HKA = Hauptkomponentenanalyse

$(Y = \delta' X) \mid \max_{\delta} Var(\delta' X)$ unter der NB: $\delta'\delta = 1 \Rightarrow Y_1 = \gamma_1 X$, wobei γ_1 das δ ist welches $Var(\delta' X)$ maximiert.

HKA.log

```
      log:  U:\einfuehrung\HKA.log
  log type:  text
 opened on:   1 Jul 2010, 14:08:12

. pca length heightleft heightright frmlowborder frmuppborder diagonal, vce(norm
> al) cov
```

HKA mit Kovarianzmatrix, i.e. Maximieren Varianz!

```
Principal components/covariance                 Number of obs    =       200
                                                Number of comp.  =         6
                                                Trace            =  4.494725
                                                Rho              =    1.0000
SEs assume multivariate normality               SE(Rho)          =    0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Eigenvalues						
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						
length	-.0437744	.0159486	-2.74	0.006	-.0750331	-.0125156
heightleft	.1121618	.0134047	8.37	0.000	.0858891	.1384345
heightright	.1391907	.0142387	9.78	0.000	.1112834	.1670981
frmlowborder	.7683049	.0328803	23.37	0.000	.7038607	.8327491
frmuppborder	.2017662	.0407064	4.96	0.000	.121983	.2815493
diagonal	-.5789019	.0313059	-18.49	0.000	-.6402603	-.5175436
Comp2						
length	-.0107095	.0331183	-0.32	0.746	-.0756202	.0542012
heightleft	-.0714465	.0270489	-2.64	0.008	-.1244614	-.0184316
heightright	-.0662823	.029525	-2.24	0.025	-.1241502	-.0084143
frmlowborder	.5630723	.0460388	12.23	0.000	.4728379	.6533067
frmuppborder	-.65929	.034834	-18.93	0.000	-.7275634	-.5910166
diagonal	.4885424	.0452855	10.79	0.000	.3997844	.5773004
Comp3						
length	.3263183	.1838544	1.77	0.076	-.0340297	.6866664
heightleft	.2589621	.1475784	1.75	0.079	-.0302863	.5482105
heightright	.3447349	.1370982	2.51	0.012	.0760274	.6134423
frmlowborder	.2180215	.0673138	3.24	0.001	.0860889	.3499541
frmuppborder	.5566841	.1460904	3.81	0.000	.2703522	.843016
diagonal	.591762	.086124	6.87	0.000	.4229622	.7605619
Comp4						
length	.5616896	.1206168	4.66	0.000	.325285	.7980942
heightleft	.4554578	.0912776	4.99	0.000	.2765569	.6343586
heightright	.41533	.1199291	3.46	0.001	.1802733	.6503866
frmlowborder	-.1861093	.0745136	-2.50	0.013	-.3321532	-.0400654
frmuppborder	-.4507004	.1780348	-2.53	0.011	-.7996422	-.1017585
diagonal	-.2584502	.1883464	-1.37	0.170	-.6276023	.1107018
Comp5						
length	-.7525739	.0518166	-14.52	0.000	-.8541325	-.6510153
heightleft	.3468013	.0729277	4.76	0.000	.2038656	.489737
heightright	.5346499	.0643185	8.31	0.000	.4085881	.6607118
frmlowborder	-.0999675	.026521	-3.77	0.000	-.1519478	-.0479873
frmuppborder	-.1019017	.0542561	-1.88	0.060	-.2082418	.0044384
diagonal	.0844595	.0459164	1.84	0.066	-.0055349	.1744539
Comp6						
length	-.0980959	.0633015	-1.55	0.121	-.2221646	.0259729
heightleft	.7665121	.0329975	23.23	0.000	.7018381	.8311861

Kennzeichnet vergleichsweise hohe Werte

HKe = Hauptkomponente

heightright	-.6316969	.0459089	-13.76	0.000	-.7216766	-.5417172
frmlowborder	.022217	.0160259	1.39	0.166	-.0091931	.0536271
frmuppborder	.0348594	.0272464	1.28	0.201	-.0185426	.0882613
diagonal	.0456795	.0235441	1.94	0.052	-.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

erste HKe „erklärt" 67% der Gesamtvarianz
zweite " " 21% " "

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

zusammen erklären 1ste & 2te HKE über 80% der Gesamtvarianz

```
. pca length heightleft heightright frmlowborder frmuppborder diagonal, vce(norm
> al)
(with PCA/correlation, SEs and tests are approximate)
```

Standardbefehl pca ohne cov führt HKA unter Maximierung der Korrelation durch

Principal components/correlation — Number of obs = 200
Number of comp. = 6
Trace = 6
Rho = 1.0000
SEs assume multivariate normality — SE(Rho) = 0.0000

	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

Augenmaß für Anzahl der HKen: a) Größe der Eigenwerte
b) Wieviele HKen braucht man um etwa 80% der Gesamtvarianz zu erklären? (Meistens reichen 2-3 HKen)

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
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

weniger als vorher!

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

```
. log close
      log:  U:\einfuehrung\HKA.log
  log type:  text
 closed on:   1 Jul 2010, 14:22:48
```

Problem: Skalierung der Daten (durch verschiedene Einheiten, wie Jahr, Kilogramm, Dollar)

Bsp: $X' = (X_1, aX_2, \dots, aX_p)$

Für $a \to 0$ gilt $\mathrm{Var}(aX_j) = a^2 \mathrm{Var}(X_j) \to 0 \quad \forall j \in \{2, \dots, p\}$

$\Rightarrow$ lediglich $\mathrm{Var}(X_1) > 0$ $\Rightarrow$ erste HKe $Y_1 \approx X_1$ für $a \to 0$

Idee: Anstatt Kovarianzmatrix nehmen wir Korrelationsmatrix, da diese skalierungsinvariant ist.

Nachteil: Maximieren nicht mehr Varianz, sondern Korrelation

$\Rightarrow$ Gesamtvarianz keine richtige Varianz mehr, sondern immer gleich der Anzahl an Elementen in X.

Fazit: HKA mit Korrelationsmatrix ist nur sinnvoll im Falle stark unterschiedlicher Varianzen oder unproportionalen Maßeinheiten