

Hauptkomponentenanalyse PCA

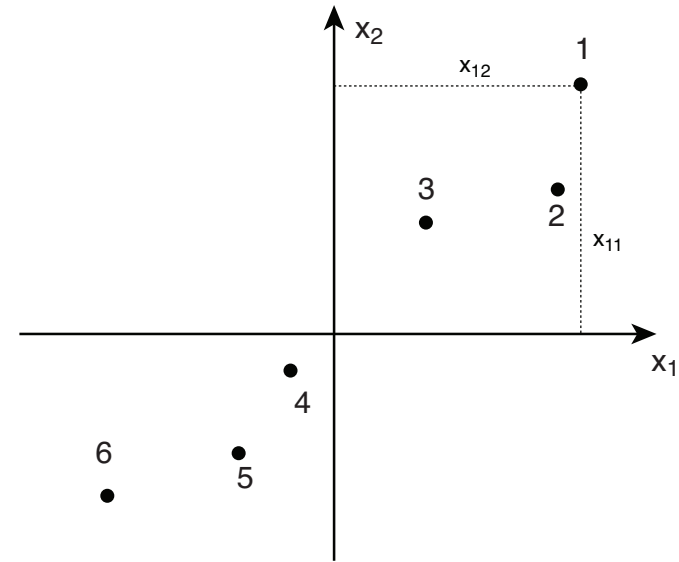
Die Hauptkomponentenanalyse (Principal Component Analysis, PCA) ist eine Methode zur linearen Transformation der Variablen, so dass:

- möglichst wenige neue Variablen die relevante Information beschreiben. (Mass für die Relevanz ist die Varianz)
- die neuen Variablen orthogonal und damit unkorreliert sind

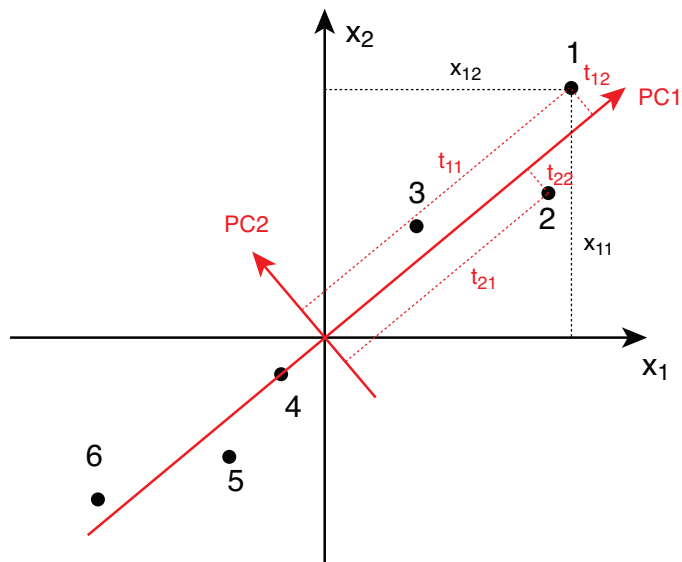
Zweck:

- explorative Datenanalyse
Entdeckung von Zusammenhängen in 2D- oder 3D-Plots
- Modellbildung (z.B. Regression) mit den transformierten Daten
eliminiert irrelevante Information wie Rauschen
eliminiert die Gefahr von Artefakten wie Ausreißern

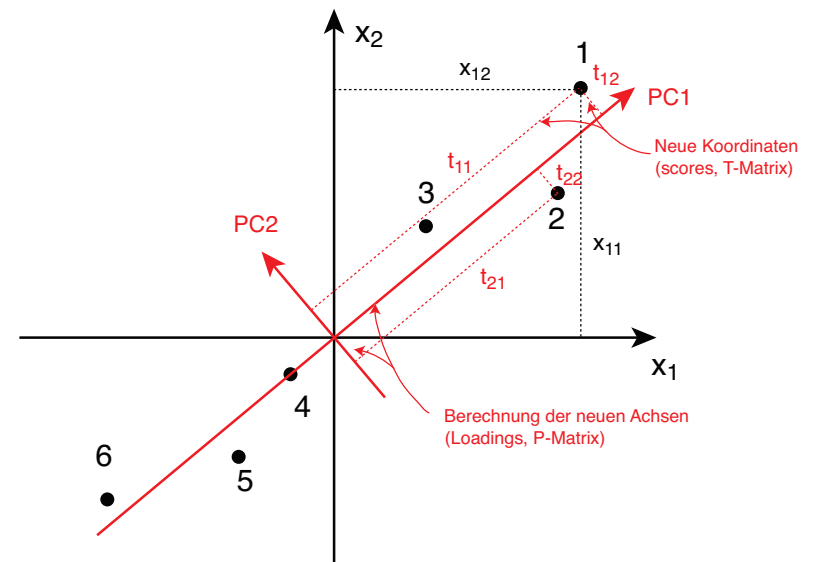
Hauptkomponentenanalyse



Hauptkomponentenanalyse



Hauptkomponentenanalyse

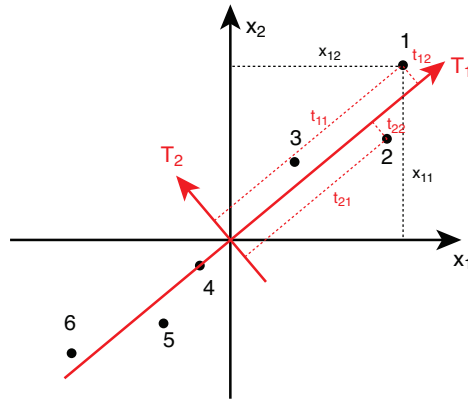


Hauptkomponentenanalyse

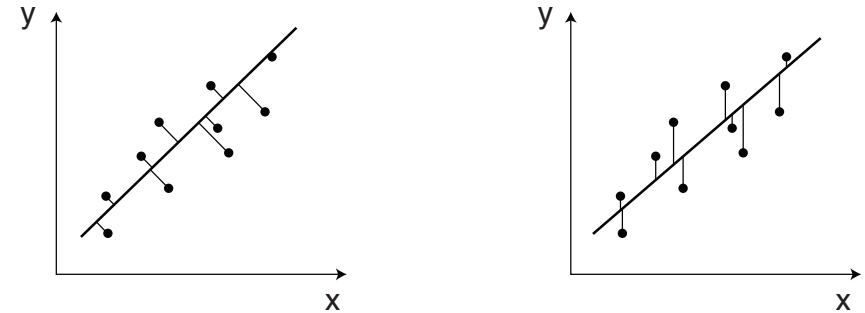
Es handelt sich um eine Rotation des Koordinatensystems (Multiplikation mit einer orthogonalen Matrix).

Mass für den Informationsgehalt ist der Anteil an der totalen Varianz, den die einzelnen Hauptkomponenten beschreiben.

Der Informationsgehalt der höheren Hauptkomponenten nimmt oft rapide ab, so dass sie ohne Verlust an Information weggelassen können.



Hauptkomponentenanalyse und Lineare Regression



Bei der Hauptkomponentenanalyse werden die Fehlerquadrate senkrecht zur Geraden minimiert (orthogonale Regression), bei der linearen Regression diejenige in der y-Richtung.

Hauptkomponentenanalyse

$$\begin{matrix} m & & m & & m & & m \\ \square & = & \square & + & \square & + \dots + & \square \\ n & \mathbf{X} & n & \mathbf{M}_1 & n & \mathbf{M}_2 & n & \mathbf{M}_r \\ & & & 1 & & 2 & & r \end{matrix}$$

- Eine Matrix vom Rang r kann als eine Summe von r Matrizen vom Rang 1 ausgedrückt werden.
- Für eine $n \times m$ Matrix mit $n > m$, ist der Rang $r \leq m$.
- Der Rang einer Matrix entspricht der Ordnung der grössten (quadratischen) Submatrix, deren Determinante (D) ungleich null ist.
(Eine Submatrix kann aus einer Matrix durch Weglassen einer beliebigen Anzahl von Zeilen und/oder Spalten gebildet werden).
- Der Rang entspricht der Anzahl unabhängiger Informationen. Wenn eine Spalte oder Zeile einer Matrix als Linearkombination anderer Spalten/Zeilen ausgedrückt werden kann, ist die entsprechende Information nicht unabhängig.
- Eine quadratische Matrix mit $D = 0$ ist singulär und kann nicht invertiert werden.

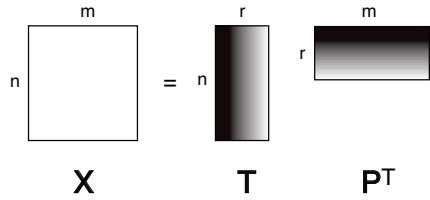
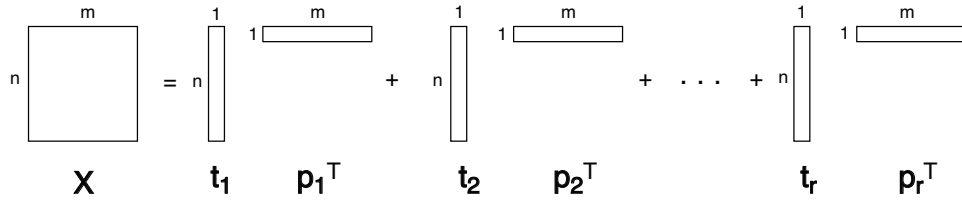
Hauptkomponentenanalyse

$$\begin{matrix} m & & 1 & & m & & 1 & & m & & 1 & & m \\ \square & = & \square & + & \square & + & \dots & + & \square & + & \dots & + & \square \\ n & \mathbf{X} & n & \mathbf{t}_1 & n & \mathbf{p}_1^T & n & \mathbf{t}_2 & n & \mathbf{p}_2^T & n & \mathbf{t}_r & n & \mathbf{p}_r^T \end{matrix}$$

$$\begin{matrix} m & & r & & m \\ \square & = & \square & \square \\ n & \mathbf{X} & n & \mathbf{T} & r & \mathbf{P}^T \end{matrix}$$

- Bei der Hauptkomponentenanalyse werden die Vektoren \mathbf{t} und \mathbf{p} so ausgewählt, dass
1. die \mathbf{p} Vektoren paarweise orthonormal sind,
 2. die \mathbf{t} Vektoren orthogonal sind,
 3. jeder \mathbf{t} -Vektor (Scores, neue Koordinaten) das Maximum der verbleibenden Varianz beschreibt.

Hauptkomponentenanalyse



Die weniger relevanten Hauptkomponenten können oft ohne wesentlichen Verlust an Information weggelassen werden. Sie beschreiben viel Rauschen und enthalten kaum relevante Information.

Hauptkomponentenanalyse

X					
110.71	102.05	88.52	504.01	76.54	66.49
102.05	267.23	240.15	1.24E+3	187.19	159.24
88.52	240.15	220.72	1.1E+3	165.44	141.35
504.01	1.24E+3	1.1E+3	5.85E+3	886.6	752.89
76.54	187.19	165.44	886.6	134.42	113.99
66.49	159.24	141.35	752.89	113.99	97.16

t1						p1^T												
587.8	0.082	0.199	0.177	0.942	0.143	0.121	44.01	107.36	95.3	507.49	76.87	65.31	66.7	-5.31	-6.78	-3.49	-0.33	1.18
1314							107.4	261.86	232.46	1.24E+3	187.5	159.3	-5.31	5.37	7.69	-2.06	-0.31	-0.06
1167							95.3	232.46	206.36	1.1E+3	166.45	141.4	-6.78	7.69	14.36	-3.57	-1.01	-0.06
6212							507.5	1.24E+3	1.1E+3	5.85E+3	886.34	753	-3.49	-2.06	-3.57	1.38	0.27	-0.10
941.0							76.87	187.5	166.45	886.34	134.25	114.1	-0.33	-0.31	-1.01	0.27	0.16	-0.07
799.4							65.31	159.29	141.41	753	114.06	96.9	1.19	-0.06	-0.06	-0.10	-0.07	0.26

t2						p2^T												
-67.3	-0.985	0.098	0.135	0.041	0.003	-0.017	66.28	-6.61	-9.09	-2.77	-0.17	1.17	0.42	1.30	2.31	-0.72	-0.16	0.01
6.71							-6.61	0.66	0.91	0.28	0.02	-0.12	1.30	4.71	6.79	-2.34	-0.33	0.06
9.23							-9.09	0.91	1.25	0.38	0.02	-0.16	2.31	6.79	13.11	-3.96	-1.03	0.10
2.81							-2.77	0.28	0.38	0.17	0.007	-0.05	-0.72	-2.34	-3.96	1.27	0.26	-0.05
0.17							-0.17	0.02	0.02	0.007	0.0004	-0.003	-0.16	-0.33	-1.03	0.26	0.17	-0.06
-1.19							1.17	-0.12	-0.16	-0.05	-0.003	0.021	0.013	0.061	0.101	-0.054	-0.06	0.239

t3						p3^T												
2.78	0.149	0.464	0.831	-0.259	-0.06	0.007	0.42	1.29	2.31	-0.72	-0.17	0.02	0.0004	0.01	-0.005	-0.001	0.004	-0.007
8.65							1.29	4.02	7.19	-2.24	-0.52	0.062	0.01	0.69	-0.4	-0.1	0.19	-0.001
15.48							2.31	7.19	12.87	-4.01	-0.92	0.111	-0.006	-0.4	0.24	0.06	-0.11	-0.01
-4.83							-0.72	-2.24	-4.01	1.25	0.29	-0.035	-0.0001	-0.1	0.06	0.02	-0.03	-0.02
-1.11							-0.17	-0.52	-0.92	0.29	0.07	-0.008	0.004	0.19	-0.11	-0.03	0.1	-0.05
0.13							0.02	0.06	0.11	-0.04	-0.008	0.001	0.007	-0.001	-0.01	-0.02	-0.05	0.24

M1 = t1*p1^T						X - M1					
66.28	-6.61	-9.09	-2.77	-0.17	1.17	0.42	1.30	2.31	-0.72	-0.16	0.01
-6.61	0.66	0.91	0.28	0.02	-0.12	1.30	4.71	6.79	-2.34	-0.33	0.06
-9.09	0.91	1.25	0.38	0.02	-0.16	2.31	6.79	13.11	-3.96	-1.03	0.10
-2.77	0.28	0.38	0.17	0.007	-0.05	-0.72	-2.34	-3.96	1.27	0.26	-0.05
-0.17	0.02	0.02	0.007	0.0004	-0.003	-0.16	-0.33	-1.03	0.26	0.17	-0.06
1.17	-0.12	-0.16	-0.05	-0.003	0.021	0.013	0.061	0.101	-0.054	-0.06	0.239

M2 = t2*p2^T						X - M1 - M2					
0.42	1.29	2.31	-0.72	-0.17	0.02	0.0004	0.01	-0.005	-0.001	0.004	-0.007
1.29	4.02	7.19	-2.24	-0.52	0.062	0.01	0.69	-0.4	-0.1	0.19	-0.001
2.31	7.19	12.87	-4.01	-0.92	0.111	-0.006	-0.4	0.24	0.06	-0.11	-0.01
-0.72	-2.24	-4.01	1.25	0.29	-0.035	-0.0001	-0.1	0.06	0.02	-0.03	-0.02
-0.17	-0.52	-0.92	0.29	0.07	-0.008	0.004	0.19	-0.11	-0.03	0.1	-0.05
0.02	0.06	0.11	-0.04	-0.008	0.001	0.007	-0.001	-0.01	-0.02	-0.05	0.24

M3 = t3*p3^T						X - M1 - M2 - M3					
0.42	1.30	2.31	-0.72	-0.16	0.01	0.0004	0.01	-0.005	-0.001	0.004	-0.007
1.30	4.71	6.79	-2.34	-0.33	0.06	0.01	0.69	-0.4	-0.1	0.19	-0.001
2.31	6.79	13.11	-3.96	-1.03	0.10	-0.006	-0.4	0.24	0.06	-0.11	-0.01
-0.72	-2.34	-3.96	1.27	0.26	-0.05	-0.0001	-0.1	0.06	0.02	-0.03	-0.02
-0.16	-0.33	-1.03	0.26	0.17	-0.06	0.004	0.19	-0.11	-0.03	0.1	-0.05
0.013	0.061	0.101	-0.054	-0.06	0.239	0.007	-0.001	-0.01	-0.02	-0.05	0.24

Hauptkomponentenanalyse: Mathematische Grundlagen

Bei der Hauptkomponentenanalyse sucht man nach einer linearen Transformation der Koordinaten **X** (Transformationsmatrix **P**) in neue Koordinaten **T** (**XP = T**), so dass ein möglichst grosser Teil der Varianz beschrieben ist, und dass die Transformationsmatrix orthonormal ist.

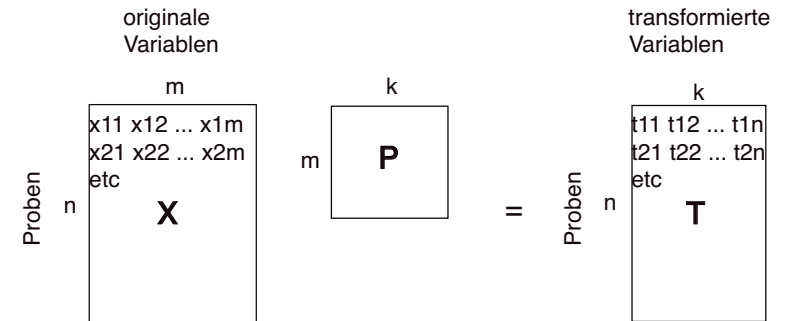
- Man sucht also bei jedem Schritt nach den **p**- und **t**-Vektoren, die die Bedingungen erfüllen:
- $t^T t = \max$ d.h. $(Xp)^T (Xp) = p^T X^T X p = \max$ ($X^T X$ ist die Varianz-Kovarianzmatrix * (n-1))
 - $p^T p = 1$

Optimierung mit Nebenbedingungen, Methode der Lagrange-Multiplikatoren:
 Optimierung der Funktion $f(x,y)$ unter der Bedingung $g(x,y) = 0$
 Vorgehen: Definition einer allgemeineren Funktion: $u(x,y,\lambda) = f(x,y) - \lambda g(x,y)$
 $p^T X^T X p$ maximal, Nebenbedingung: $p^T p = 1$
 Funktion: $u = p^T X^T X p - \lambda(p^T p - 1)$

$du/dp = 2X^T X p - 2\lambda p = 0$
 $(A - \lambda I)p = 0$ (mit $A = X^T X$): klassisches Eigenwertproblem. **p** ist jeweils ein Eigenvektor der Varianz-Kovarianzmatrix.

Hauptkomponentenanalyse

$X = T P^T$
 $T = X P$ **P** ist eine orthogonale Matrix, daher ist die Inverse die Transponierte.



Die zu den einzelnen Eigenvektoren **p_i** gehörenden Eigenwerte λ_i geben den Anteil der Gesamtvarianz an, der durch sie repräsentiert ist.

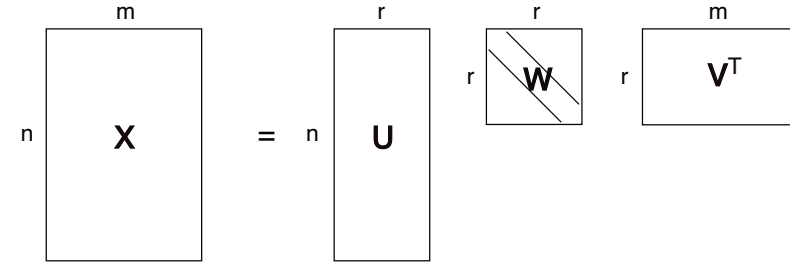
Hauptkomponentenanalyse: Das Vorgehen

1. Berechnung der Korrelationsmatrix $\mathbf{X}^T\mathbf{X}$ der zentrierten und mit der Varianz skalierten Daten.
2. Berechnung der Eigenvektoren und Eigenwerte der $\mathbf{X}^T\mathbf{X}$ -Matrix. Die Eigenvektoren solcher Matrizen sind reell, und die Eigenwerte sind nicht negativ.
3. Auswahl der Anzahl signifikanter Eigenwerte. Der Anteil der Varianz, die die ausgewählten r Komponenten beschreiben, ist: $\sum \lambda_i / k$ (k ist die Dimension der $\mathbf{X}^T\mathbf{X}$ -Matrix). Die den ausgewählten Eigenwerten entsprechenden Eigenvektoren bilden die Koeffizienten für die Linearkombination der ursprünglichen X -Variablen für die Berechnung der neuen Variablen \mathbf{T} ("scores"), d.h. sie bilden die "loadings"-Matrix.
4. Berechnung der \mathbf{T} -Matrix ("scores"): $\mathbf{T} = \mathbf{X}\mathbf{P}$

Man kann die Hauptkomponenten auch ohne Skalierung oder Zentrierung der Variablen berechnen. Je nach Aufgabestellung kann die eine oder die andere Art sinnvollere Resultate geben.

Singularwert-Zerlegung

Allgemein gilt, dass jede Matrix \mathbf{X} in das Produkt dreier Matrizen zerlegt werden kann: Singularwert-Zerlegung (Singular value decomposition): $\mathbf{X}_{n \times m} = \mathbf{U}_{n \times r} \mathbf{W}_{r \times r} \mathbf{V}_{r \times m}^T$ mit \mathbf{U} und \mathbf{V} orthonormal und \mathbf{W} diagonal.



Zusammenhang mit der Hauptkomponentenanalyse: $\mathbf{U}\mathbf{W} = \mathbf{T}$; $\mathbf{V} = \mathbf{P}$
Die Eigenwerte der $\mathbf{X}^T\mathbf{X}$ und $\mathbf{X}\mathbf{X}^T$ Matrizen sind $\mathbf{\Lambda} = \mathbf{W}^2$. \mathbf{U} ist die Matrix der Eigenvektoren von $\mathbf{X}\mathbf{X}^T$ und \mathbf{V} die Matrix der Eigenvektoren von $\mathbf{X}^T\mathbf{X}$

Singularwert-Zerlegung (SVD): BSP

Bestimmung von drei Spurenelementen in Luftproben in Abhängigkeit von der Windrichtung:

Windrichtung	Na	Cl	Si
0°	0.212	0.399	0.190
90°	0.072	0.133	0.155
180°	0.036	0.063	0.213
270°	0.078	0.141	0.273

$$\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \begin{bmatrix} 0.753 & 0.618 \\ 0.343 & -0.127 \\ 0.302 & -0.567 \\ 0.473 & -0.529 \end{bmatrix} \begin{bmatrix} 0.626 & 0 \\ 0 & 0.214 \end{bmatrix} \begin{bmatrix} 0.371 & 0.690 & 0.622 \\ 0.280 & 0.556 & -0.783 \end{bmatrix} = \begin{bmatrix} 0.212 & 0.399 & 0.190 \\ 0.072 & 0.133 & 0.155 \\ 0.036 & 0.063 & 0.213 \\ 0.078 & 0.141 & 0.273 \end{bmatrix} = \mathbf{X}$$

Singularwert-Zerlegung (SVD): BSP

$$\mathbf{U}^T\mathbf{U} = \begin{bmatrix} 0.753 & 0.343 & 0.302 & 0.473 \\ 0.618 & -0.127 & -0.567 & -0.529 \end{bmatrix} \begin{bmatrix} 0.753 & 0.618 \\ 0.343 & -0.127 \\ 0.302 & -0.567 \\ 0.473 & -0.529 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{V}^T\mathbf{V} = \begin{bmatrix} 0.371 & 0.690 & 0.622 \\ 0.280 & 0.556 & -0.783 \end{bmatrix} \begin{bmatrix} 0.371 & 0.280 \\ 0.690 & 0.556 \\ 0.622 & -0.783 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$\mathbf{X}^T\mathbf{X} = (\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T)^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^T$
 \mathbf{V} und $\mathbf{\Lambda}^2$ sind Eigenvektoren und Eigenwerte der Matrix $\mathbf{X}^T\mathbf{X}$

$\mathbf{X}\mathbf{X}^T = (\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T)^T\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T\mathbf{V}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T$
 \mathbf{U} und $\mathbf{\Lambda}^2$ sind Eigenvektoren und Eigenwerte der Matrix $\mathbf{X}\mathbf{X}^T$

SVD und PCA

$$T = U \lambda = \begin{bmatrix} 0.753 & 0.618 \\ 0.343 & -0.127 \\ 0.302 & -0.567 \\ 0.473 & -0.529 \end{bmatrix} \begin{bmatrix} 0.626 & 0 \\ 0 & 0.214 \end{bmatrix} = \begin{bmatrix} 0.472 & 0.132 \\ 0.215 & -0.027 \\ 0.189 & -0.122 \\ 0.296 & -0.113 \end{bmatrix} =$$

$$= X V = \begin{bmatrix} 0.212 & 0.399 & 0.190 \\ 0.072 & 0.133 & 0.155 \\ 0.036 & 0.063 & 0.213 \\ 0.078 & 0.141 & 0.273 \end{bmatrix} \begin{bmatrix} 0.371 & 0.280 \\ 0.690 & 0.556 \\ 0.622 & -0.783 \end{bmatrix}$$

SVD und PCA

$$T = U \lambda = \begin{bmatrix} 0.753 & 0.618 \\ 0.343 & -0.127 \\ 0.302 & -0.567 \\ 0.473 & -0.529 \end{bmatrix} \begin{bmatrix} 0.626 & 0 \\ 0 & 0.214 \end{bmatrix} = \begin{bmatrix} 0.472 & 0.132 \\ 0.215 & -0.027 \\ 0.189 & -0.122 \\ 0.296 & -0.113 \end{bmatrix} =$$

$$= X V = \begin{bmatrix} 0.212 & 0.399 & 0.190 \\ 0.072 & 0.133 & 0.155 \\ 0.036 & 0.063 & 0.213 \\ 0.078 & 0.141 & 0.273 \end{bmatrix} \begin{bmatrix} 0.371 & 0.280 \\ 0.690 & 0.556 \\ 0.622 & -0.783 \end{bmatrix}$$

Teeproben: PCA

Teeproben: zentrierte Daten X-Matrix

-2.3	-4.38	-4.05	5.54	0.68	0.4
-1.74	-4.17	-4.01	4.35	0.53	0.28
-1.01	-3.82	-2.96	3.04	0.15	0.04
-0.49	-4.36	-2.56	1.67	-0.04	-0.13
-0.3	-3.2	-2.1	-0.21	-0.26	-0.32
0.3	-3.64	-1.97	-1.26	-0.38	-0.4
1.5	-3.6	-1.26	-2.39	-0.62	-0.55
-2.73	-3.95	-3.16	3.74	0.44	-0.24
-1.05	-3.48	-2.29	2.5	0.24	-0.42
-1.02	-3.56	-1.79	1.28	0.1	-0.51
0.2	-2.6	-0.38	0.56	-0.27	-0.61
0.37	-3.42	0.13	-0.47	-0.34	-0.82
-1.48	1.38	-2.51	-1.94	0.47	1.01
-0.81	0.83	-1.98	-2.85	0.38	0.93
0.52	0.84	-1.05	-3.43	0.26	0.7
1.24	-1.58	0.12	-4.15	-0.02	0.03
-0.85	-1.44	-2.36	3.19	1.27	1.9
-1.1	-1.48	-1.76	0.96	0.56	1.3
-0.99	-0.85	-1.58	0.25	0.35	1.08
-1.15	-0.87	-1.18	-0.28	0.23	0.86
-0.56	-1.15	0.03	-0.81	0.05	0.67
-0.69	-0.75	0.39	-0.95	-0.01	0.55
0.03	0.5	1.09	-1.33	-0.17	0.46
0.35	3.56	2.37	-2.84	-0.67	-0.45
0.33	3.07	2.97	-2.88	-0.79	-0.93
0.1	6.55	3.6	-2.97	-0.82	-1.52
0.12	6.3	4.29	-3.07	-0.93	-1.63
0.31	4.74	3.41	-4.53	-0.89	-0.62
0.94	4.95	3.58	-4.85	-1.04	-1.19
0.21	5.17	4.5	-4.63	-1.1	-1.58
0.05	5.14	5.02	-4.65	-1.12	-1.66

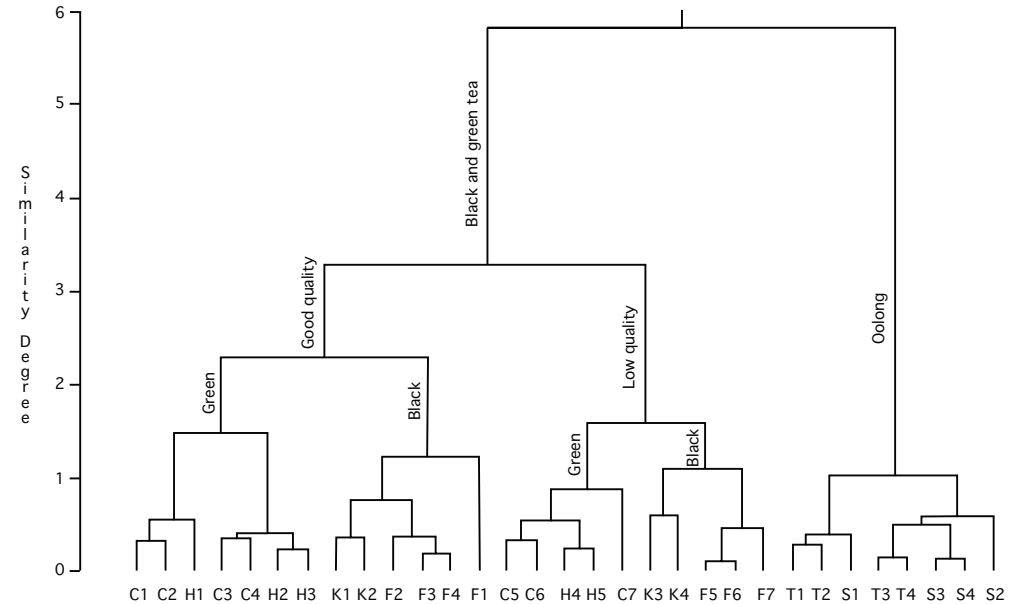
Teeproben: Kreuzprodukt-Matrix (X^TX)

32.74	46.84	51.35	-56.45	-10.77	-11.27	
46.84	383.95	256.25	-232.75	-41.13	-44.22	
51.35	256.25	217.2	-180.97	-40.61	-49.59	
-56.45	-232.75	-180.97	268.55	41.71	38.28	
-10.77	-41.13	-40.61	41.71	11.33	15.06	
-11.27	-44.22	-49.59	38.28	15.06	25.99	
Eigenwerte	Eigenvektoren (P-Matrix)					
7.277	-0.46	-0.02	0.79	0.34	-0.2	0.12
0.569	-0.3	-0.04	0.04	-0.41	0.54	0.67
17.95	0.58	0.07	0.04	0.62	0.16	0.5
45.252	-0.07	-0.08	0.12	0.28	0.79	-0.52
93.109	0.15	0.93	0.24	-0.21	0.08	-0.09
775.61	0.58	-0.36	0.56	-0.46	0.03	-0.1
0.77% 0.06% 1.9% 4.8% 9.9% 82.5%						

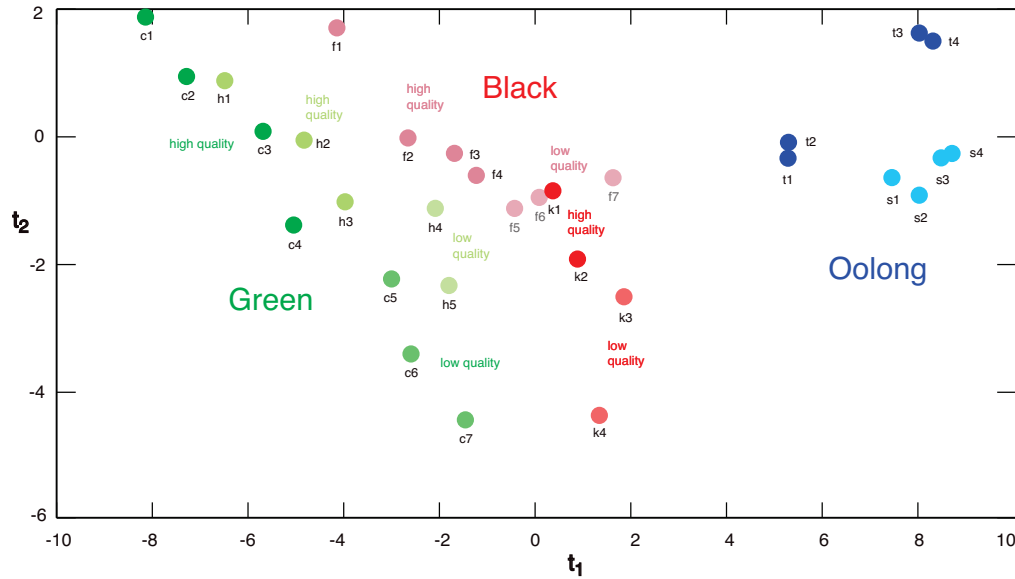
Neue Koordinaten: T = XP

-0.051	-0.047	-1.101	-0.273	1.908	-8.201
-0.355	-0.063	-0.896	-0.39	0.953	-7.334
-0.29	-0.174	-0.648	0.188	0.084	-5.75
-0.168	-0.141	-0.551	0.566	-1.359	-5.108
-0.342	-0.14	-0.718	0.049	-2.211	-3.068
-0.402	-0.124	-0.459	0.281	-3.397	-2.665
-0.597	-0.174	0.237	0.914	-4.428	-1.519
0.254	0.159	-2.015	-0.204	0.905	-6.502
-0.2	0.154	-0.938	0.491	-0.05	-4.874
0.113	0.192	-1.128	0.571	-1.002	-4.022
0.022	-0.012	-0.305	1.389	-1.109	-2.11
0.422	0.145	-0.443	1.918	-2.336	-1.824
-0.384	0.019	-0.759	-3.726	-0.823	0.342
-0.217	0.084	-0.41	-3.146	-1.905	0.804
-0.403	0.145	0.444	-2.155	-2.499	1.77
0.279	0.345	0.427	-0.035	-4.386	1.287
0.521	0.138	0.921	-1.412	1.707	-4.209
0.697	-0.078	-0.024	-1.308	-0.014	-2.687
0.455	-0.15	-0.161	-1.469	-0.253	-1.753
0.658	-0.107	-0.486	-1.29	-0.604	-1.278
1.07	-0.076	-0.2	-0.252	-1.123	-0.478
1.151	-0.063	-0.369	-0.208	-0.941	0.045
0.806	-0.156	0.143	-0.07	-0.616	1.54
-0.003	-0.198	-0.232	-0.315	-0.082	5.19
0.204	-0.072	-0.546	0.484	-0.303	5.241
-0.697	0.041	-0.9	-0.369	1.642	7.972
-0.306	0.046	-0.967	0.212	1.522	8.226
0.256	-0.174	-0.522	-0.516	-0.631	7.406
-0.328	-0.089	-0.403	-0.077	-0.902	7.945
0.223	0.049	-1.136	0.407	-0.327	8.399
0.558	0.103	-1.295	0.721	-0.247	8.642

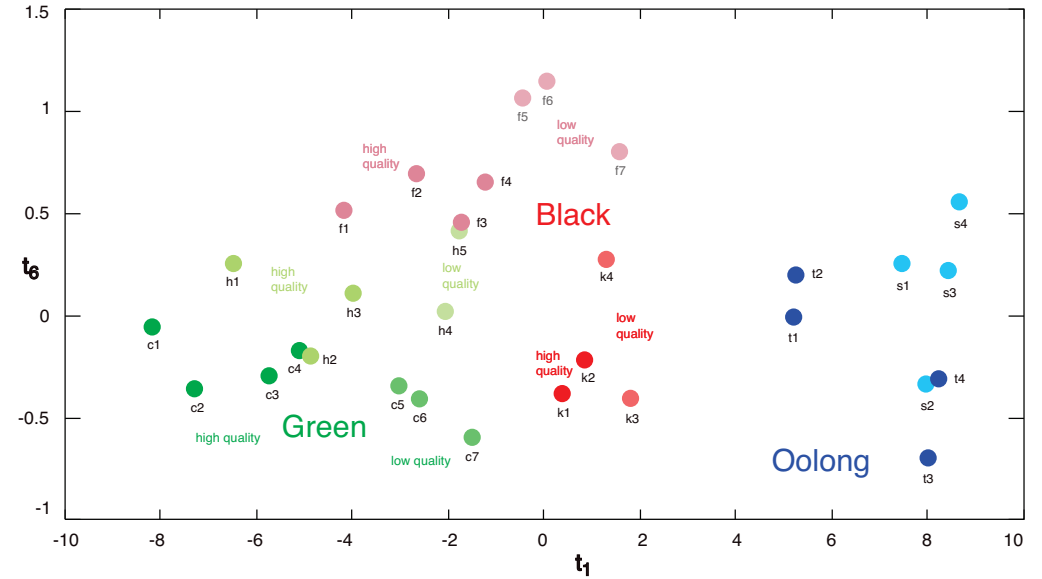
Teeproben: Clustering



Teeproben: Score Plots t_1 vs t_2



Teeproben: Score Plots t_1 vs t_6



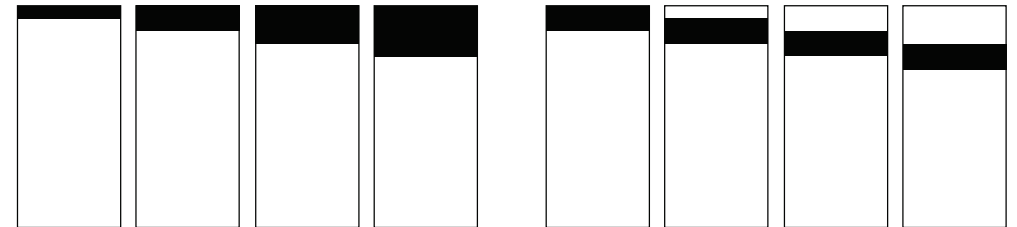
Evolvierende Faktoranalyse EFA

Die Evolvierende Faktoranalyse EFA ist eine Anwendung der Hauptkomponentenanalyse. Ein sich entwickelndes System wird beschrieben. Ein typisches Beispiel ist ein HPLC-UV-Experiment. Der Detektor liefert nicht nur einen Skalar, sondern ein ganzes Spektrum. Dieses kann als Vektor aufgefasst werden. Bei p verschiedenen Wellenlängen wird die Absorbanz gemessen. Das System entwickelt sich zeitlich. Es kommen also dauernd neue Datenvektoren dazu. Die Datenmatrix des ganzen Experiments besteht aus n Spektren, die zu n Zeitpunkten gemessen wurden.

Unter der Annahme, dass die chemischen Komponenten paarweise unterschiedliche UV-Spektren aufweisen, wird die PCA der Datenmatrix so viele Hauptkomponenten mit nennenswertem Eigenwert ergeben, wie chemische Komponenten eluiert wurden. Die anderen Hauptkomponenten beschreiben Rauschen und ähnliche irrelevante Phänomene. Ihre Eigenwerte sind deutlich kleiner als die der relevanten Komponenten.

Evolvierende Faktoranalyse EFA

Die zeitliche Entwicklung wird berücksichtigt, indem nur Teile der Datenmatrix verwendet werden. Bei der klassischen EFA wird die Datenmatrix sukzessive vergrößert, indem Vektoren so hinzugefügt werden, wie sie zeitlich entstanden sind. Nach jedem Hinzufügen wird eine PCA durchgeführt. Bei einer alternativen Methode wird eine fixe Anzahl Datenvektoren als Zeitfenster systematisch verschoben, indem ein Vektor dazukommt und der zeitlich früheste entfernt wird (Moving Window Factor Analysis).

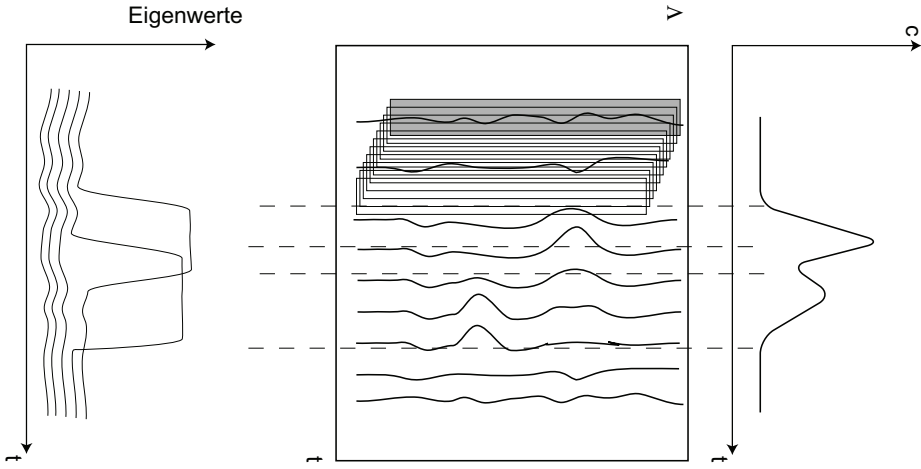


Evolving factor analysis

Moving window factor analysis

Die Anzahl chemischer Komponenten, die in den reduzierten Datenmatrizen vorhanden sind, lässt sich anhand der relevantesten Hauptkomponenten erkennen.

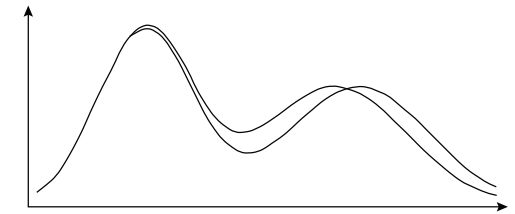
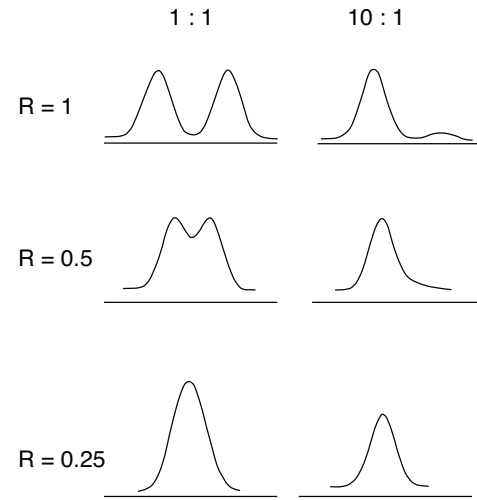
Evolvierende Faktoranalyse



Evolvierende Faktoranalyse

Zur Definition der Auflösung R siehe Vorlesung Analytische Chemie III

Angenommene Spektren

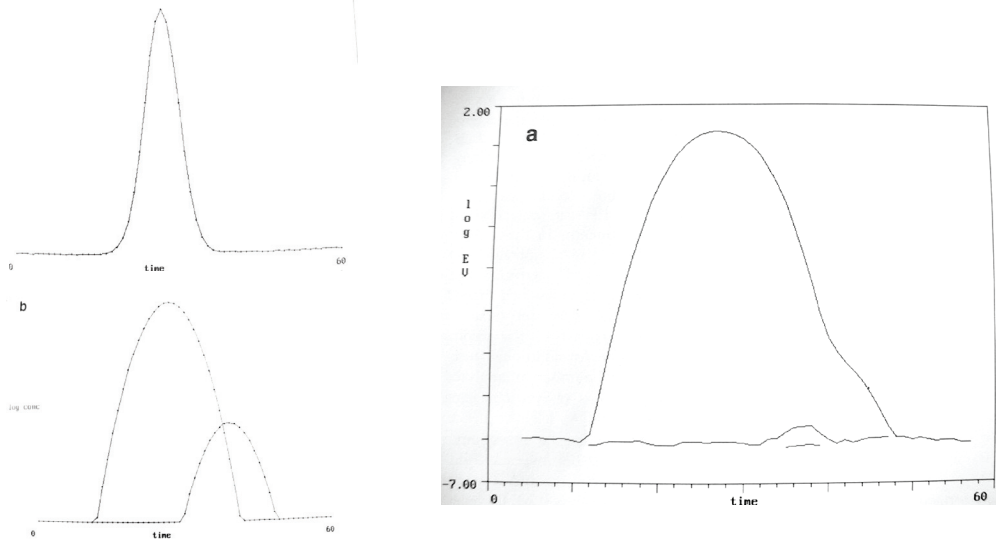


Tests mit:
 1. $R = 1.0$, 0.2% Verunreinigung
 2. $R = 0.2$, 0.5% Verunreinigung

Evolvierende Faktoranalyse

1. $R = 1.0$, 0.2% Verunreinigung

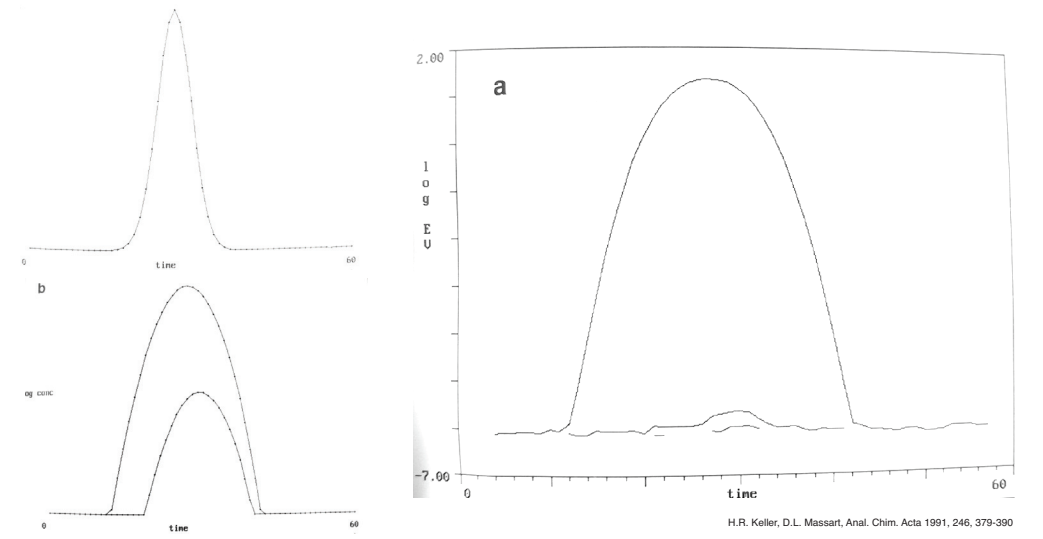
Simuliertes Chromatogramm



Evolvierende Faktoranalyse

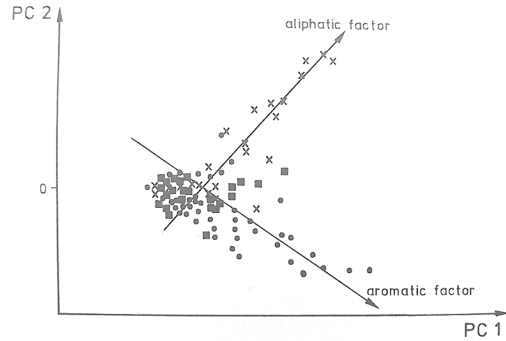
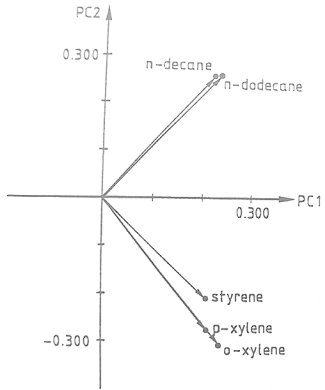
2. $R = 0.2$, 0.5% Verunreinigung

Simuliertes Chromatogramm



Faktoranalyse

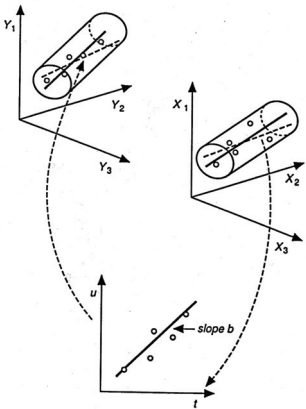
Gelegentlich wird der Begriff Faktoranalyse für die PCA verwendet. In der chemischen Literatur ist die Faktoranalyse eine nicht notwendigerweise orthogonale Rotation der Koordinatenachsen, so dass die neuen Koordinaten chemisch interpretierbar sind.



Partial least squares (PLS)

1. PLS ist eine mit der Hauptkomponentenanalyse verwandte Methode. Es werden orthogonale Linearkombinationen der ursprünglichen Variablen so ausgewählt, dass sie
 - a. ein Maximum der Varianz der X -Matrix beschreiben
 - b. ein Maximum der Korrelation mit y (oder Y) zeigen
2. PLS- und PCA-Hauptkomponenten unterscheiden sich leicht.
3. In vielen Arbeiten wurden die beiden Methoden verglichen. PLS ist in manchen Fällen PCR überlegen, aber nicht immer. Beispiel: Korrelation von Strukturvektoren mit Eigenschaften: PLS war signifikant besser.

Partial least squares (PLS)



PLS ist mit der Hauptkomponentenanalyse verwandt. Neben dem Kriterium, dass eine Hauptkomponente das Maximum der (Rest)-Varianz beschreiben soll, muss sie zusätzlich eine möglichst optimale Korrelation mit den abhängigen Variablen haben.