_j}, 0, \dots, 0)'$$

Wenden wir das Delta-Theorem auf die Funktion $g_j(\lambda)$ an erhalten wir mit Hilfe von Theorem 1a.1

$$\sqrt{n-1} (\log \hat{\lambda}_j - \log \lambda_j) = \sqrt{n-1} (g_j(\hat{\lambda}) - g_j(\lambda))$$

$$\overset{a}{\sim} N(0, \nabla g_j(\lambda)' \Sigma \nabla g_j(\lambda)) \sim N(0, 2)$$

Somit gilt

$$\sqrt{\frac{n-1}{2}} (\log \hat{\lambda}_j - \log \lambda_j) \sim N(0, 1)$$

ii) Konfidenzintervall

Wir bestimmen das Konfidenzintervall $[\log \hat{\lambda}_j, \bar{\lambda}_j]$ für ein Signifikanzniveau von α . Dazu betrachten wir

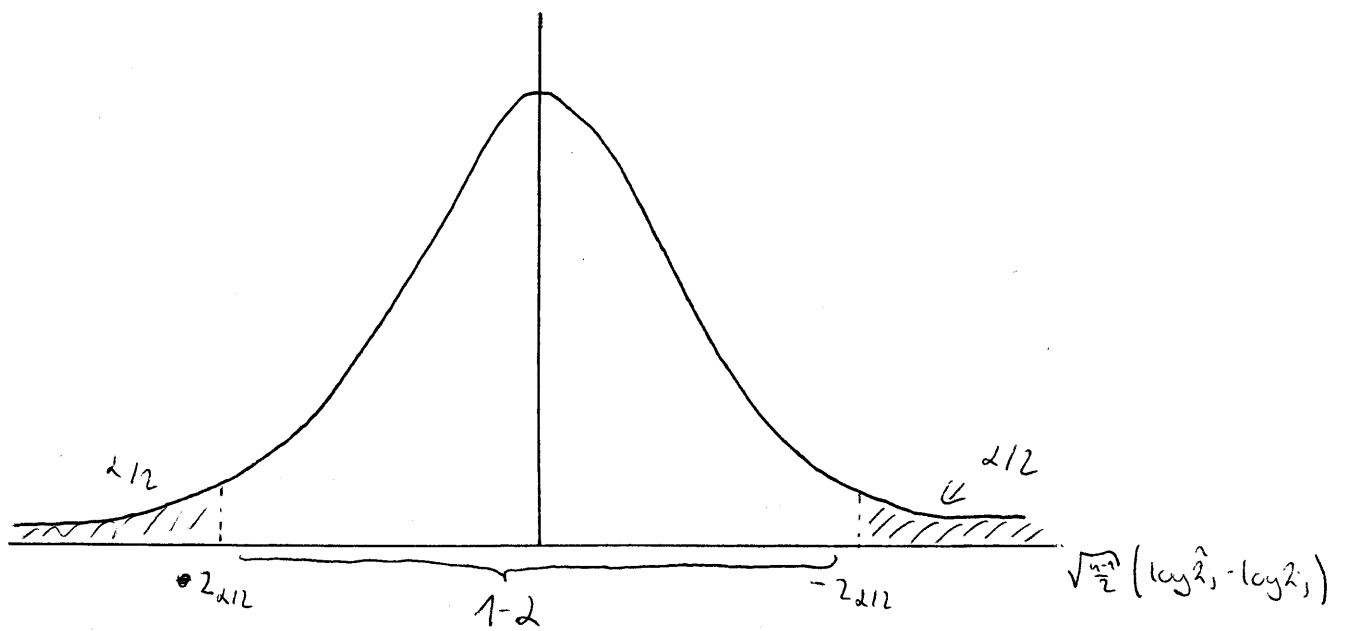
$$P \left(-z_{\alpha/2} \leq \sqrt{\frac{n-1}{2}} (\log \hat{\lambda}_j - \log \lambda_j) \leq z_{\alpha/2} \right) = 1 - \alpha$$

$$\Rightarrow P \left(\sqrt{\frac{2}{n-1}} z_{\alpha/2} - \log \hat{\lambda}_j \leq -\log \lambda_j \leq -z_{\alpha/2} \sqrt{\frac{2}{n-1}} - \log \hat{\lambda}_j \right) = 1 - \alpha$$

$$\Rightarrow P \left(z_{\alpha/2} \sqrt{\frac{2}{n-1}} + \log \hat{\lambda}_j \leq \log \lambda_j \leq \log \hat{\lambda}_j - \sqrt{\frac{2}{n-1}} z_{\alpha/2} \right) = 1 - \alpha$$

mit $z_{\alpha/2} = \Phi^{-1} \left(\frac{\alpha}{2} \right)$. Somit erhalten wir die untere und obere Schranke d. Konfidenzintervalls

$$\underline{\lambda}_j = \exp \left\{ z_{\alpha/2} \sqrt{\frac{2}{n-1}} + \log \hat{\lambda}_j \right\}, \quad \bar{\lambda}_j = \exp \left\{ -z_{\alpha/2} \sqrt{\frac{2}{n-1}} + \log \hat{\lambda}_j \right\}$$



Aufgabe 2:

1. a) Die Variable „length“ hat einen negativen und signifikanten Einfluss auf die erste Hauptkomponente. Jedoch erklärt die Variable „length“ nur etwa ~~20~~⁴⁰% der erklärten Varianz in der ersten Hauptkomponente. Somit ist die Bedeutung der Variable „length“ für die erste Hauptkomponente relativ ^{hoch} ~~gering~~.

$$\beta_{\text{length},1} \approx -0,44$$

$$\beta_{\text{length}} \approx 0,006$$

$$\lambda_1 \approx 3,00$$

$$\text{Var}(X_{\text{length}}) \approx (0,38)^2 \approx 0,14$$

$$S_{X_{\text{length}},Y_1} \approx \beta_{\text{length},1} \cdot \sqrt{\frac{\lambda_1}{\text{Var}(X_{\text{length}})}} \approx -0,20$$

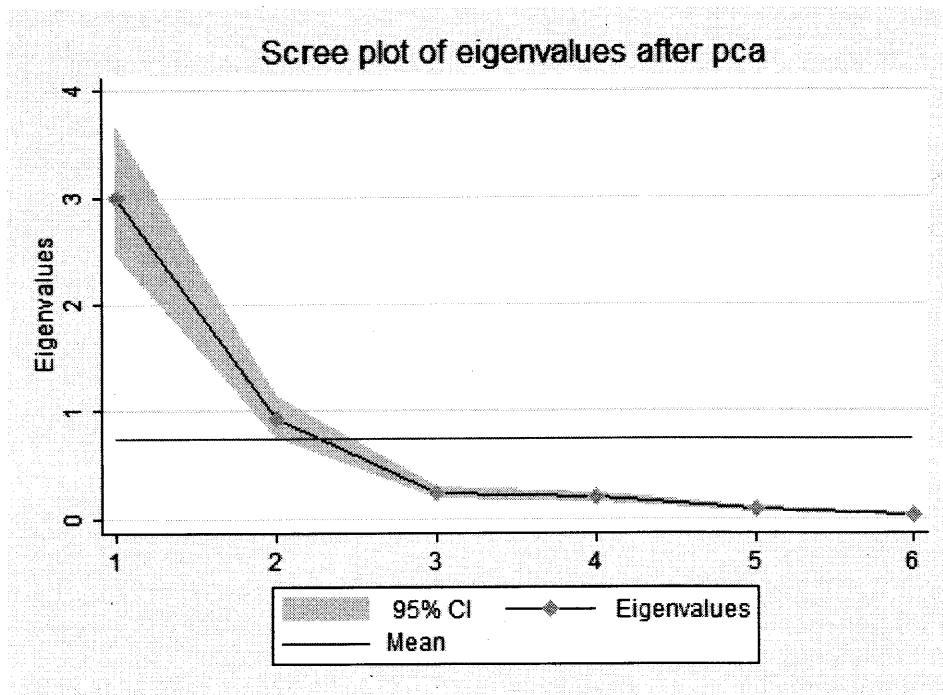
- b) Die ersten zwei Hauptkomponenten beschreiben überdurchschnittlich viel von der Varianz. Die restl. Hauptkomponenten beschreiben nur einen unterdurchschnittlichen Anteil an der Varianz. Somit sollte man auf jeden Fall die ersten zwei Hauptkomponenten beibehalten. ~~Erstmal~~
- c) Unter Nullhypothese $H_0: \text{Var}(Y_{p-2+1}) = \dots = \text{Var}(Y_p)$ gilt

$$U_n = \left(n - \frac{2p+1}{6} \right) \left(2 \ln \bar{\lambda}_n - \sum_{j=p-2+1}^p \lambda_j \right) \stackrel{a}{\sim} \chi^2_v$$

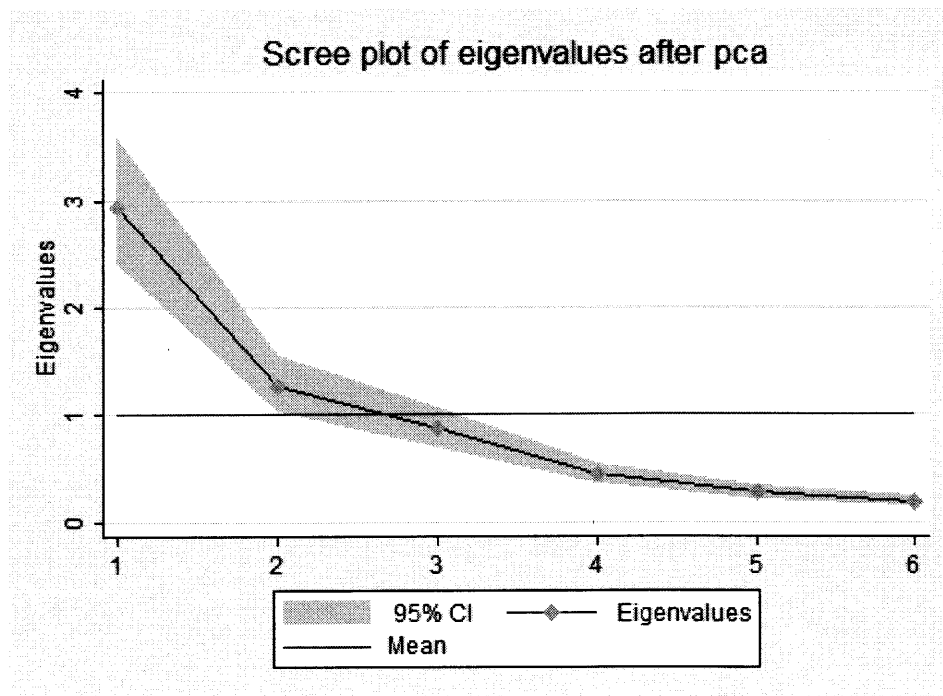
mit $\bar{\lambda}_n = \frac{1}{n} \sum_{j=p-2+1}^p \lambda_j$ und $v = \frac{1}{2} (2-1)(2+2)$. Die Ergebnisse dieses Tests zeigen, dass jede Hauptkomponente einen signifikanten Anteil an der Varianz erklärt. Somit sollten alle ~~Eigenwert~~-HK genutzt werden.

- 2a.) Der Unterschied von λ_1 gegenüber dem Fall „ohne standardisierte Daten“ ist relativ gering ($3 - 2,95$); jedoch ist der Unterschied zu den anderen HK geringer geworden.
- b.) Die Entscheidung für zwei Hauptkomponenten bleibt unverändert. Jedoch wird hier noch deutlicher, dass die anderen HK einen signifikanten Erklärungsgehalt an der Varianz haben. Somit sollte man seine Entscheidung mittels eines tests nochmals formal überprüfen (Dies haben wir getan). Dieser zeigt uns, dass man keine HK vernachlässigen sollte.
- c.) Der screeplot zeigt uns, dass die Streuung in den ersten zwei HK relativ groß ist. Man sieht auch, dass diese HK unkorreliert sind.

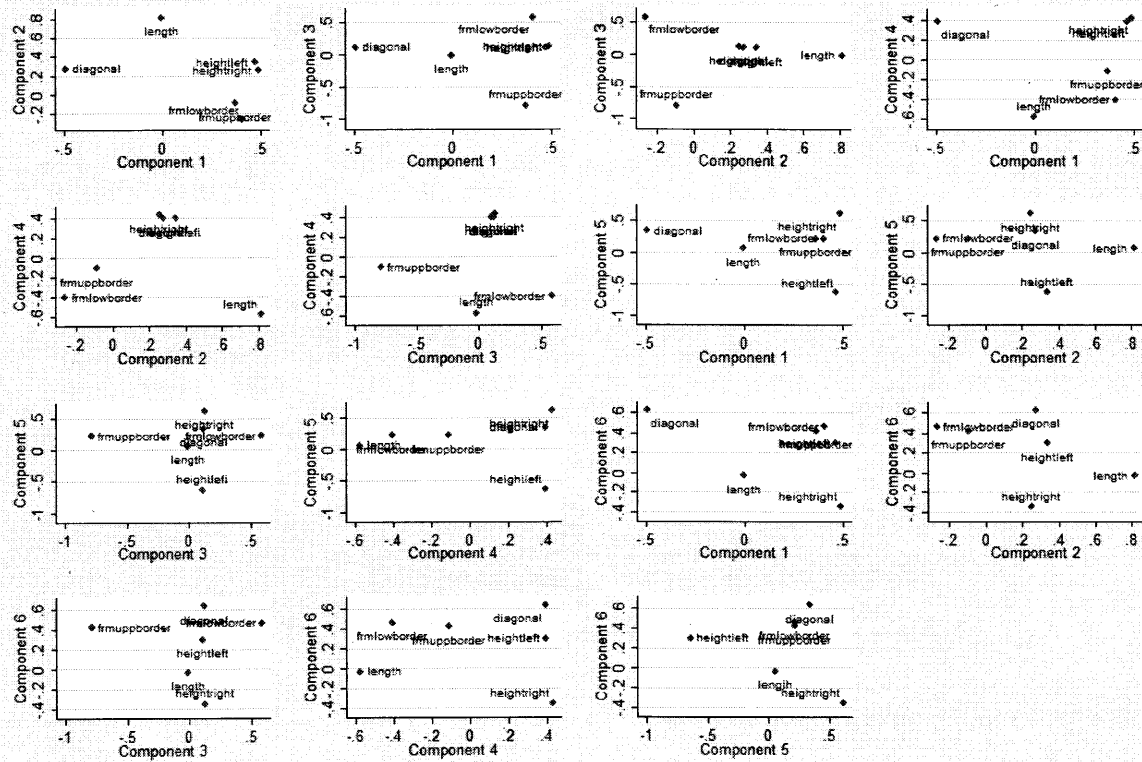
Hauptkomponentenanalyse mit Kovarianzmatrix:



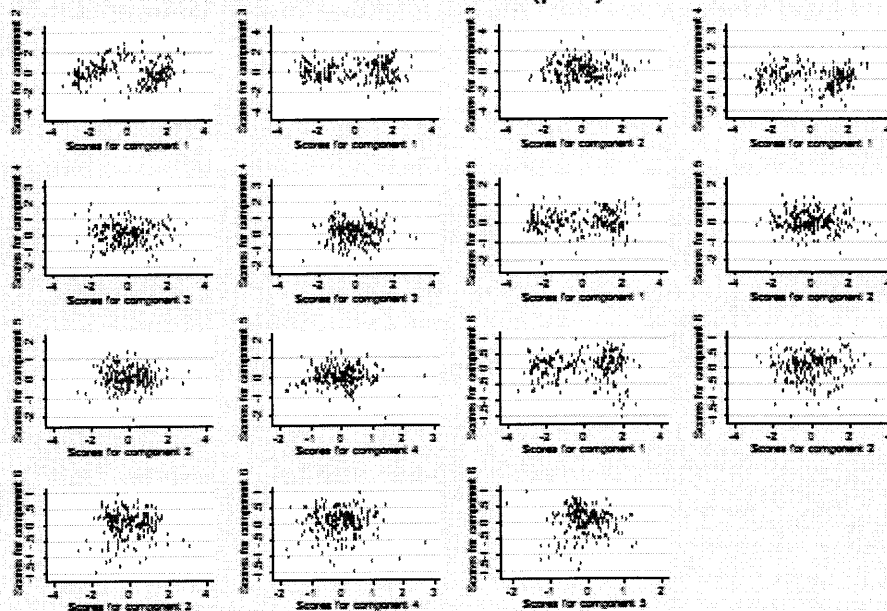
Hauptkomponentenanalyse mit Korrelationsmatrix:



Component loadings



Score variables (pca)



Bank.log

log: U:\Übung4\Bank.log
log type: text
opened on: 18 Jun 2011, 14:00:21

```

. /*****
. /*
. /*      */
. /*  Ü B U N G  4  -  H A U P T K O M P O N E N T E N A N A L Y S E      */
. /*
. /*      */
. /*****
.
. /*****
. /* 0. Preliminary */
. /*****

```

```

. * Daten einlesen
. use "U:\Übung4\Bank.dta", clear

```

```

. /*****
. /* 1. Hauptkomponentenanalyse */
. /*****

```

Verwende nicht
standardisierte Daten

Verwende geschätzte
Standardfehler basierend
auf Normalverteilung

```

. * 1.1 Hauptkomponentenanalyse mit Kovarianzmatrix
. pca length heightleft heightright frmlowborder frmuppborder diagonal, vce(normal)

```

COV

Principal components/covariance

Number of obs = 200
Number of comp. = 6
Trace = 4.494725 = $\sum \lambda_i$
Rho = 1.0000
SE(Rho) = 0.0000

SEs assume multivariate normality

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Eigenvalues					λ_1	λ_2
Comp1	3.000305	.2998663	10.01	0.000	2.412578	3.588032
Comp2	.9356206	.0930188	10.06	0.000	.7533071	1.117934
Comp3	.2434134	.0231389	10.52	0.000	.198062	.2887648
Comp4	.1946587	.0180351	10.79	0.000	.1593106	.2300068
Comp5	.0852121	.0083423	10.21	0.000	.0688616	.1015627
Comp6	.0355148	.0034805	10.20	0.000	.0286931	.0423365

Comp1		λ_1	λ_2
length	-.0437744	.0159486	-2.74
heightleft	.1121618	.0134047	8.37
heightright	.1391907	.0142387	9.78
frmlowborder	.7683049	.0328803	23.37
frmuppborder	.2017662	.0407064	4.96
diagonal	-.5789019	.0313059	-18.49

Comp2		λ_1	λ_2
length	-.0107095	.0331183	-0.32
heightleft	-.0714465	.0270489	-2.64
heightright	-.0662823	.029525	-2.24
frmlowborder	.5630723	.0460388	12.23

Seite 1

Starker Einfluss auf 2. HK

Keine Signifikanz $\Rightarrow \lambda_{length,2} = 0$

Der Einfluss
von length
auf 2. HK
ist relativ
gering

Einfluss
auf
2. HK

		Sender Einfluss auf 2. HK		Bank.log		
{	frmuppborder	<u>- .65929</u>	.034834	-18.93	0.000	-.7275634
	diagonal	<u>.4885424</u>	.0452855	10.79	0.000	.3997844

Comp3						
{	length	.3263183	.1838544	1.77	0.076	-.0340297
	heightleft	.2589621	.1475784	1.75	0.079	-.0302863
	heightright	.3447349	.1370982	2.51	0.012	.0760274
	frmlowborder	.2180215	.0673138	3.24	0.001	.0860889
	frmuppborder	.5566841	.1460904	3.81	0.000	.2703522
{	diagonal	.591762	.086124	6.87	0.000	.4229622
						.7605619

Comp4						
{	length	.5616896	.1206168	4.66	0.000	.325285
	heightleft	.4554578	.0912776	4.99	0.000	.2765569
	heightright	.41533	.1199291	3.46	0.001	.1802733
	frmlowborder	-.1861093	.0745136	-2.50	0.013	-.3321532
	frmuppborder	-.4507004	.1780348	-2.53	<u>0.011</u>	-.7996422
{	diagonal	-.2584502	.1883464	-1.37	<u>0.170</u>	-.6276023
						.1107018

Comp5						
{	length	-.7525739	.0518166	-14.52	0.000	-.8541325
	heightleft	.3468013	.0729277	4.76	0.000	.2038656
	heightright	.5346499	.0643185	8.31	0.000	.4085881
	frmlowborder	-.0999675	.026521	-3.77	0.000	-.1519478
	frmuppborder	-.1019017	.0542561	-1.88	<u>0.060</u>	-.2082418
{	diagonal	.0844595	.0459164	1.84	<u>0.066</u>	-.0055349
						.1744539

Comp6						
{	length	-.0980959	.0633015	-1.55	<u>0.121</u>	-.2221646
	heightleft	.7665121	.0329975	23.23	0.000	.7018381
	heightright	-.6316969	.0459089	-13.76	0.000	-.7216766
	frmlowborder	.022217	.0160259	1.39	<u>0.166</u>	-.0091931
	frmuppborder	.0348594	.0272464	1.28	<u>0.201</u>	-.0185426
{	diagonal	.0456795	.0235441	1.94	<u>0.052</u>	-.000466
						.0918251

LR test for independence: chi2(15) = 511.57 Prob > chi2 = 0.0000
 LR test for sphericity: chi2(20) = 1203.40 Prob > chi2 = 0.0000

Explained variance by components $\psi_1 - \psi_{q-1}$ ψ_q						
Components	Eigenvalue	Proportion	SE_Prop	Cumulative	SE_Cum	Bias
Comp1	3.000305	0.6675	0.0266	0.6675	0.0266	.009782
Comp2	.9356206	0.2082	0.0216	0.8757	0.0108	-.003271
Comp3	.2434134	0.0542	0.0064	0.9298	0.0066	.002753
Comp4	.1946587	0.0433	0.0051	0.9731	0.0027	-.006154
Comp5	.0852121	0.0190	0.0023	0.9921	0.0010	-.002016
Comp6	.0355148	0.0079	0.0010	1.0000	0.0000	-.001094

```

. * 1.1.1 Erklärungsgehalt der Variablen
. mat MEig=e(L) /* Speichern der Eigenvektoren (gamma)*/
. mat MEV=e(Ev) /* Speichern der Eigenwerte (lambda)*/

```

```

. sum length

```

Variable	Obs	Mean	Std. Dev.	Min	Max
length	200	214.896	.3765542	213.8	216.3

Bank.log

```
. scalar lvar=r(Var) /* speichern der Varianz der Länge*/
. scalar rho=MEig[1,1]*(MEV[1,1]/lvar)^(1/2) /*Berechnung von rho(x1,y1)*/
```

```
. display rho
-.20136085
```

$$= S_{x_{length}, y_1} = \sigma_{length, 1} \sqrt{\frac{\lambda_1}{\text{Var}(X_{length})}}$$

```
. * 1.1.2 screeplot
. screeplot, scheme(s2mono) ci mean
. graph save scree, replace
(file scree.gph saved)
```

mit Mittelwert
mit Konfidenzintervall

```
. * 1.1.3 Test-Statistics
. mat UMat=J(5,1,0)
. mat pMat=J(5,1,0)
. local lnbar=MEV[1,6]
. local sumln=ln(MEV[1,6])
```

```
. forvalues k=2/6 {
2. local lnbar=ln(((`k'-1)/(`k'))*(exp(`lnbar')+(1/(`k'-1))*MEV[1,6-`k'+1])))
3. local sumln=sumln'+ln(MEV[1,6-`k'+1])
4. mat UMat[`k'-1,1]=(200-(2*6+11)/6)*(`k'*`lnbar'-`sumln')
5.
6. local v=(1/2)*(`k'-1)*(`k'+2)
7. mat pMat[`k'-1,1]=chi2tail(`v',UMat[`k'-1,1])
}
```

Berechnung von U_R und $P(U_R \leq X)$

```
. matlist UMat
```

	c1
r1	910.8444
r2	973.9544
r3	996.9272
r4	1067.31
r5	1430.177

$$\left. \begin{array}{l} r1 \\ r2 \\ r3 \\ r4 \\ r5 \end{array} \right\} U_R = \left(n - \frac{2 \cdot 6 + 11}{6} \right) \left(R \ln \bar{R} - \sum_{j=p-R+1}^p Z_j \right) \approx \chi^2_v$$

```
. matlist pMat
```

	c1
r1	1.6e-198
r2	2.6e-208
r3	7.9e-209
r4	5.6e-219
r5	3.8e-291

$$\left. \begin{array}{l} r1 \\ r2 \\ r3 \\ r4 \\ r5 \end{array} \right\} P(U_R \leq X) \text{ mit } X \sim \chi^2_v$$

standardisierte
Daten

```
. * 1.2 Hauptkomponentenanalyse mit Korrelationsmatrix
. pca length heightleft heightright frm1lowborder frm1uppborder diagonal, vce(normal)
(with PCA/correlation, SEs and tests are approximate)
```

Principal components/correlation		Bank.log		Number of obs = 200		
				Number of comp. = 6		
				Trace = 6		
				Rho = 1.0000		
SEs assume multivariate normality		Standardfehler		SE(Rho) = 0.0000		
				gelten nicht, da nicht Konvergenz erreicht wird		
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Eigenvalues						
Comp1	2.945558	2939589	10.02	0.000	2.36941	3.521707
Comp2	1.278082	.1252163	10.21	0.000	1.032663	1.523502
Comp3	.8690328	.0839859	10.35	0.000	.7044235	1.033642
Comp4	.4497695	.0437462	10.28	0.000	.3640284	.5355105
Comp5	.2686771	.0257229	10.45	0.000	.2182611	.3190931
Comp6	.1888797	.0179952	10.50	0.000	.1536098	.2241497
Comp1						
length	-.006988	.0696884	-0.10	0.920	-.1435747	.1295988
heightleft	.4677581	.0353205	13.24	0.000	.3985312	.5369851
heightright	.4866785	.0304384	15.99	0.000	.4270204	.5463366
frmlowborder	.4067585	.0420445	9.67	0.000	.3243527	.4891643
frmuppborder	.367891	.0447397	8.22	0.000	.2802029	.4555791
diagonal	-.4934585	.0303297	-16.27	0.000	-.5529036	-.4340134
Comp2						
length	.8154953	.0374262	21.79	0.000	.7421414	.8888493
heightleft	.3419663	.0571723	5.98	0.000	.2299107	.4540219
heightright	.25246	.0603703	4.18	0.000	.1341363	.3707837
frmlowborder	-.2662277	.1157599	-2.30	0.021	-.493113	-.0393425
frmuppborder	-.0914868	.1476963	-0.62	0.536	-.3809662	.1979925
diagonal	.2739402	.0576495	4.75	0.000	.1609494	.3869311
Comp3						
length	-.0176821	.1604891	-0.11	0.912	-.3322349	.2968706
heightleft	.1033842	.0880152	1.17	0.240	-.0691225	.2758909
heightright	.1234737	.0794644	1.55	0.120	-.0322737	.2792212
frmlowborder	.583538	.0720062	8.10	0.000	.4424085	.7246675
frmuppborder	-.7875716	.0358132	-21.99	0.000	-.8577641	-.7173791
diagonal	.1138757	.0774871	1.47	0.142	-.0379963	.2657476
Comp4						
length	-.574617	.0534713	-10.75	0.000	-.6794188	-.4698152
heightleft	.3949244	.0944952	4.18	0.000	.2097171	.5801316
heightright	.4302763	.0916367	4.70	0.000	.2506716	.609881
frmlowborder	-.4036752	.0803006	-5.03	0.000	-.5610615	-.2462889
frmuppborder	-.1102266	.0951894	-1.16	0.247	-.2967944	.0763412
diagonal	.3919296	.0730782	5.36	0.000	.248699	.5351602
Comp5						
length	.0587933	.085109	0.69	0.490	-.1080172	.2256038
heightleft	-.6394962	.0822113	-7.78	0.000	-.8006273	-.4783651
heightright	.6140996	.0925006	6.64	0.000	.4328018	.7953974
frmlowborder	.2154742	.1132892	1.90	0.057	-.0065686	.437517
frmuppborder	.2198466	.0965222	2.28	0.023	.0306665	.4090267
diagonal	.3401588	.137983	2.47	0.014	.0697171	.6106005
Comp6						
length	-.0310563	.0536237	-0.58	0.562	-.1361568	.0740442
heightleft	.2977457	.1322403	2.25	0.024	.0385594	.5569319
heightright	-.3491508	.127912	-2.73	0.006	-.5998536	-.0984479
frmlowborder	.4623537	.060032	7.70	0.000	.344693	.5800143

			Bank.log			
frmuppborder	.4189688	.0562163	7.45	0.000	.308787	.5291507
diagonal	.6317998	.0758806	8.33	0.000	.4830765	.7805231

LR test for independence: chi2(15) = 511.57 Prob > chi2 = 0.0000
 LR test for sphericity: chi2(20) = 513.16 Prob > chi2 = 0.0000

Explained variance by components

Components	Eigenvalue	Proportion	SE_Prop	Cumulative	SE_Cum	Bias
Comp1	2.945558	0.4909	0.0284	0.4909	0.0284	.022594
Comp2	1.278082	0.2130	0.0201	0.7039	0.0199	.008567
Comp3	.8690328	0.1448	0.0147	0.8488	0.0115	-.011928
Comp4	.4497695	0.0750	0.0081	0.9237	0.0066	-.005821
Comp5	.2686771	0.0448	0.0050	0.9685	0.0035	-.00528
Comp6	.1888797	0.0315	0.0035	1.0000	0.0000	-.008132

```
. * 1.2.2 screeplot
. screeplot, scheme(s2mono) ci mean
(caution is advised in interpreting an asymptotic theory-based confidence interval of
eigenvalues of a correlation matrix)

. graph save scree2, replace
(file scree2.gph saved)

. * 1.2.3 vorhersage der Hauptkomponenten
. loadingplot, scheme(s2mono) factors(6) combined

. graph save loading, replace
(file loading.gph saved)

. scoreplot, scheme(s2mono) factors(6) combined

. graph save score, replace
(file score.gph saved)

. log close
  log: u:\Übung4\Bank.log
  log type: text
  closed on: 18 Jun 2011, 14:06:28
```