Chemometrie: von Daten zu Information

Variablen	Eigenschaften
direkt messbar nicht unmittelbar relevant können nicht interpretiert werden	nicht direkt messbar unmittelbar relevant können interpretiert werden
Daten Spektren	**Information** chemische Struktur

Spannung → Absorbanz → chemische Zusammensetzung → Geschmack

Chemometrie: Definition

D.L. Massart et al., Handbook of Chemometrics and Qualimetrics

Chemometrics is a chemical discipline that uses mathematics, statistics, and formal logic to design or select optimal experimental procedures ...

→ Optimierung, Versuchsplanung, Kalibration

... to provide maximum relevant chemical information by analyzing chemical data ...

→ Explorative Datenanalyse

... to obtain knowledge about chemical systems.

Warnungen

Vorsicht mit Zahlen!

Wenn immer möglich: **Graphische Darstellung als Kontrolle**

Auf die Anzahl Freiheitsgrade achten

Vorsicht bei Datentransformationen

Man kann auch mit guten mathematischen Modellen Dummheiten machen!

Literatur dazu:
Use and Abuse of Chemometrics, Trends Anal. Chem. (TrAC) 2006, **25** (11).
darin:
M. Badertscher, E. Pretsch, Bad results from good data, Trends Anal. Chem. (TrAC) 2006, **25** 1131-1138.

Daten von Anscombe: numerisch

x	y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

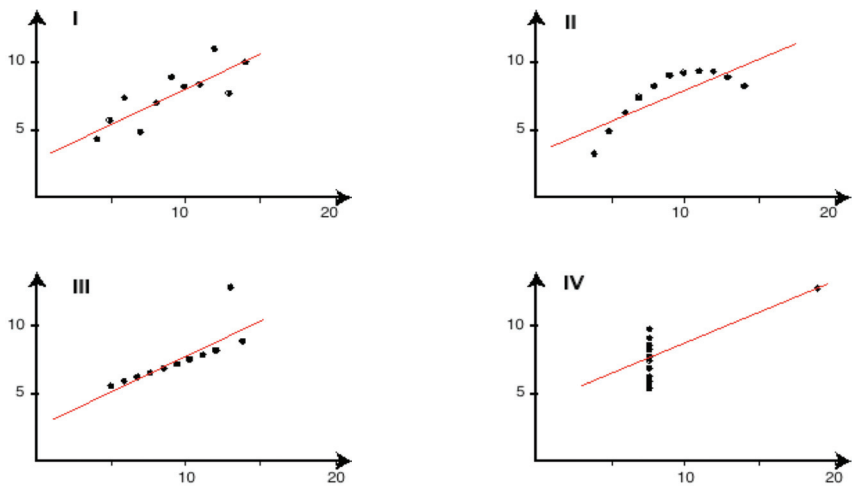
x	y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

x	y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

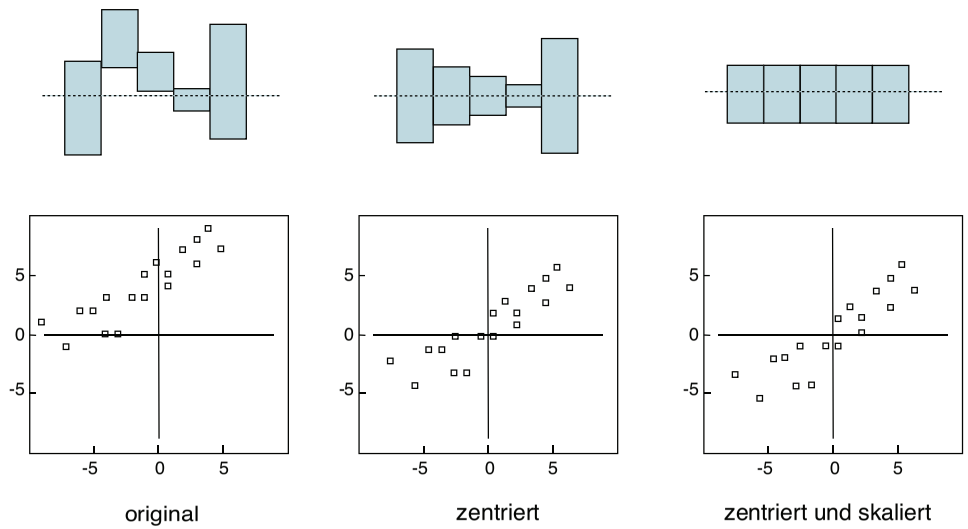
x	y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

für alle Datensätze gilt:
Stichprobenumfang = 11
Mittelwert der x = 9.0
Mittelwert der y = 7.5
Regressionsgerade (y auf x) : $y = 3 + 0.5\,x$
Standardabweichung der Steigung = 0.118
nicht erfasste Quadratsumme = 13.75
Korrelationskoeffizient (zwischen x und y) = 0.82

Daten von Anscombe: graphisch



Zentrieren und Skalieren



Datentransformationen

Zentrierung (Mittelwert = 0 setzen: $x_i' = x_i - \bar{x}$)
Häufig der erste Schritt, da die Variation der Daten wichtiger ist als ihre Absolutwerte

Skalierung bezüglich Standardabweichung s:
(Varianz = 1 setzen: $x_i' = (x_i - \bar{x})/s$)
Notwendig um verschiedenartige Daten zu vergleichen. Vorsicht, wenn Absolutwerte in der gleichen Grössenordnung liegen wie die Fehler.

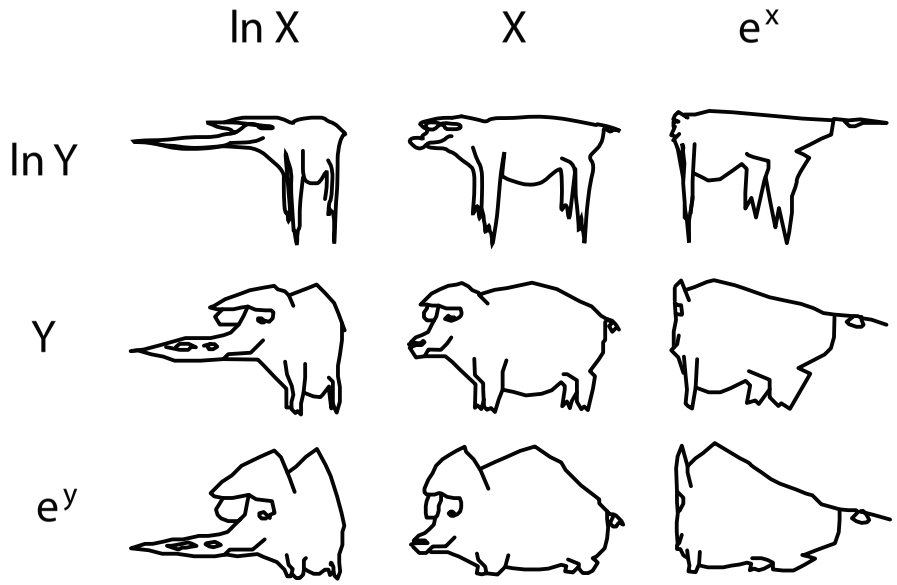
Skalierung bezüglich Bereich: $x_i' = (x_i - \min(x))/(\max(x)-\min(x))$

Logarithmierung

Linearisierung (z.B. Gran-Plot, Scatchard-Plot)
Gut um Zusammenhänge zu erkennen. Die Parameter sollten aber nicht mit den linearisierten Variablen angepasst werden.

Achtung: Fehlerstruktur, Distanzen und Winkel (Korrelationskoeffizient) können durch die Transformation verändert werden.

Transformationen



Scatchard-Plot

Bestimmung der Komplexbildungskonstante (K) zwischen Ligand (L) und Substrat (S). Zur Lösung des Liganden (totale Konzentration: L_{tot}) wird schrittweise Substrat-Lösung zugegeben und die Konzentration des Komplexes (LS) bestimmt. Ein Plot von [LS]/[S] gegen [LS] hat die Steigung –K.

$L + S \rightleftharpoons LS$; $\quad K = \frac{[LS]}{[L]\,[S]}$; $\quad L_{tot} = [L] + [LS]$; $\quad [L] = L_{tot} - [LS]$

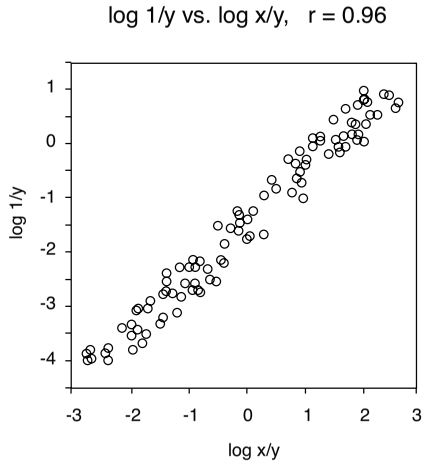
Daraus folgt: $\frac{[LS]}{[S]} = K\,[L] = K\,(L_{tot} - [LS])$

Problem: Die Voraussetzung der Linearen Regression (Fehler nur in y) ist NICHT erfüllt. Zudem kann der Fehler entlang der x-Achse variieren (heteroskedastische Daten). Die Methode sollte für die quantitative Auswertung NICHT eingesetzt werden.

G. Scatchard, Ann. N. Y. Acad. Sci. **1949**, *51*, 660.

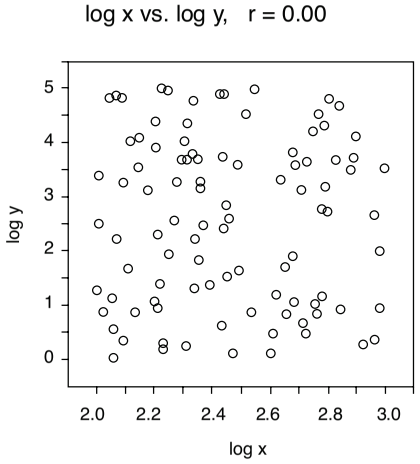
Datentransformation: Warnungen

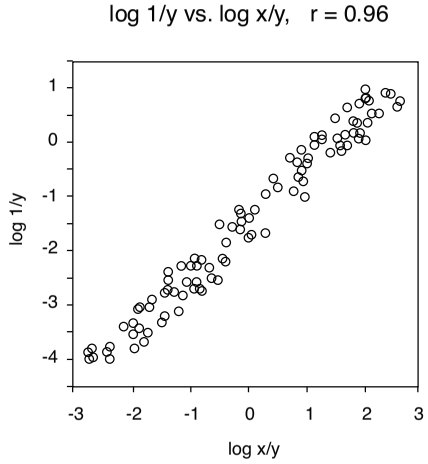
Wie kann man für gute Korrelationen sorgen?



Datentransformation: Warnungen

Wie kann man für gute Korrelationen sorgen?





Charakterisierung von 1D Daten

Mittelwert: μ, Schätzung $m_x = \sum x_i / n$
Varianz σ^2, Schätzung: s^2 = (s: Standardabweichung) $s^2 = \frac{\sum (x_i - m_x)^2}{(n-1)}$
Höhere Momente

Oft wird eine Normalverteilung angenommen. Sie ist durch Mittelwert und Varianz vollständig charakterisiert.

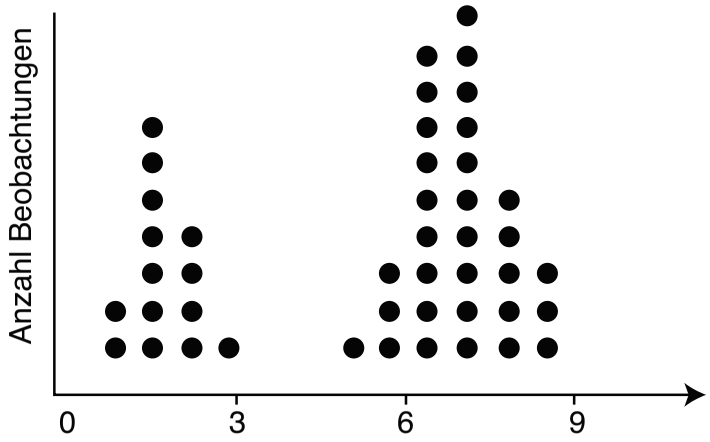
Mittelwert und Varianz können für beliebig verteilte Daten berechnet werden. Sie haben dann aber nicht die gleiche statistische Bedeutung.

Wichtigkeit von Graphiken

Mittelwert = 3.6
s = 4.2

Wichtigkeit von Graphiken

Mittelwert = 3.6
s= 4.2



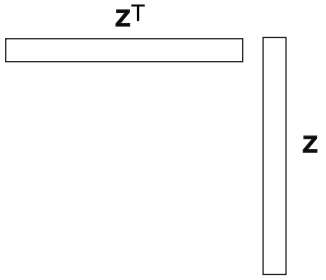
Varianz

$$s^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\sum z_i^2$$

z: zentrierte Variablen

Mit Vektornotation:

$$s^2 = \frac{1}{n-1}\mathbf{z}^T\mathbf{z}$$



$$= \sum z_i^2$$

Berechnung der Varianz: Matrixnotation

Eindimensionale Daten (Datenvektor $\mathbf{x}$)

Mittelwert: $\bar{x} = \frac{1}{n}\Sigma\, x_i = \frac{1}{n}\mathbf{1}^T\mathbf{x}$

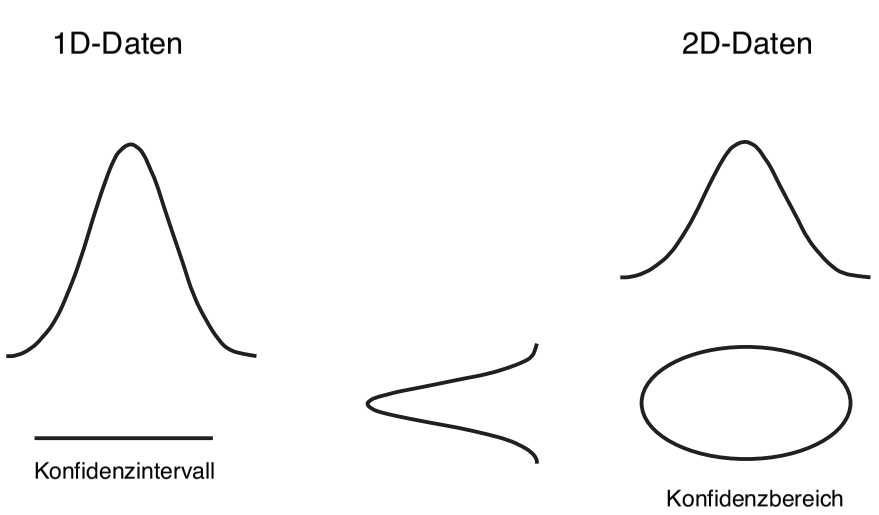
Summe der Fehlerquadrate:

$$\Sigma\ (x_i-\bar{x})^2 = \Sigma\ x_i^2 - \frac{(\Sigma x_i)^2}{n} = \mathbf{x}^T\mathbf{x} - \frac{1}{n}(\mathbf{x}^T\mathbf{1})(\mathbf{1}^T\mathbf{x}) = \mathbf{x}^T\mathbf{H}_n\mathbf{x} \text{ mit } \mathbf{H}_n=\mathbf{I} - \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T\right)$$

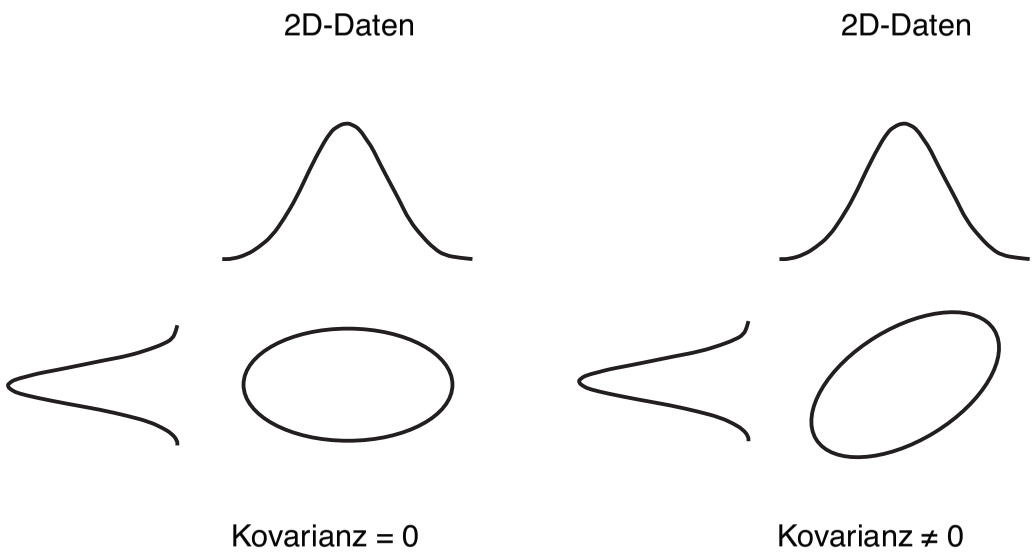
Varianz: $s^2 = \frac{1}{n-1}\mathbf{x}^T\mathbf{H}_n\mathbf{x}$ $\mathbf{H}_n$: Zentrierungsmatrix

Da $\mathbf{H}_n$ symmetrisch und idempotent ist: $(\mathbf{H}_n\mathbf{x})^T\mathbf{H}_n\mathbf{x} = \mathbf{x}^T\mathbf{H}_n^T\mathbf{H}_n\mathbf{x} = \mathbf{x}^T\mathbf{H}_n\mathbf{x}$

Varianz und Kovarianz



Varianz und Kovarianz



Varianz und Kovarianz

1D-Wahrscheinlichkeitsdichtefunktion für die Normalverteilung

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)}$$

σ^2: Varianz

nD-Wahrscheinlichkeitsdichtefunktion für die Normalverteilung

$$f(\mathbf{x}) = \frac{1}{\sqrt{2\pi\Sigma}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

Σ: Varianz-Kovarianzmatrix

Berechnung der Varianz: Matrixnotation

$$s^2 = \frac{1}{n-1}\mathbf{y}^T\mathbf{H}_n\mathbf{y}$$

Für mehrdimensionale Daten erhält man mit der analogen Matrixoperationen die Varianz-Kovarianz-matrix **S**:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A}^T\mathbf{A} - \frac{1}{n}(\mathbf{A}^T\mathbf{1})(\mathbf{1}^T\mathbf{A}) = \mathbf{A}^T\mathbf{H}_n\mathbf{A}$$

Die Diagonalelemente sind die Varianzen, die Ausserdiagonalelemente die Kovarianzen der entsprechenden Variablen.

$$\mathbf{H}_n\mathbf{A} = \begin{bmatrix} x_1-\bar{x} & y_1-\bar{y} & z_1-\bar{z} \\ x_2-\bar{x} & y_2-\bar{y} & z_2-\bar{z} \\ x_3-\bar{x} & y_3-\bar{y} & z_3-\bar{z} \end{bmatrix} \qquad \mathbf{S} = \frac{1}{n-1}\mathbf{A}^T\mathbf{H}_n\mathbf{A} = \frac{1}{n-1}\begin{bmatrix} \Sigma(x_i-\bar{x})^2 & \Sigma(x_i-\bar{x})(y_i-\bar{y}) & \Sigma(x_i-\bar{x})(z_i-\bar{z}) \\ \Sigma(x_i-\bar{x})(y_i-\bar{y}) & \Sigma(y_i-\bar{y})^2 & \Sigma(y_i-\bar{y})(z_i-\bar{z}) \\ \Sigma(x_i-\bar{x})(z_i-\bar{z}) & \Sigma(y_i-\bar{y})(z_i-\bar{z}) & \Sigma(z_i-\bar{z})^2 \end{bmatrix} =$$

$$\begin{bmatrix} \text{var}(x) & \text{cov}(x,y) & \text{cov}(x,z) \\ \text{cov}(x,y) & \text{var}(y) & \text{cov}(y,z) \\ \text{cov}(x,z) & \text{cov}(y,z) & \text{var}(z) \end{bmatrix}$$

Teeproben

Category	Variety	Samples	Source
Green tea	Chunmee	C1, C2, C3, C4, C5, C6, C7	Shanghai Tea Inst.
	Hyson	H1, H2, H3, H4, H5	Shanghai Tea Inst.
Black tea	Keemun	K1, K2, K3, K4	Shanghai Tea Inst.
	Feng Quing	F1, F2, F3, F4, F5, F6, F7	Yunnan Tea Inst.
Oolong tea	Tikuanyin	T1, T2, T3, T4	Xia Men Tea Inst.
	Se Zhong	S1, S2, S3, S4	Xia Men Tea Inst.

High quality
Low quality

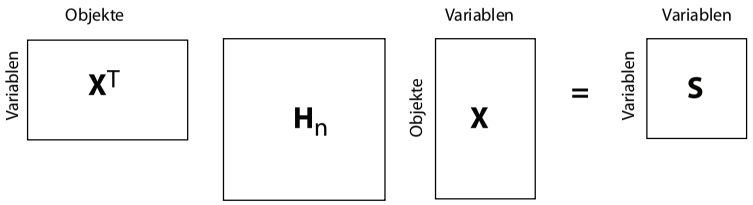
X. Liu, P. van Espen, F. Adams, S.H. Yan, M. Vanbella, Anal.Chim. Acta **1987**, 200, 421

Teeproben: Messdaten

Probe	Cellulose	Hemicellulose	Lignin	Polyphenole	Coffein	Aminosäuren
c1	9.50	4.90	3.53	29.03	4.44	3.82
c2	10.06	5.11	3.57	27.84	4.29	3.70
c3	10.79	5.46	4.62	26.53	3.91	3.46
c4	11.31	4.92	5.02	25.16	3.72	3.29
c5	11.50	6.08	5.48	23.28	3.50	3.10
c6	12.10	5.64	5.61	22.23	3.38	3.02
c7	13.30	5.68	6.32	21.10	3.14	2.87
h1	9.07	5.33	4.42	27.23	4.20	3.18
h2	10.75	5.80	5.29	25.99	4.00	3.00
h3	10.78	5.72	5.79	24.77	3.86	2.91
h4	12.00	6.68	7.20	24.05	3.49	2.81
h5	12.17	5.86	7.71	23.02	3.42	2.60
k1	10.32	10.66	5.07	21.55	4.23	4.43
k2	10.99	10.11	5.60	20.64	4.14	4.35
k3	12.32	10.12	6.53	20.06	4.02	4.12
k4	13.04	7.70	7.70	19.34	3.74	3.45
f1	10.95	7.84	5.22	26.68	5.03	5.32
f2	10.70	7.80	5.82	24.45	4.32	4.72
f3	10.81	8.43	6.00	23.74	4.11	4.50
f4	10.65	8.41	6.40	23.21	3.99	4.28
f5	11.24	8.13	7.61	22.68	3.81	4.09
f6	11.11	8.53	7.97	22.54	3.75	3.97
f7	11.83	9.78	8.67	22.16	3.59	3.88
t1	12.15	12.84	9.95	20.65	3.09	2.97
t2	12.13	12.35	10.55	20.61	2.97	2.49
t3	11.90	15.83	11.18	20.52	2.94	1.90
t4	11.92	15.58	11.87	20.42	2.83	1.79
s1	12.11	14.02	10.99	18.96	2.87	2.80
s2	12.74	14.23	11.16	18.64	2.72	2.23
s3	12.01	14.45	12.08	18.86	2.66	1.84
s4	11.85	14.42	12.60	18.84	2.64	1.76

Teeproben: Varianz-Kovarianzmatrix

	Cellulose	Hemicellulose	Lignin	Polyphenole	Coffein	Aminosäuren
Cellulose	1.06	1.51	1.66	-1.82	-0.35	-0.36
Hemicellulose	1.51	12.39	8.27	-7.51	-1.33	-1.43
Lignin	1.66	8.27	7.01	-5.83	-1.31	-1.60
Polyphenole	-1.82	-7.51	-5.83	8.67	1.35	1.24
Coffein	-0.35	-1.33	-1.31	1.35	0.37	0.49
Aminosäuren	-0.36	-1.43	-1.60	1.24	0.49	0.84



Korrelationskoeffizient

Ein normiertes Mass für den Zusammenhang zwischen zwei Zufallsvariablen ist der Korrelationskoeffizient r mit Werten zwischen -1 und +1.

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$ (σ_{xy} ist die Kovarianz)

Für eine Abschätzung des Korrelationskoeffizienten gilt: $$r = \frac{cov_{xy}}{s_x s_y}$$

Aus der Varianz-Kovarianzmatrix **S** kann man die Korrelationsmatrix wie folgt ableiten. Zuerst erzeugt man eine Diagonalmatrix **D**, die die jeweiligen reziproken Standardabweichungen enthält (vgl. Gleichung 6.20 im Skript). Die Korrelationsmatrix ist dann:

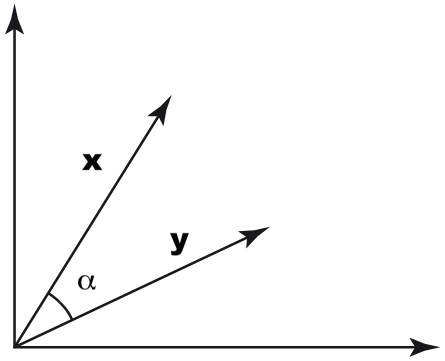
R = D S D

Die Diagonalelemente der Korrelationsmatrix haben den Wert 1, die Ausserdiagonalelmente sind die Korrelationskoeffizienten der entsprechenden Paare von Variablen.

Teeproben: Korrelationsmatrix

	Cellulose	Hemicellulose	Lignin	Polyphenole	Coffein	Aminosäuren
Cellulose	1.00	0.42	0.61	-0.60	-0.56	-0.38
Hemicellulose	0.42	1.00	0.89	-0.72	-0.62	-0.44
Lignin	0.61	0.89	1.00	-0.75	-0.82	-0.66
Polyphenole	-0.60	-0.72	-0.75	1.00	0.76	0.46
Coffein	-0.56	-0.62	-0.82	0.76	1.00	0.88
Aminosäuren	-0.38	-0.44	-0.66	0.46	0.88	1.00

Winkel zwischen zwei Vektoren



Skalarprodukt:

$\mathbf{x}^T\mathbf{y} = ||x||\ ||y||\ \cos\alpha$

Länge eines Vektors:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2} = \sqrt{\sum x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

$$\cos\alpha = \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{\mathbf{x}^T\mathbf{x}}\sqrt{\mathbf{y}^T\mathbf{y}}}$$

Korrelationskoeffizient

$$r = \frac{cov(xy)}{s_x s_y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}}$$

Für zentrierte Variablen: $r = \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{\mathbf{x}^T\mathbf{x}}\sqrt{\mathbf{y}^T\mathbf{y}}}$

$$\cos\alpha = \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{\mathbf{x}^T\mathbf{x}}\sqrt{\mathbf{y}^T\mathbf{y}}}$$

Der Korrelationskoeffizient entspricht dem Cosinus des Winkels zwischen den beiden Vektoren

Objektraum und Variablenraum

Multivariate Daten:

p Variablen

n Objekte

Datenmatrix

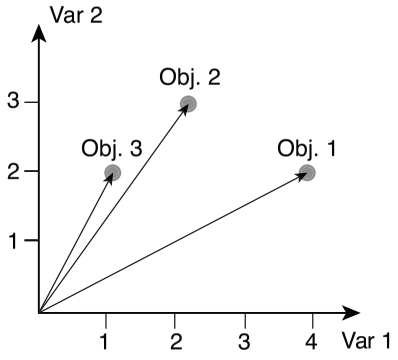
Geometrische Repräsentation

	Variable 1	Variable 2
Objekt 1	4	2
Objekt 2	2	3
Objekt 3	1	2

Variablenraum

Jede Variable spannt eine Dimension auf. Jedes Objekt ist ein Punkt im p-dimensionalen Raum (hier ist p=2).

So lassen sich Objekte vergleichen (z.B. über ihren Abstand).



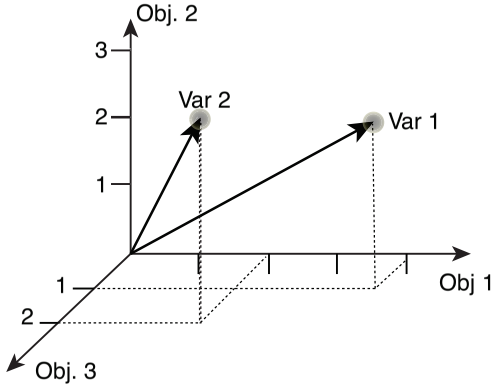
Geometrische Repräsentation

	Variable 1	Variable 2
Objekt 1	4	2
Objekt 2	2	3
Objekt 3	1	2

Objektraum

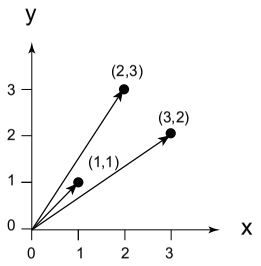
Jedes Objekt spannt eine Dimension auf. Jede Variable ist ein Punkt im n-dimensionalen Raum (hier ist n=3).

So lassen sich Variablen vergleichen (z.B. über den Winkel zwischen ihnen).

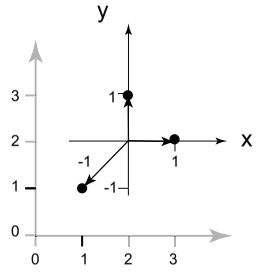


Bedeutung des Korrelationskoeffizienten

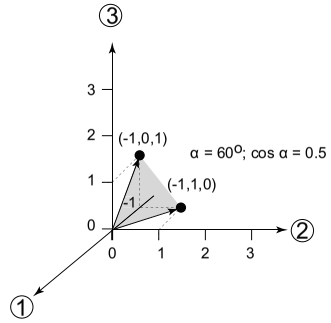
Variablenraum



Variablenraum zentriert



Objektraum (zentrierte Variablen)



$$r = \frac{\begin{bmatrix}-1 & 0 & 1\end{bmatrix}\begin{bmatrix}-1\\1\\0\end{bmatrix}}{\sqrt{\begin{bmatrix}-1 & 0 & 1\end{bmatrix}\begin{bmatrix}-1\\0\\1\end{bmatrix}}\sqrt{\begin{bmatrix}-1 & 1 & 0\end{bmatrix}\begin{bmatrix}-1\\1\\0\end{bmatrix}}} = \frac{1}{\sqrt{2}\ \sqrt{2}} = 0.5$$

Teeproben: Korrelationsmatrix

	Cellulose	Hemicellulose	Lignin	Polyphenole	Coffein	Aminosäuren
Cellulose	1.00	0.42	0.61	-0.60	-0.56	-0.38
Hemicellulose	0.42	1.00	0.89	-0.72	-0.62	-0.44
Lignin	0.61	0.89	1.00	-0.75	-0.82	-0.66
Polyphenole	-0.60	-0.72	-0.75	1.00	0.76	0.46
Coffein	-0.56	-0.62	-0.82	0.76	1.00	0.88
Aminosäuren	-0.38	-0.44	-0.66	0.46	0.88	1.00

Geometrische Interpretation:
Im 31-dimensionalen Objektraum, ist der Winkel zwischen: Cellulose und Aminosäuren: 112°
Coffein und Aminosäuren: 28°
Coffein und Polyphenolen: 41°

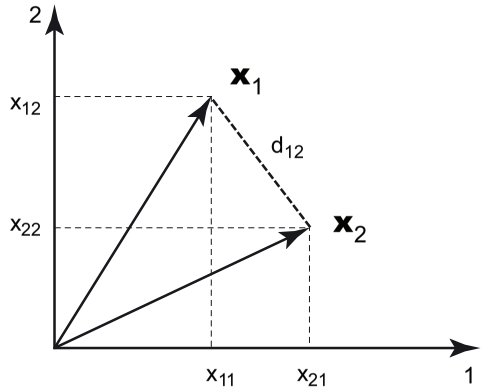
Teeproben

Category	Variety	Samples	Source
Green tea	Chunmee	C1, C2, C3, C4, C5, C6, C7	Shanghai Tea Inst.
	Hyson	H1, H2, H3, H4, H5	Shanghai Tea Inst.
Black tea	Keemun	K1, K2, K3, K4	Shanghai Tea Inst.
	Feng Quing	F1, F2, F3, F4, F5, F6, F7	Yunnan Tea Inst.
Oolong tea	Tikuanyin	T1, T2, T3, T4	Xia Men Tea Inst.
	Se Zhong	S1, S2, S3, S4	Xia Men Tea Inst.

High quality
Low quality

X. Liu, P. van Espen, F. Adams, S.H. Yan, M. Vanbella, Anal.Chim. Acta **1987**, 200, 421

Distanz zwischen zwei Vektoren



Euklid: $$d_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)}$$

Distanz zwischen Teeproben

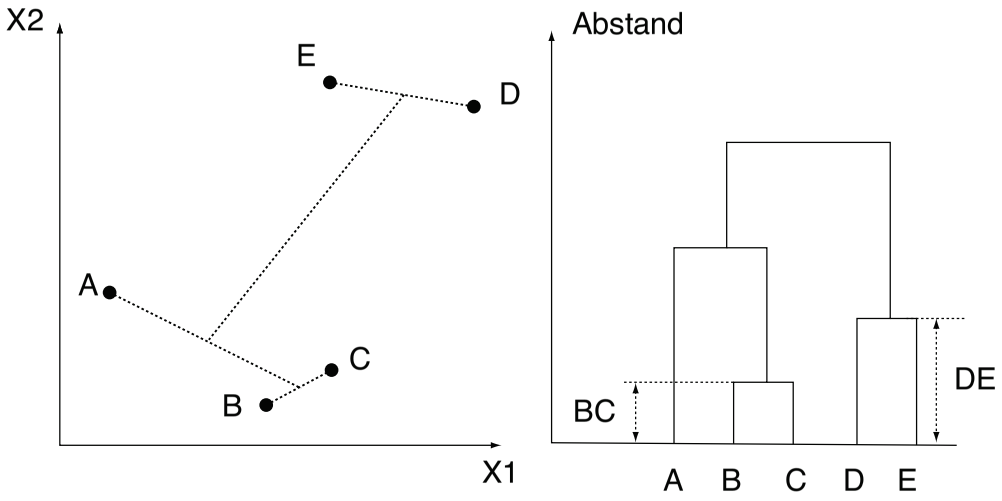
	c1	c2	c3	c4	c5	c6	c7	h1	h2	h3	h4	h5	k1	k2	k3	k4
c2	1.35															
c3	3.13	1.92														
c4	4.61	3.37	1.63													
c5	6.61	5.33	3.53	2.28												
c6	7.72	6.43	4.66	3.20	1.30											
c7	9.40	8.11	6.30	4.82	3.01	1.82										
h1	2.21	1.55	1.92	3.17	4.87	6.03	7.77									
h2	3.95	2.81	1.04	1.42	2.88	4.06	5.68	2.33								
h3	5.17	4.01	2.20	1.35	1.77	2.91	4.54	3.35	1.33							
h4	7.04	5.93	4.04	3.14	2.06	2.64	3.51	5.37	3.17	2.25						
h5	8.01	6.88	4.99	3.74	2.41	2.29	2.66	6.28	4.15	3.00	1.44					
k1	9.62	8.56	7.30	6.97	5.28	5.64	6.25	8.01	6.76	6.16	5.71	6.29				
k2	10.22	9.07	7.63	7.01	5.05	5.11	5.37	8.53	7.01	6.21	5.46	5.76	1.37			
k3	11.18	9.99	8.37	7.56	5.46	5.22	4.91	9.50	7.69	6.78	5.50	5.57	2.96	1.74		
k4	11.50	10.26	8.44	7.20	5.06	4.26	3.14	9.73	7.69	6.52	5.00	4.30	5.40	4.13	2.96	
f1	4.66	3.94	3.28	4.10	4.72	5.84	7.22	4.01	3.34	3.96	4.69	5.92	6.01	6.60	7.42	8.34
f2	6.07	5.02	3.61	3.50	2.89	3.92	5.25	4.56	3.12	2.81	3.06	4.10	4.17	4.49	5.31	6.09
f3	6.99	5.91	4.43	4.14	2.97	3.82	4.94	5.39	3.84	3.32	3.04	4.00	3.30	3.56	4.37	5.38
f4	7.49	6.41	4.86	4.40	2.94	3.65	4.65	5.77	4.17	3.46	2.93	3.70	3.13	3.21	3.96	4.86
f5	8.42	7.34	5.61	4.88	3.20	3.53	4.04	6.66	4.82	3.89	2.54	2.92	3.91	3.52	3.63	3.87
f6	8.84	7.79	6.08	5.39	3.70	4.00	4.39	7.06	5.26	4.34	2.92	3.25	3.87	3.47	3.51	3.87
f7	10.18	9.15	7.45	6.82	5.08	5.23	5.18	8.49	6.64	5.81	4.07	4.34	4.14	3.62	3.10	3.86
t1	13.57	12.59	10.96	10.42	8.56	8.56	8.12	11.88	10.13	9.35	7.57	7.72	6.01	5.54	4.65	5.89
t2	13.66	12.68	11.01	10.41	8.56	8.51	8.01	11.92	10.14	9.31	7.46	7.50	6.51	5.97	5.02	5.81
t3	16.19	15.33	13.83	13.47	11.71	11.80	11.40	14.57	12.99	12.35	10.64	10.89	8.70	8.49	7.79	9.16
t4	16.44	15.58	14.05	13.65	11.90	11.96	11.51	14.79	13.20	12.52	10.76	10.94	9.12	8.86	8.09	9.25
s1	15.83	14.84	13.21	12.59	10.63	10.50	9.87	14.08	12.35	11.51	9.72	9.70	7.79	7.24	6.28	7.28
s2	16.41	15.40	13.74	13.08	11.10	10.91	10.17	14.66	12.87	12.02	10.17	10.10	8.44	7.84	6.77	7.60
s3	16.77	15.82	14.18	13.57	11.65	11.52	10.87	14.99	13.29	12.44	10.62	10.55	9.10	8.57	7.62	8.35
s4	17.02	16.09	14.45	13.84	11.94	11.82	11.16	15.24	13.54	12.70	10.87	10.77	9.50	8.97	8.03	8.65

Distanz zwischen Teeproben

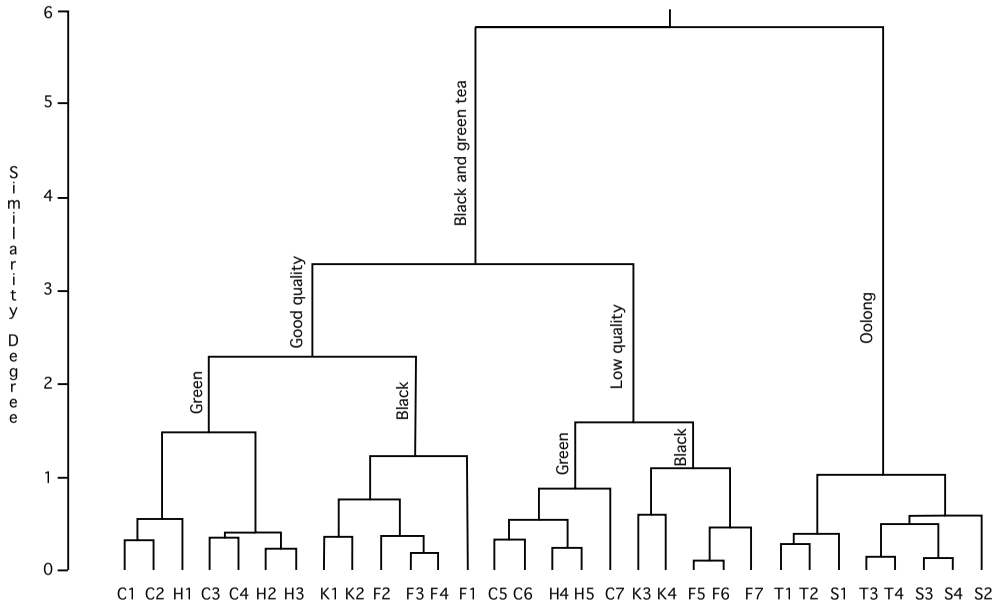
	f1	f2	f3	f4	f5	f6	f7	t1	t2	t3	t4	s1	s2	s3
c2														
c3														
c4														
c5														
c6														
c7														
h1														
h2														
h3														
h4														
h5														
k1														
k2														
k3														
k4														
f1														
f2	2.50													
f3	3.34	1.02												
f4	4.00	1.60	0.73											
f5	4.99	2.72	2.06	1.50										
f6	5.35	3.14	2.41	1.81	0.59									
f7	6.40	4.45	3.63	3.14	2.13	1.66								
t1	9.72	7.97	7.06	6.60	5.86	5.35	3.80							
t2	9.96	8.16	7.28	6.78	5.90	5.37	3.87	0.92						
t3	12.42	10.95	10.06	9.67	9.09	8.56	7.07	3.42	3.59					
t4	12.71	11.21	10.34	9.91	9.25	8.70	7.20	3.57	3.57	0.76				
s1	11.98	10.17	9.23	8.73	7.95	7.43	5.95	2.33	2.41	2.57	2.53			
s2	12.60	10.80	9.85	9.36	8.55	8.04	6.53	2.91	2.88	2.64	2.52	0.96		
s3	13.07	11.29	10.37	9.86	9.03	8.49	7.03	3.44	3.22	2.36	1.95	1.54	1.28	
s4	13.36	11.58	10.69	10.16	9.29	8.75	7.31	3.81	3.51	2.63	2.10	1.99	1.78	0.55

Clusteranalyse

Prinzip der hierarchischen Clusteranalyse: Avarage-Linkage-Methode:
links: Koordinaten im 2D-Raum, rechts: Dendrogramm

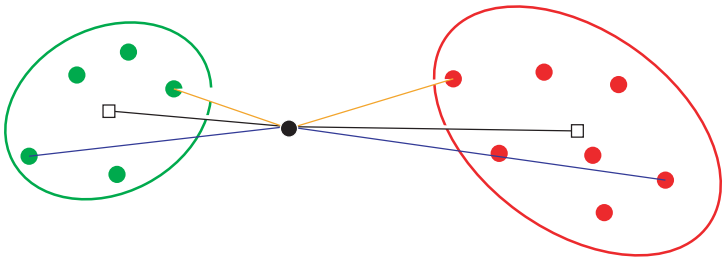


Teeproben: Clustering



Clusterungsmethoden

"Single linkage" (nächster Nachbar): kann zu elongierten Clustern führen
"Average linkage" (Mitte der Cluster)
"Complete linkage" (am weitesten entfernter Mitglied der Cluster): führt zu kompakten Clustern



Distanzmasse

Minkowski $$d_{ij} = \left(\sum_k \left(x_{ik} - x_{jk} \right)^n \right)^{1/n}$$

Euklid $$d_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2} = \sqrt{\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \left(\mathbf{x}_i - \mathbf{x}_j\right)}$$

City block $$d_{ij} = \sum_k \left| x_{ik} - x_{jk} \right|$$

Distanzmasse

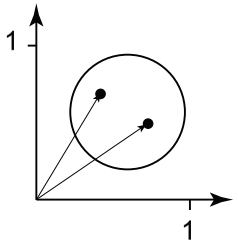
Euklid $$d_{ij} = \sqrt{\sum_k (x_{ik} - x_{jk})^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)}$$

Euklid gewichtet $$d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} (\mathbf{x}_i - \mathbf{x}_j)}$$

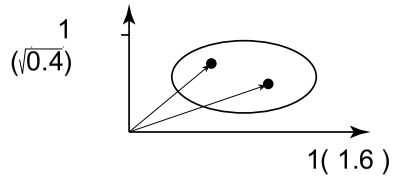
Gewichtsmatrix: **I:** Euklid
diag W: individuell gewichtet,
z.B $w_i = 1/s_i^2$ Kolonnenvarianz
Mahalanobis Distanz:
$\mathbf{W} = \mathbf{C}^{-1}$ (Inverse der Varianz-Kovarianz Matrix)
$\mathbf{C} = (1/n)\, \mathbf{X}_z^T \mathbf{X}_z$, mit $\mathbf{X}_z$: zentrierte Daten

Gewichtung = Skalierung der Koordinaten

Durch die Gewichtung ändern sich die Distanzen und die Winkel (Korrelationskoeffizienten) zwischen den Vektoren



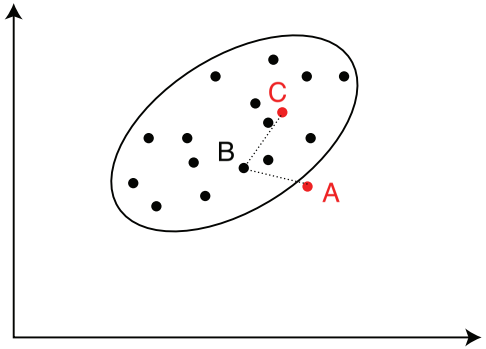
$$\mathbf{W} = 0.5 \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}$$



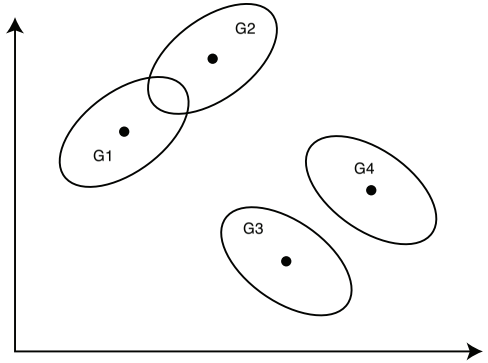
$$\mathbf{W} = 0.5 \begin{vmatrix} 1.6 & 0 \\ 0 & 0.4 \end{vmatrix}$$

Mahalanobis-Distanz

Abstand eines Punktes von einem andren, der sich in einer bekannten Verteilung befindet. Das Abstandsmass berücksichtigt die Korrelation.



Die Mahalanobis-Distanz von C zu B ist kleiner als von A zu B.
Die euklidschen Distanzen sind gleich.



Die Mahalanobis-Distanz zwischen den beiden Clustern G1 und G2 ist keiner als zwischen G3 und G4.
Die euklidschen Distanzen sind gleich.

Distanzmasse für binäre Variablen

Hamming-Distanz: $$d_{ij} = \sum_k \mathrm{XOR}(x_{ik}, x_{jk})$$

Die Hamming-Distanz ist die Cityblock-Distanz für binäre Variablen
XOR: exklusives OR: 0 XOR 0 = 0; 1 XOR 1 = 0
0 XOR 1 = 1; 1 XOR 0 = 1

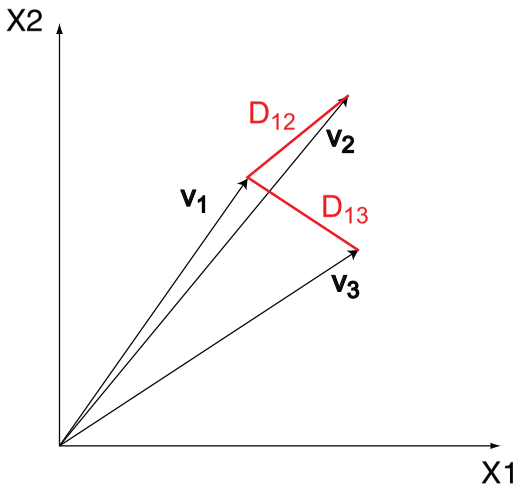
Tanimoto-Koeffizient T zwischen zwei binären Vektoren A und B:
Anzahl der 1 in A: N_A, in B: N_B, gleichzeitiges Vorkommen in A und B: $N_{A\&B}$ (Tanimoto Abstand: 1–T)

$$T = \frac{N_{A\&B}}{N_A + N_B - N_{A\&B}}$$

Beispiel:
A: 1 1 0 1 1 0 0 1 0 1 0 0 1 1
B: 1 1 0 1 0 0 0 1 1 0 0 0 1 1
A&B 1 1 0 1 0 0 0 1 0 0 0 0 1 1
T = 6 / (8 + 7 – 6) = 0.667; 1 – T = 0.333

Hamming-Distanz: 3

Distanzmasse: Abstand und Winkel



Die Abstände D_{12} und D_{13} sind gleich, aber bei der Benützung des Winkels zwischen den Vektoren als Distanzmass ist v_2 viel ähnlicher zu v_1 als v_3.

Der Cosinus des Winkels zwischen zwei Datenvektoren entspricht dem Korrelationskoeffizienten.

Abstand und Winkel: Ein Beispiel

Retentionsindizes von fünf Substanzen (1–5) mit drei stationären Phasen (SF1–SF3) in der Gaschromatographie

Stationäre Phase	1	2	3	4	5
SF1	100	130	150	160	170
SF2	120	110	170	150	145
SF3	190	260	310	320	350

D12 = 43.9 D13 = 329.5
r12 = 0.658 r13 = 0.997

Interpretation: Die absoluten Retentionsindizes sind für SF1 und SF2 ähnlich, die relativen für SF1 und SF3. Die Phasen SF1 und SF2 sind ähnlich polar. SF1 und SF3 zeigen ähnliche spezifische Wechselwirkungen mit den Proben. Man würde bei Erhöhung der Temperatur für SF3 ähnliche Werte bekommen wie für SF1.