8. 6. Die Regressionsanalyse

(vgl. Hartung & Elpelt 1999, S. 77 ff.)

Ziel:
Quantifizierung eines (vermuteten) linearen Zusammenhangs zwischen einem Attribut y (Regressand; endogene Variable; abhängige Variable) und *h* anderen Attributen x_j ($j = 1, ..., h$) (Regressoren; exogene Variablen; unabhängige Variablen)
→ Prognose von y

Voraussetzung:
alle Attribute sind numerische Größen (im Folgenden zunächst: reelle Zahlen).

$h+1$ Attribute $x_1, x_2, \ldots, x_h, y$

„Regressionsfunktion":

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_h x_h$$

$\beta_0, \beta_1, \ldots, \beta_h$: <u>Regressionskoeffizienten</u> (unbekannt)
— werden anhand einer Lernstichprobe geschätzt

Lernstichprobe aus n Instanzen $(n > h+1)$:

$$(x_{i1}, x_{i2}, \ldots, x_{ih}, y_i) \qquad (i = 1, \ldots, n)$$

Regressionsmodell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_h x_{ih} + e_i \qquad (i = 1, \ldots, n)$$

mit stochastisch unabhängigen Zufallsvariablen $e_1, \ldots, e_n$,
Erwartungswert $E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_h x_{ih}$
Varianz $Var(y_i) = Var(e_i) = \sigma^2$ (= unbekannt)

Matrixschreibweise des multiplen Regressionsmodells:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$

Dann soll gelten: ← X heißt „Designmatrix".

$$y = X\beta + e$$

$$E(y) = X\beta, \quad E(e) = 0,$$

Kovarianzmatrix $Cov(y) = Cov(e) = \sigma^2 \cdot I_n$

(I_n = $n \times n$ - Einheitsmatrix $\begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$.)

Es wird vorausgesetzt: $X^T \cdot X$ ist invertierbar. (zunächst; aber s. Bemerkung unten)

Schätzung der β_j durch „kleinste-Quadrate-Schätzer" $\hat{\beta}_j$:

Mit der geschätzten Regressionsfunktion

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$$

soll die Fehlerquadratsumme

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$$

minimal werden (bzgl. $\beta_0, \ldots, \beta_k$).

$\Rightarrow$ Extremalproblem in den $k+1$ Variablen $\beta_0, \ldots, \beta_k$.

Dieses kann gelöst werden (partielle Ableitungen $= 0$ setzen ...) und erweist sich als äquivalent zum Normalengleichungssystem:

$$X^T X \hat{\beta} = X^T y$$

$\Rightarrow$ eindeutige Lösung $\hat{\beta} = (X^T X)^{-1} X^T y$.

Bemerkung:

- Für eine beliebige Matrix X existiert die Matrix

$$X^+ = \lim_{\alpha \to 0} (X^T X + \alpha^2 I)^{-1} X^T .$$

X^+ heißt die <u>Pseudoinverse</u> von X.

- Die Pseudoinverse wird eindeutig charakterisiert durch die folgenden Eigenschaften („Penrose-Axiome"):

(1.) $(X^+ X)^T = X^+ X$
(2.) $(X X^+)^T = X X^+$
(3.) $X^+ X X^+ = X^+$
(4.) $X X^+ X = X$

- Es gilt: $X^+ X X^T = X^T$
  $X^T X X^+ = X^T$
- Falls $X^T X$ invertierbar ist, so gilt
  $X^+ = (X^T X)^{-1} X^T$.
- Falls sogar X invertierbar ist, so gilt
  $X^+ = X^{-1}$.
- In jedem Fall ist $\hat{\beta} = X^+ y$ Lösung der Minimierungsaufgabe $\| y - X\beta \|^2 \to \min$.

Schätzwert für die Varianz:

$$s^2 = \frac{1}{n-k-1} y^T (I_n - X(X^T X)^{-1} X^T) y .$$

Für die Erwartungswerte, Varianzen und Kovarianzen der $\hat{\beta}_j$ gilt:

$$E(\hat{\beta}_j) = \beta_j \qquad \text{(erwartungstreuer Schätzer)}$$

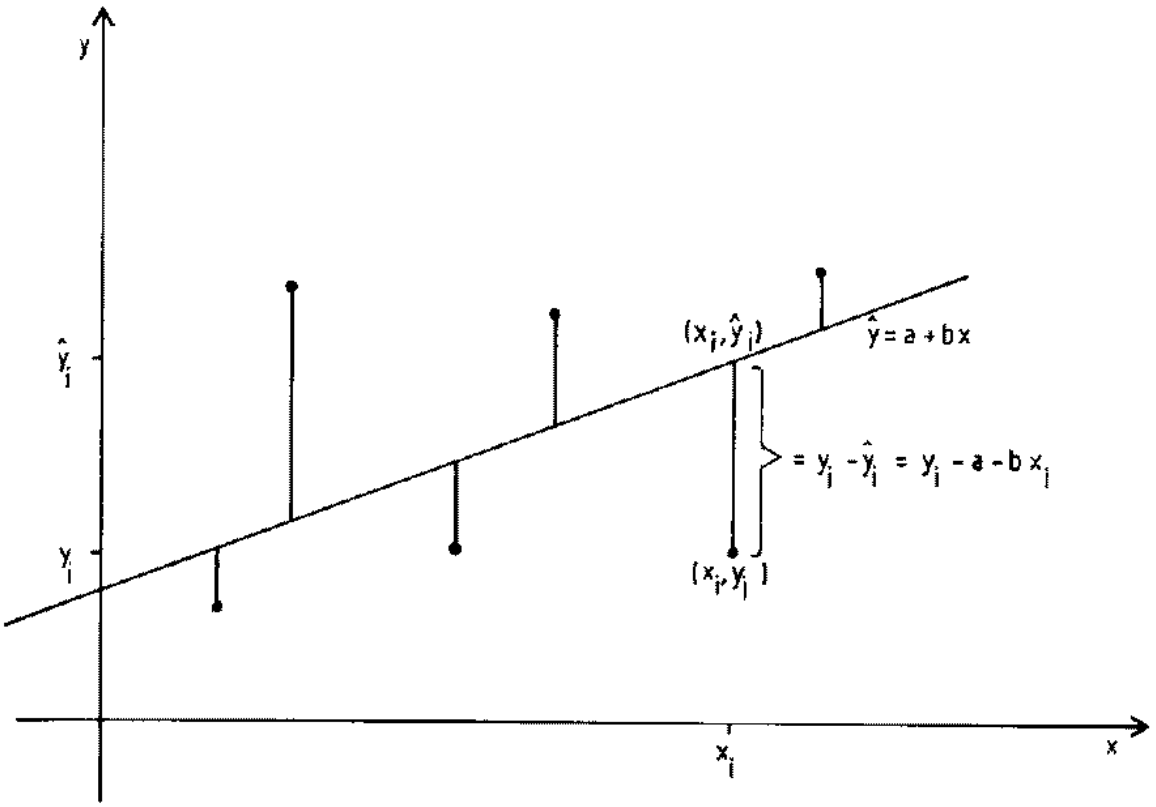
$$Var(\hat{\beta}_j) = \sigma^2 c_{jj}$$

$$Cov(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 c_{jk} \,,$$

wobei $$C = (X^T X)^{-1} = \begin{pmatrix} c_{00} & \cdots & c_{0k} \\ \vdots & & \vdots \\ c_{0k} & \cdots & c_{kk} \end{pmatrix}$$ (symmetr. Matrix).

Der Fall zweier Merkmale (= univariater Fall: 1 abhängige Variable):

x : Regressor, y : Regressand



Graphische Veranschaulichung der Methode der kleinsten Quadrate

Regressionsgerade $\hat{y} = a + b \cdot x$

Beachte: Diese Gerade ist i. allg. verschieden von der "umgekehrten" Regressionsgeraden für x aus y: $\hat{x} = c + d \cdot y$.

Elementare Formeln:

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$ = Steigung der Regressionsgeraden,

$$a = \frac{\sum y - b\sum x}{n}$$ = Achsenabschnitt der Regressionsgeraden auf der y-Achse,

ferner gilt auch $$b = \frac{s_{xy}}{s_x^2} = r \cdot \frac{s_y}{s_x}$$

und $$a = \bar{y} - b \cdot \bar{x}$$,

wobei $s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1}$ die empirische Covarianz der x_i, y_i und s_x^2 die empirische Varianz der x_i ist. (Alle Summen mit Summationsindex $i = 1, ..., n$).

Das Bestimmtheitsmaß ist

$$r^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2}$$

= Verhältnis der Varianz der geschätzten Werte $\hat{y}_i$ zur Varianz der beobachteten Werte y_i

= Anteil an der Varianz von y, der durch die Regression erklärt werden kann.

Beachte: Die Regressionsgerade $\hat{y} = a + b \cdot x$ läuft stets durch den Schwerpunkt $(\bar{x}, \bar{y})$ der Punktwolke.

Für die einzelnen Punkte gilt:

$$y_i = a + b \cdot x_i + \varepsilon_i \quad \text{mit } \textit{Residuen} \quad \varepsilon_i = y_i - \hat{y}_i$$

Residuenanalyse:

Durch Scatterplot der ε_i überprüfen, dass kein systematisches Muster und keine Abhängigkeit der Varianz der ε_i von x vorliegt (*Homoskedastizität*, d.h. konstante Breite des Varianzbandes).

Sonst ist ggf. durch eine geeignete *Transformation* der y-Werte (z.B. Übergang von y zu $\log y$) dafür zu sorgen, dass diese Voraussetzungen erfüllt werden.

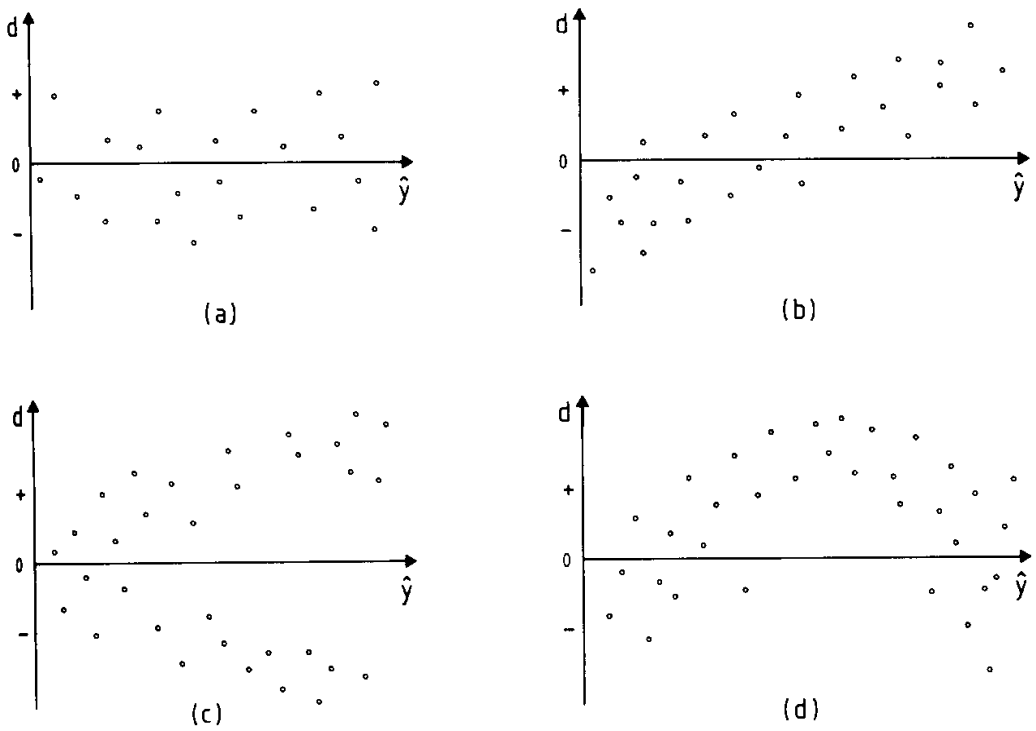
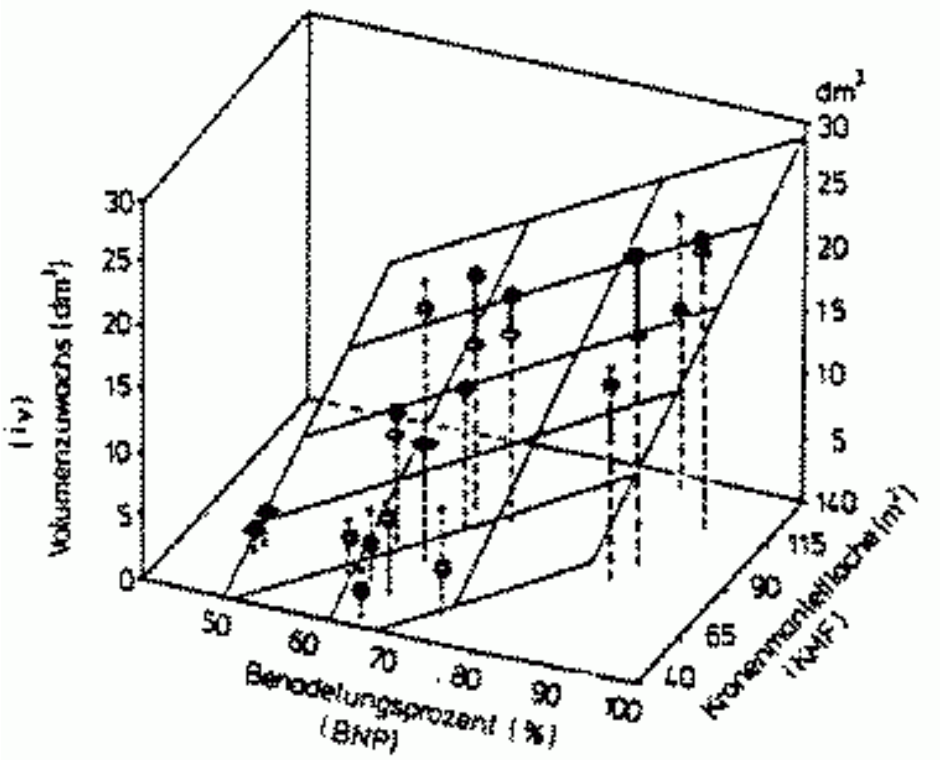


Abb. 12: Graphische Darstellung der normierten Residuen in Abhängigkeit von den geschätzten Werten ŷ des Regressanden: (a) idealer Verlauf; (b) linearer Trend, auf einen Rechenfehler hindeutend; (c) ansteigende Varianzen (mit ŷ), Wechsel zu einem Modell für ungleiche Varianz der Beobachtungen eventuell angebracht; (d) nichtlinearer Verlauf der Residuen, inadäquates Modell, d. h. Transformation der Daten oder Änderung der Regressionsfunktion angezeigt

Abbildung zur Residuenanalyse

Der bivariate Fall (2 unabhängige Variablen, 1 abhängige Variable), Beispiel:



Quasilineare Regression

Es werden Funktionen auf (einige oder alle) Variablen angewandt, bevor die lineare Regression durchgeführt wird.

Beispiel:

$$\ln y_i = a + b_1 \ln x_i + b_2 x_i + \ln \varepsilon_i$$ bzw.

$$\ln \vec{y} = X \cdot \vec{\beta} + \ln \vec{\varepsilon}$$ mit

$$\vec{\beta} = \begin{pmatrix} a \\ b_1 \\ b_2 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & \ln x_1 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & \ln x_n & x_n \end{pmatrix}.$$

Lösungsformel: $\hat{\vec{\beta}} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot \ln \vec{y}$.

(Oft statt $\ln x$ auch x^2, e^x etc.)

Anderes Beispiel:

$y = a + bx + cx^2$ (quadratische Regression).

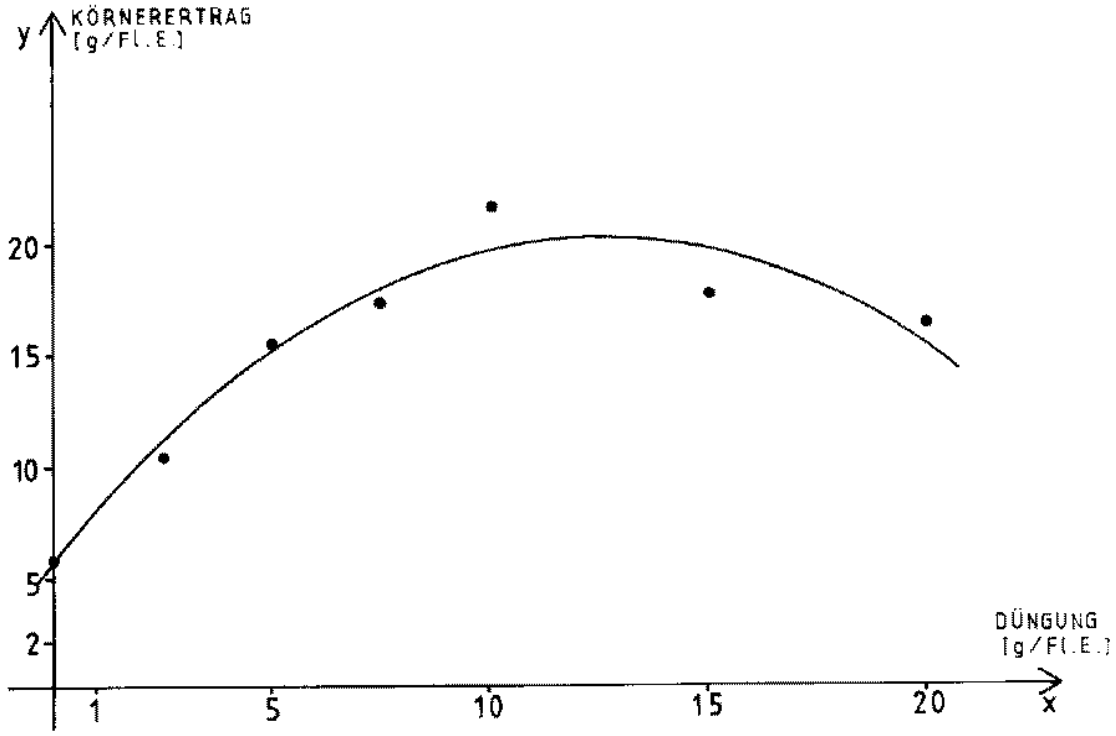
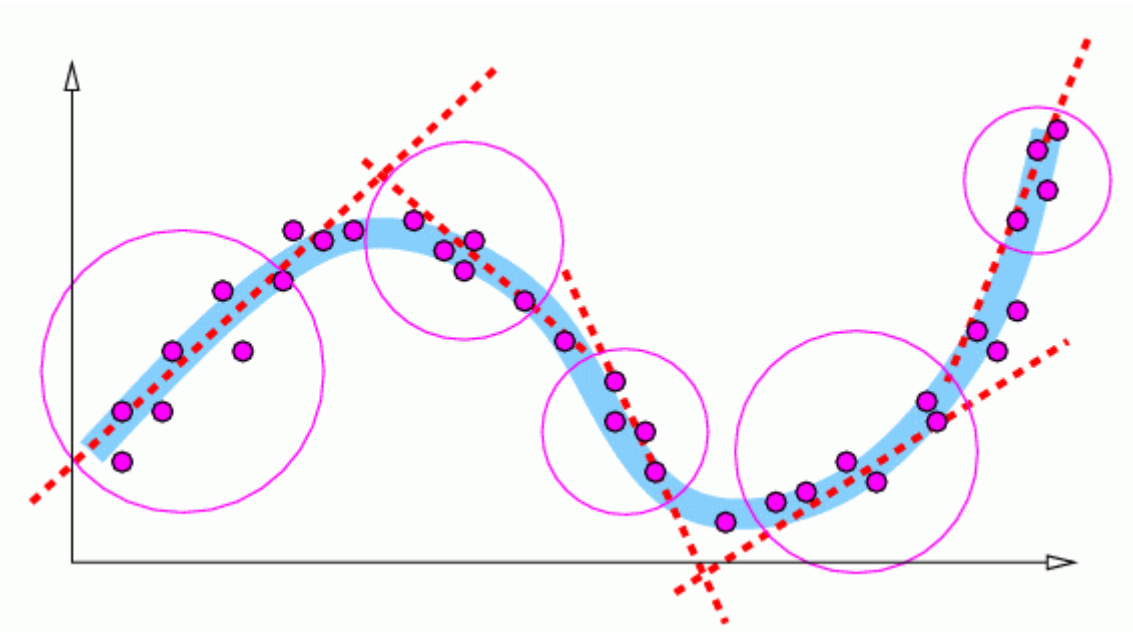


Abb. 6: Quadratische Regressionsfunktion für den Zusammenhang zwischen Körnerertrag y und Düngung x: $y = 5{,}778 + 2{,}299\,x - 0{,}091\,x^2$

Regressionsanalyse als Gegenstück zur Varianzanalyse:

Varianzanalyse:	1 abhängige Variable (Effektvariable), *k* Einflußgrößen (i. allg. nominal skaliert; qualitativ). Jede Einflußgröße liegt in diskreten "Faktorstufen" vor.
Regressionsanalyse:	1 abhängige Variable (Regressand), *k* unabhängige Variablen (metrisch skaliert; quantitativ). Jede unabh. Variable kann sich stetig ändern.
Kombination beider Situationen:	die Einflußfaktoren sind gemischt (mixed), d.h. es gibt qualitative und quantitative Faktoren • **Kovarianzanalyse (*ANCOVA*)**

Lokale lineare Regression:



Nichtlineare Regression

Gehen die Parameter nichtlinear in die Regressionsfunktion ein (Beisp.: $y = a \cdot e^{bx}$) oder wird ein anderes Fehlermaß als die Summe der Fehlerquadrate verwendet, lassen sich die Parameter im Allgemeinen nicht mehr direkt aus den partiellen Ableitungen bzw. mit dem obigen Normalengleichungs-Ansatz berechnen.

Man hat ein *nichtlineares Optimierungsproblem* zu lösen (Fehler $\rightarrow$ min.).

Lösungsansätze:

Gradientenverfahren: Dazu müssen die partiellen Ableitungen der Fehlerfunktion (in Abhängigkeit von den Parametern) existieren und berechnet werden. Man folgt dem maximalen Gradienten, bis ein lokales Minimum des Fehlers erreicht ist.

Es gibt verschiedene Sub-Varianten von Gradientenverfahren, die z.T. in Statistik-Softwarepaketen angeboten werden.

"Direkte Methoden":
Unter *direkten Methoden* (Methoden nullter Ordnung) versteht man in der Optimierung Verfahren, die nur Werte der Zielfunktion (hier der Fehlerfunktion) selbst, aber nicht deren (partielle) Ableitungen benutzen.

Zufallssuche (*random search*): Es werden viele verschiedene Parameterkombinationen zufällig erzeugt und die beste gewählt.

Hillclimbing: Es wird eine zufällige Anfangs-Parameterkombination erzeugt und diese jeweils zufällig geringfügig verändert. Ist die neue Konfiguration besser als die vorhergehende, wird diese weiter verändert, ansonsten die ursprüngliche.

Simulated Annealing (simuliertes Ausglühen): Wie Hill-climbing, allerdings wird eine Verschlechterung der bisher gefundenen Lösung mit einer gewissen Wahrscheinlichkeit in Kauf genommen, um aus nicht-optimalen lokalen Minima wieder herauszukommen. Die Akzeptanzwahrscheinlichkeit ist umso kleiner, je größer die Verschlechterung ist. Generell wird sie während des Verfahrens immer weiter abgesenkt.

Threshold Accepting: wie Simulated Annealing, allerdings wird eine Verschlechterung immer bis zu einem Maximalwert akzeptiert. Dieser wird während des Verfahrens immer weiter abgesenkt.

Sintflutalgorithmus: Prinzip wie Hillclimbing; es wird eine Toleranzschranke festgelegt. Eine veränderte Lösung wird akzeptiert, wenn sie besser als die Schranke ist. Die Schranke wird während des Verfahrens auf immer bessere Werte gesetzt.

Evolutionsstrategien: Prinzip ähnlich wie Hillclimbing, aber es wird gleichzeitig mit einer Population von *m* Lösungen gearbeitet, die jeweils *n* "Kinder" erzeugen. Die besten *m* bilden die nächste Generation. Adaptive Evolutionsstrategien passen die Schrittweite (Größe der Veränderung) an die Verbesserungsquote der Kinder an.

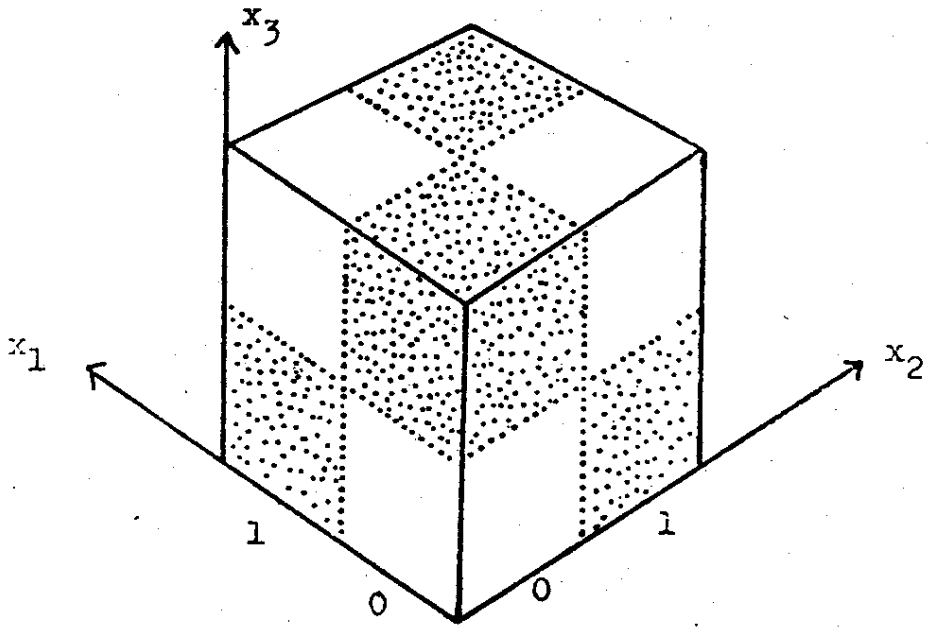
8. 7. Die Häufungsanalyse von Binärmustern und die Konfigurationsfrequenzanalyse (KFA)

Ziel: Auffinden besonders häufiger Merkmalskombinationen
Entdecken von "Typen" oder "Syndromen"

Voraussetzung: binäre Merkmale, d.h. als Werte nur 0 und 1

Motivation:
Korrelationsanalyse je zweier Merkmale reicht manchmal nicht aus

"Meehl'sches Paradoxon":



(aus Krauth & Lienert 1973)

je zwei der Merkmale x_1, x_2, x_3 sind völlig unkorreliert
die Dreierkombinationen (0, 0, 0), (0, 0, 1), ..., (1, 1, 1) sind jedoch sehr ungleich verteilt

Klassisches Beispiel aus der Psychologie:
Versuche von *Leuner* zur Wirkung von LSD

"psychotoxisches Basissyndrom": Bewusstseinstrübung, Denkstörung und Affektivitätsbeeinflussung
in den Versuchsprotokollen gab es keine paarweisen Korrelationen zwischen diesen drei Merkmalen

⇒ Syndrom nicht existent?

B	D	A	f_{ijk}	e_{ijk}	$\chi^2_{ijk} = (f—e)^2/e$	Typ
+	+	+	20	12,506	4,491*	trisympt.
+	+	—	1	6,848	4,995	
+	—	+	4	11,402	4,805	
+	—	—	12	6,244	5,306*	monosympt.
—	+	+	3	9,464	4,415	
—	+	—	10	5,182	4,479*	monosympt.
—	—	+	15	8,629	4,704*	monosympt.
—	—	—	0	4,725	4,725	

die Merkmalskombinationen (+, +, +), (+, –, –), (–, +, –)
und (–, –, +) sind häufiger als erwartet.

Quantifizierung:

(a) Häufungsanalyse (Clusteranalyse) von Binärmustern

Fragestellung: Gibt es Merkmalskombinationen, die besonders häufig sind?

Nullhypothese: jedes der *k* Merkmale hat Wahrscheinlichkeit ½, die Merkmale sind unabhängig voneinander
⇒ für jede der 2^k Merkmalskombinationen ist der Erwartungswert e = $N/2^k$ (*N* = Anzahl der Objekte)

f sei die absolute Häufigkeit (Frequenz) einer Merkmalskombination

man betrachtet die Testgröße $T = (f - e)^2 / e$

eine signifikante Häufung ("Überfrequentierung") liegt vor, wenn *T* größer ist als das α-Quantil einer Chiquadratverteilung mit 1 Freiheitsgrad.
Insgesamt verwirft man die Nullhypothese, wenn die Summe aller *T*-Werte größer ist als das α-Quantil einer Chiquadratverteilung mit $2^k - k - 1$ Freiheitsgraden.

Beispiel:
5 binäre Merkmale (Q, G, A, N und D)

QGAND	f	e	$(f-e)^2/e$	
+++++	12	4,7	11,3	*
++++-	4	4,7	0,1	
+++-+	7	4,7	1,1	
+++--	1	4,7	2,9	
++-++	7	4,7	1,1	
++-+-	2	4,7	1,6	
++--+	7	4,7	1,1	
++---	1	4,7	2,9	
+-+++	0	4,7	4,7	
+-++-	2	4,7	1,6	
+-+-+	1	4,7	2,9	
+-+--	0	4,7	4,7	
+--++	0	4,7	4,7	
+--+-	2	4,7	1,6	
+---+	0	4,7	4,7	
+----	0	4,7	4,7	
-++++	7	4,7	1,1	
-+++-	4	4,7	0,1	
-++-+	11	4,7	8,4	*
-++--	7	4,7	1,1	
-+-++	7	4,7	1,1	
-+-+-	8	4,7	2,3	
-+--+	9	4,7	3,9	
-+---	17	4,7	32,2	*
--+++	2	4,7	1,6	
--++-	1	4,7	2,9	
--+-+	2	4,7	1,6	
--+--	9	4,7	3,9	
---++	0	4,7	4,7	
---+-	5	4,7	0,0	
----+	4	4,7	0,1	
-----	11	4,7	8,4	*

(aus Lautsch & Lienert 1993)

die signifikanten Kombinationen zum Niveau α = 0,01 sind mit einem Stern markiert; Schwellenwert für *T*: 5,41.

Häufungsanalyse macht Voraussetzung der Gleichwahrscheinlichkeit für + und – bei jedem Merkmal (oft unrealistisch)
$\Rightarrow$ suche nur Binärmuster, die signifikant häufiger auftreten als aufgrund der Gesamt-Häufigkeiten der einzelnen Merkmalsausprägungen zu erwarten ist
$\Rightarrow$ Auffinden von *Typen* oder *Syndromen*

⇒ (b) Konfigurationsfrequenzanalyse (KFA)

berechne Erwartungswert jeder Merkmalskombination einzeln aus dem Produkt der eindim. Randsummen:

$$e = \frac{\prod_{i=1}^{k} f_{(i)}}{N^{k-1}}$$, wobei die $f_{(i)}$ die Summe aller Häufigkeiten von Konfigurationen mit festgehaltener *i*-ter Merkmalsausprägung (übereinstimmend mit der Ausprägung in der betrachteten Kombination) ist.
Teststatistik ist die gleiche wie bei der Häufungsanalyse.

Beispiel:

QGAND	f	e	$\frac{(f-e)^2}{e}$		QGAND	f	e	$\frac{(f-e)^2}{e}$	
+++++	12	3,4	21,8	*	-++++	7	7,6	0,1	
++++-	4	3,3	0,1		-+++-	4	7,4	1,6	
+++-+	7	4,7	1,1		-++-+	11	10,6	0,0	
+++--	1	4,6	2,7		-++--	7	10,3	0,7	
++-++	7	3,9	2,5		-+-++	7	8,7	0,3	
++-+-	2	3,8	0,9		-+-+-	8	8,5	0,0	
++--+	7	5,3	0,5		-+--+	9	12,1	0,8	
++---	1	5,2	3,4		-+---	17	11,7	2,4	
+-+++	0	1,2	1,2		--+++	2	2,7	0,2	
+-++-	2	1,2	0,6		--++-	1	2,6	1,0	
+-+-+	1	1,6	0,2		--+-+	2	3,7	0,8	
+-+--	0	1,6	1,6		--+--	9	3,6	8,1	*
+--++	0	1,4	1,4		---++	0	3,1	3,1	
+--+-	2	1,3	0,4		---+-	5	3,0	1,3	
+---+	0	1,9	1,9		----+	4	4,2	0,0	
+----	0	1,8	1,8		-----	11	4,1	11,6	*

⇒ entdeckt werden das "Vollsyndrom" (+++++), ein "Monosyndrom" (– – + – –) und das "Nullsyndrom" (– – – – –).
(aus Lautsch & Lienert 1993).

Achtung: Wenn man ohne Typenhypothesen testet (explorativ), ist der Fehler erster Art durch die Zahl der Konfigurationen zu teilen ("α-Adjustierung nach Bonferroni"):
neues $\alpha = \alpha/2^k$.

8. 8. *Die Faktorenanalyse*

- Beschäftigt sich *nicht* mit dem Zusammenhang zwischen abhängiger und unabhängiger Variablen
- Untersucht *Zusammenhänge* in einer Gruppe von Variablen (welche Variablen variieren gemeinsam)
- Versucht, jeweils ein Bündel von Variablen durch eine synthetische Variable („Faktor“) zu ersetzen

- Dieser Faktor repräsentiert eine inhaltliche Dimension
- Grobunterteilung zwischen
  - explorativer (Struktur-suchender) Faktoranalyse
  - konfirmatorischer (Struktur-prüfender) Faktorenanalyse

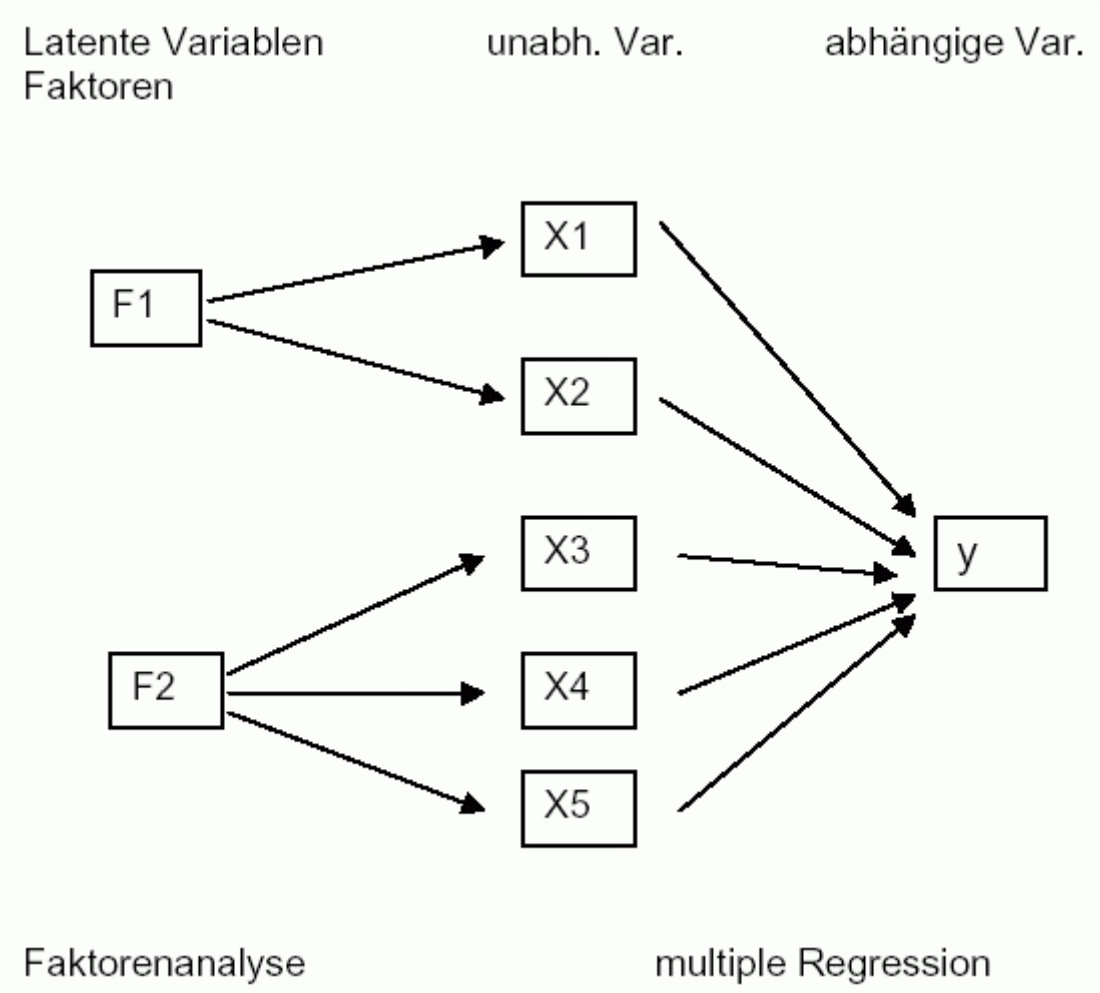
- Ausgangspunkt: Korrelationen bzw. Kovarianzen zwischen Variablen
- Faktorenanalyse versucht, Ausgangsvariablen durch möglichst wenige Faktoren zu ersetzen

	V1	V2	V3	V4
V1	1.0	1.0	1.0	1.0
V2		1.0	1.0	1.0
V3			1.0	1.0
V4				1.0

- Ein Faktor

	V1	V2	V3	V4
V1	1.0	0.0	1.0	0.0
V2		1.0	0.0	1.0
V3			1.0	0.0
V4				1.0

- Zwei unkorrelierte Faktoren



- "Faktorladungen": Korrelation zwischen der ursprünglichen Variablen und dem Faktor (der "latenten Dimension")
- Im Idealfall lädt jede Variable hoch auf einen einzigen Faktor ("Einfachstruktur")
- So daß sich die verborgene Struktur deutlich ausmachen läßt

Die Korrelationsmatrix gibt einen ersten Überblick darüber, welche Variablen stark und welche nur schwach miteinander korrelieren.

Weisen alle ausgegebenen Korrelationskoeffizienten nur sehr geringe (absolute) Werte auf, so wäre es wenig sinnvoll, die FA fortzusetzen, weil gemeinsame Faktoren nur für solche Variablen existieren, die (relativ) stark miteinander korrelieren.

Das faktorenanalytische Modell

In der Absicht ein möglichst einfaches Modell zu formulieren, wird ein linearer Ansatz für die p Zielvariablen $Z_1, \ldots, Z_p$ gewählt:

$$\begin{aligned} Z_1 &= \alpha_{11}F_1 + \alpha_{12}F_2 + \ldots + \alpha_{1r}F_r + U_1, \\ Z_2 &= \alpha_{21}F_1 + \alpha_{22}F_2 + \ldots + \alpha_{2r}F_r + U_2, \\ &\vdots \\ Z_p &= \alpha_{p1}F_1 + \alpha_{p2}F_2 + \ldots + \alpha_{pr}F_r + U_p. \end{aligned}$$

Bezeichnungen:

- Die Faktoren $F_1, \ldots, F_p$ heißen *gemeinsame Faktoren* (common factors), da sie zur Erklärung jeder im Modell enthaltenen Variable herangezogen werden.
- Die Faktoren $U_1, \ldots, U_p$, die jeweils nur mit einer Variablen Z_j verbunden sind, heißen *spezifische Faktoren* oder *Einzelrestfaktoren* (unique factors). In ihnen sind auch mögliche Meßfehler enthalten.
- Die Koeffizienten α_{ij} werden als *Faktorladungen* bezeichnet; α_{ij} gibt den Einfluß oder die Ladung des j-ten Faktors in der i-ten Variablen an.

Bemerkungen:

- Formal ist die Gleichung

  $$Z_j = \alpha_{j1}F_1 + \alpha_{j2}F_2 + \ldots + \alpha_{jr}F_r + U_j$$

  ein multiples lineares Regressionsmodell. Der entscheidende Unterschied besteht darin, dass die Regressoren im Regressionsmodell vorgegeben sind, während die Faktoren *latente Variablen* oder *hypothetische Konstrukte* sind, die noch zu *extrahieren* sind.

Schritte der Faktorenanalyse:

1. **Korrelationsmatrizen:** Lassen Sie Korrelationsmatrizen für alle in die FA einbezogenen Variablen berechnen und ausgeben. Die Betrachtung dieser Matrizen kann zeigen, welche Variablen gleiche "Aspekte" messen.
2. **Faktoraxtraktion:** Dies wird auch als "Ziehen" oder "Extrahieren" von Faktoren bezeichnet. Da es verschiedene Methoden der Faktorextraktion gibt, müssen Sie angeben, welche Methode angewandt werden soll.

3. **Rotation:** Die im zweiten Schritt gefundenen Faktoren sind häufig schwierig zu interpretieren. Daher sollen sie so transformiert werden, dass ihre Verbindung zu den Beobachtungsvariablen eindeutiger und damit besser interpretierbar wird.

4. **Faktorwerte:** Oftmals besteht das Ziel einer Faktorenanalyse darin, die ermittelten Faktoren zur Erklärung anderer Variablen zu verwenden. Für diese Zwecke können Sie Faktorwerte ermitteln lassen und ggf. speichern.

- Vielfalt von Eingriffsmöglichkeiten
  - Extraktionsverfahren
  - Rotation der Faktoren
  - Zahl der extrahierten Faktoren

Variante:

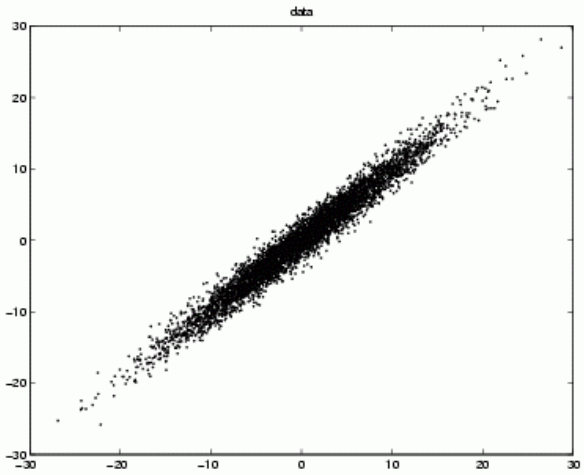
Konfirmatorische Faktorenanalyse

- Forscher definiert *vorab* theoretisch begründete Faktorenstruktur
- Zusammenhang zwischen Ausgangsvariablen und theoretischer Struktur kann mit Signifikanztests überprüft werden (LISREL)

wichtigstes Verfahren der *Faktorenextraktion* ist die *Hauptkomponentenanalyse*

8. 9. Die Hauptkomponentenanalyse
(Principal Component Analysis, PCA)

Motivation:

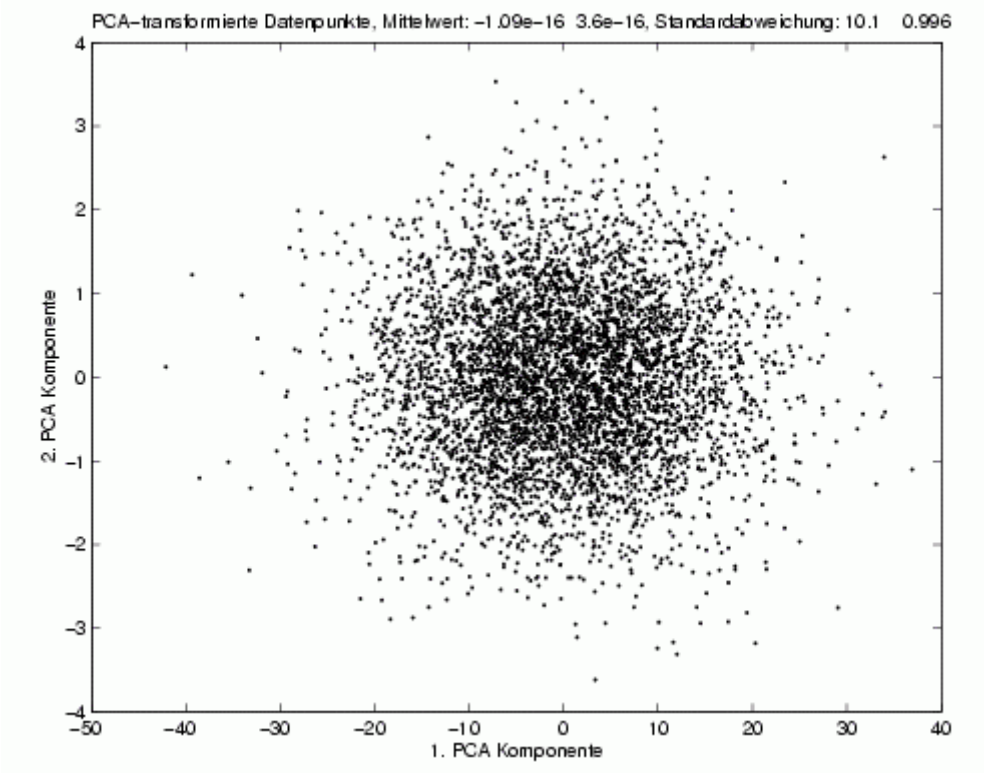


- Variation der Daten in Richtung der beiden vorgegeben Merkmale ist gleich.
- In Richtung des Vektors $(1, 1)$ ist die Variation der Daten groß; in Richtung $(1, -1)$ dagegen gering.

- Offensichtlich sind Merkmale in denen die Merkmalsausprägungen überhaupt nicht variieren bedeutungslos.
- Datenreduktion in hochdimensionalen Merkmalsräumen durch Auffinden von Richtungsvektor mit großer Variation (die sogenannten **Hauptachsen**).

- Die Hauptachsen lassen sich anordnen:
  - $1.$ Hauptachse beschreibt den Vektor $v_1 \in \mathbb{R}^d$ mit der größten Variation der Daten;
  - $2.$ Hauptachse ist der Vektor $v_2 \in \mathbb{R}^d$ der senkrecht auf v_1 steht Vektoren und in dessen Richtung die Datenpunkte am stärksten variieren.
  - $l.$ Hauptachse ist der Vektor $v_l \in \mathbb{R}^d$ der senkrecht auf $V_{l-1} := lin\{v_1, \ldots, v_{l-1}\}$ steht und in dessen Richtung die Datenpunkte am stärksten variieren

hauptachsentransformierte Daten aus dem obigen Scatterplot:

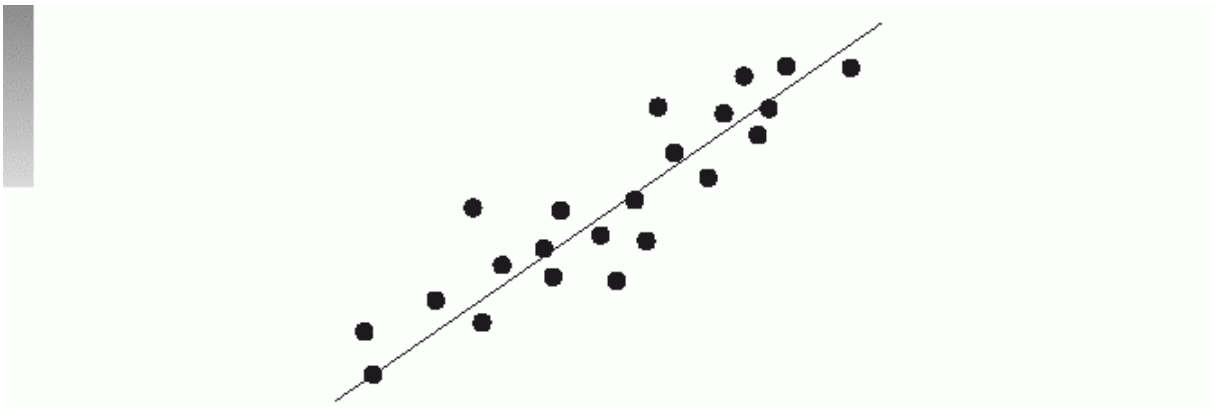


(aus Schwenker 2004)

Die **Hauptkomponentenanalyse (principle component analysis, PCA)** ist primär eine Technik zur Dimensionsreduktion, kann aber auch zur Visualisierung verwendet werden.

Das Ziel der Hauptkomponentenanalyse besteht darin, die Daten (linear) auf einen Raum mit weniger Dimensionen zu projizieren unter möglichst guter Erhaltung der Streuung/Variabilität der Daten.

Einsatz zur Dimensionsreduktion:



Eine Projektion der hier zweidimensionalen Daten auf die eingezeichnete (eindimensionale) Gerade (Hauptkomponente) führt zu einer Dimensionsreduktion mit geringem Informationsverlust.

Allgemein: Projektion p-dimensionaler Daten auf einen (linearen) q-dimensionalen Raum mit $q \ll p$.

Werden die Daten zunächst so verschoben, dass ihr Mittelwert im Koordinatenursprung liegt, lässt sich die Projektion als $q \times p$-Matrix schreiben:

$$\mathbf{y} = M \cdot (\mathbf{x} - \bar{\mathbf{x}})$$

Wobei $\bar{\mathbf{x}}$ den Mittelwert bezeichnet:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

(Klawonn 2004)

Bestimmung der Matrix M:
M soll die Summe der Varianzen der Komponenten der projizierten Daten $\mathbf{y}$ maximieren, d.h.

$$s_Y^2 = \frac{1}{n-1} \sum_{i=1}^{n} \mathbf{y_i}^\top \mathbf{y_i}$$

(Der Mittelwert der $\mathbf{y}$-Werte ist aufgrund der Definition 0.)

$$\mathbf{y_i} = M \cdot (\mathbf{x_i} - \bar{\mathbf{x}})$$

Diese Varianz kann beliebig vergrößert werden, indem in M möglichst große Werte eingetragen werden.

Um dies zu vermeiden, fordern wir: $M^\top M = \mathbf{1}$

D.h., M darf die Daten rotieren, aber nicht stauchen oder strecken.

Die Lösung dieses Optimierungsproblems mit der Randbedingung $M^\top M = \mathbf{1}$ ergibt:

$$M = (\mathbf{v_1}, \ldots, \mathbf{v_q})$$

wobei $\mathbf{v_1}, \ldots, \mathbf{v_q}$ die normierten Eigenvektoren der Kovarianzmatrix der Daten

$$C = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^\top$$

zu den q größten Eigenwerten $\lambda_1 \geq \ldots \geq \lambda_q$ sind

Die Summe der Varianzen der projizierten Daten ist dann die Summe dieser Eigenwerte:

$$\lambda_1 + \ldots + \lambda_q.$$

Wird die Hauptkomponentenanalyse zur Dimensionsreduktion verwendet, betrachtet man daher wie viel Prozent der Varianz durch eine Projektion auf q Dimensionen abgedeckt werden:

$$\frac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_p} \cdot 100\%$$

Im Beispiel des Irisdatensatzes:

q	Varianzabdeckung in %
1	72.96
2	95.81
3	99.48
4	100.00

Man wählt q so, dass eine bestimmte Prozentzahl (z.B. 95%) überschritten wird, oder erhöht q, bis eine weitere Erhöhung von q kaum noch zu einer verbesserten Varianzabdeckung führt. In unserem Beispiel also $q = 1$, evtl. $q = 2$.

Ein wesentlicher Nachteil bei der Dimensionsreduktion mittels Hauptkomponentenanalyse besteht darin, dass man zwar mit weniger Attributen arbeitet.

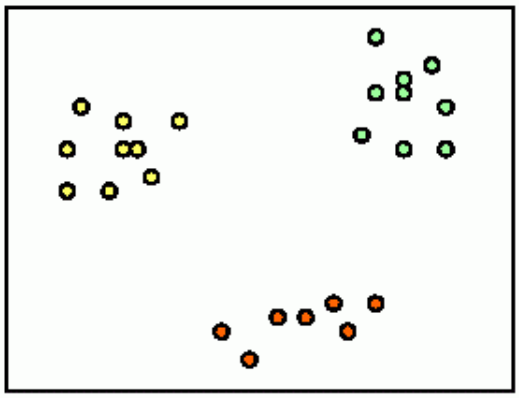
Durch die Transformation/Projektion ergeben sich die neuen Attribute aber als Linearkombinationen der ursprünglichen Attribute und sind daher schwer interpretierbar.

8. 10. Correlation Clustering
(Ergänzung zum Kapitel über Clusteranalyse)

nach Böhm 2003

Motivation:

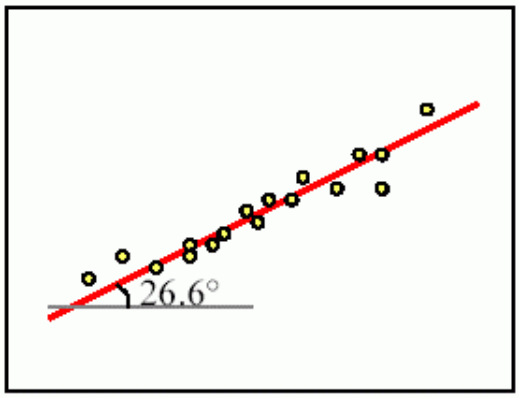
Clustering...



Einteilung der Punktemenge in Gruppen (Cluster), so dass...

- Maximale Ähnlichkeit der Punkte innerhalb der Cluster
- Minimale Ähnlichkeit der Punkte versch. Cluster

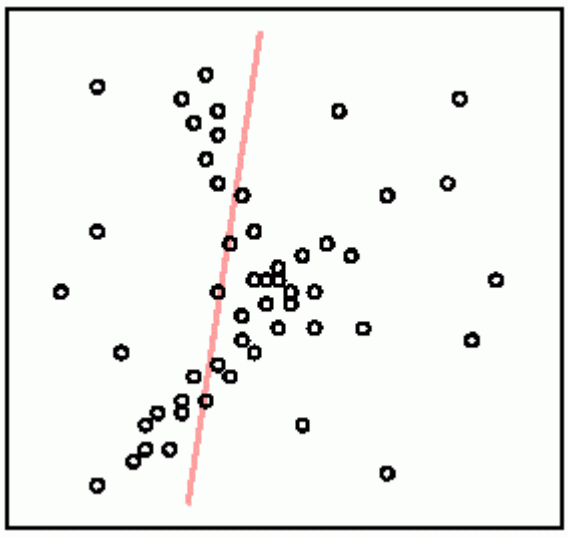
Korrelation...



$y \approx 0.5\,x + \ldots$

(lineare) Abhängigkeit zwischen den einzelnen Attributen (Dimensionen) einer Punktemenge

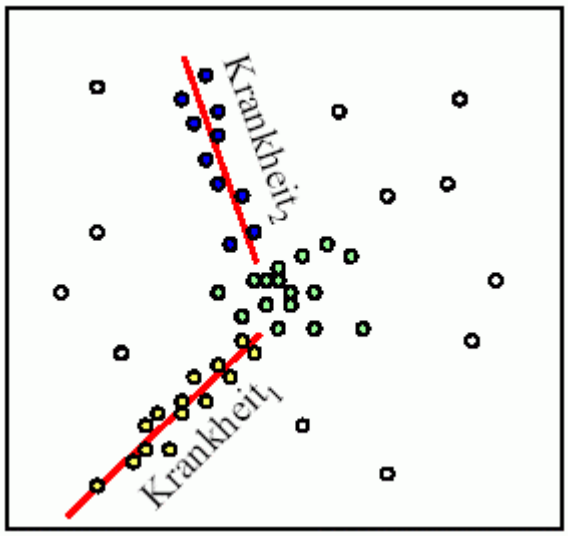
Probleme der Korrelation:



Rausch-Punkte

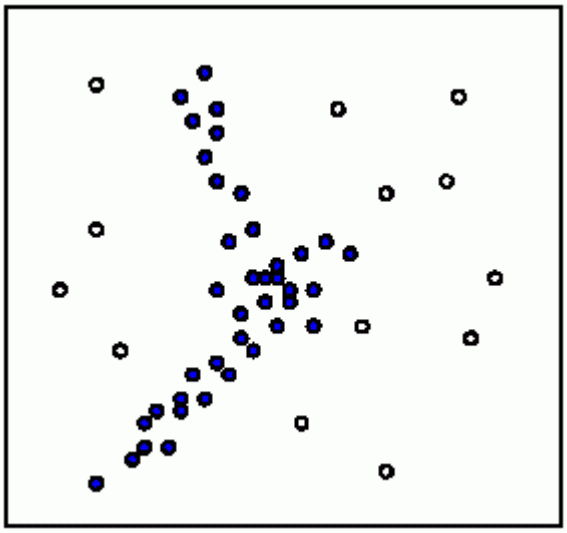
Verschiedene Teilmengen weisen unterschiedliche Korrelationen auf

→ schwache Gesamt-Korrelation



Ziel:
Suche nach Teilmengen von Punkten mit einheitlicher Korrelation

Problem des dichtebasierten Clustering (vgl. Kap. 7):



Trennt grundsätzlich auch Correlation Cluster von Rauschpunkten

Separiert aber nicht nach unterschiedlicher Regressionslinie

Informelle Def. des Correlation Clustering:

Ein Korrelations-verbundener Cluster ist eine Punktmenge mit...

- einheitlicher Punktdichte (bzw. Dichte-Schwellwert)
- einheitlicher Korrelation (Regressionslinie)

d.h. ein Correlation-Clustering-Verfahren soll

- Punktdichte und
- Korrelation

innerhalb von Clustern maximieren

zwischen separierten Clustern minimieren

Erweiterung von dichtebasiertem Clustering

- DBSCAN
- OPTICS
- Oder eines anderen Verfahrens

ggf. unter Einführung neuer Parameter (Dimension der Korrelation)

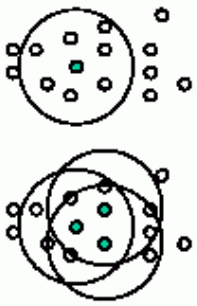
Möglichkeiten, Korrelation ins Spiel zu bringen

- Adaptives Ähnlichkeitsmaß
- Fraktale Dimension
- Hough-Transformation

Hier betrachten wir nur den Ansatz des *adaptiven Ähnlichkeitsmaßes*:

DBSCAN beruht im wesentlichen auf zwei Konzepten:

- Kernpunkte:
  Punkte, in deren ε-Umgebung sich mindestens *MinPts* Punkte befinden
- Dichte-Verbundenheit:
  Kernpunkte werden mit Nachbarn in der ε-Umgebung vereinigt



Idee von 4C (Computing Correlation Connected Clusters):

- Anpassung dieser Konzepte von DBSCAN, so dass nach korrelierten Punktmengen gesucht wird
- Dimension λ der Korrelation durch den Benutzer vorgegeben:
  - λ = 1 für Korrelations-Linien
  - λ = 2 für Korrelations-Ebenen usw.

"Kernpunkte" im neuen Verfahren:

Zusätzlich zur Forderung, dass sich in der ε-Umgebung mindestens *MinPts* Nachbarn befinden müssen:

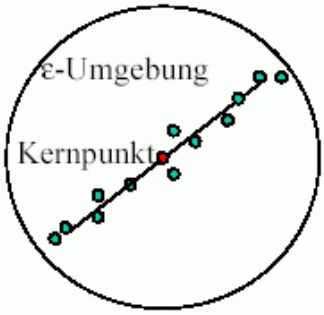
Die Punkte in der ε-Umgebung eines Kernpunktes müssen sich

- ...auf (bzw. in der Nähe) einer gemeinsamen Linie (im Fall $\lambda=1$),
- ...einer gemeinsamen Ebene (im Fall $\lambda=2$),
- ...einer gemeinsamen λ-dimensionalen Hyperebene (im Fall $\lambda>2$)

durch den Kernpunkt befinden.

Dies lässt sich mathematsch wie folgt bestimmen:

- Berechnung Kovarianzmatrix Σ der Nachbarn
- Eigenwert-Zerlegung (Principal Components) $V \cdot E \cdot V^T = \Sigma$
- Mindestens d–λ Eigenwerte müssen ≈ 0 sein

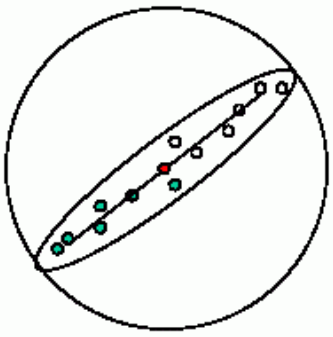


Hierdurch wird jedem Kernpunkt eine Kovarianzmatrix zugeordnet

Dichte- bzw. Korrelations-Verbundenheit:

Prinzip:

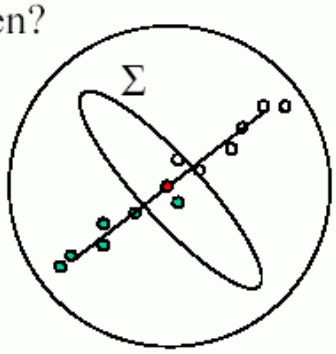
- Nur solche Punkte sollen mit einem Cluster vereinigt werden, die auch in der bisherigen Ausdehungsrichtung (d.h. nahe zur Korrelationslinie, -Ebene usw.) liegen



Wie kann dieses Ähnlichkeitsmaß erreicht werden?

- Ähnlichkeitsmaß entspricht Kovarianzmatrix: $\text{dist}^2(P,Q) = (P-Q) \cdot \Sigma \cdot (P-Q)^T$

  Richtungen starker Varianz werden durch das Ähnlichkeitsmaß stark gewichtet.
  $\Rightarrow$ Ansatz genau kontraproduktiv!

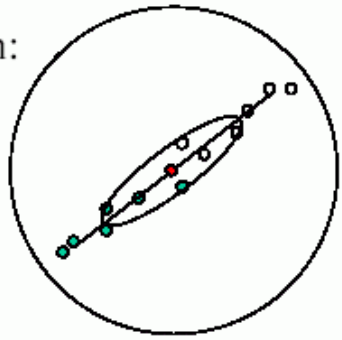


- Kovarianzmatrix mit invertierten Eigenwerten:
  $\text{dist}^2(P,Q) = (P-Q) \cdot V \cdot E^{-1} \cdot V^T \cdot (P-Q)^T$

  Anmerkung: Diagonalmatrizen werden elementweise invertiert:
  $\text{diag}(a_1,a_2,\ldots)^{-1} = \text{diag}(1/a_1,1/a_2,\ldots)$

  Ausrichtung des Ellipsoids nun korrekt!

  Probleme:
  - Was macht man mit Eigenwerten =0 (also *keine* Varianz in dieser Richtung)?
  - Ausdehnung des Ellipsoids in allen Richtungen verschieden und nicht klar definiert

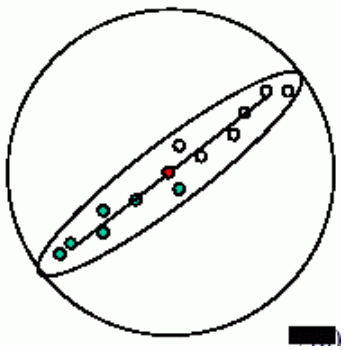


Gewünscht: Ellipsoid mit folgenden Eigenschaften

- Ausrichtung gemäß den stärksten Eigenvektoren
- Ausdehnung ε in λ Richtungen
- Eine einheitliche, wesentlich geringere Ausdehung, die eine gewisse Toleranz erlaubt, in den verbleibenden d–λ Richtungen

Die Eigenwertmatrix wird wie folgt modifiziert:

- Die ersten λ Eigenwerte werden auf 1 gesetzt (Ellipsoid ist definiert als $\{x \mid \text{dist}(P,x) \leq \varepsilon\}$)
- Die verbleibenden d–λ Eigenwerte auf κ ($\kappa >> 1$ Benutzer-definierter Wert)
- Distanzmaß mit modifiziertem E':
  $\text{dist}^2(P,Q) = (P-Q) \cdot V \cdot E' \cdot V^T \cdot (P-Q)^T$



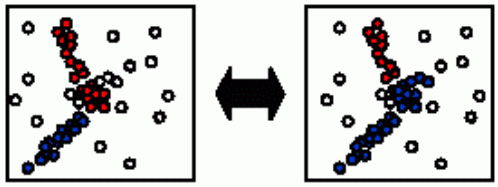
Symmetrisierung der Metrik:

Beobachtung:

Das Abstandsmaß ist nicht symmetrisch, da immer die modifiz. Kovarianzmatrix, die einem der beiden beteiligten Kernpunkte zugeordnet ist, das Abstandsmaß definiert.

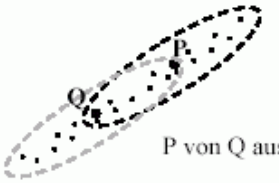
Problem:

Hierdurch wird das Clusterverfahren Reihenfolge-abhängig

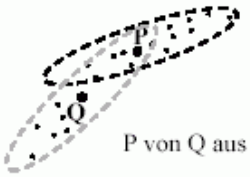


Lösung:

Vereinige Punkte nur dann, wenn sie sich „gegenseitig" finden, also: $\text{dist}_P(P,Q) \leq \varepsilon$ **und** $\text{dist}_Q(Q,P) \leq \varepsilon$



P von Q aus direkt dichte-erreichbar



P von Q aus nicht direkt dichte-erreichbar

Algorithmus 4C

Input-Parameter: ε, MinPts, λ.

Für alle Objekte *o* aus der Datenbank:

Schritt 1: Test auf Korrelations-Kernobjekt

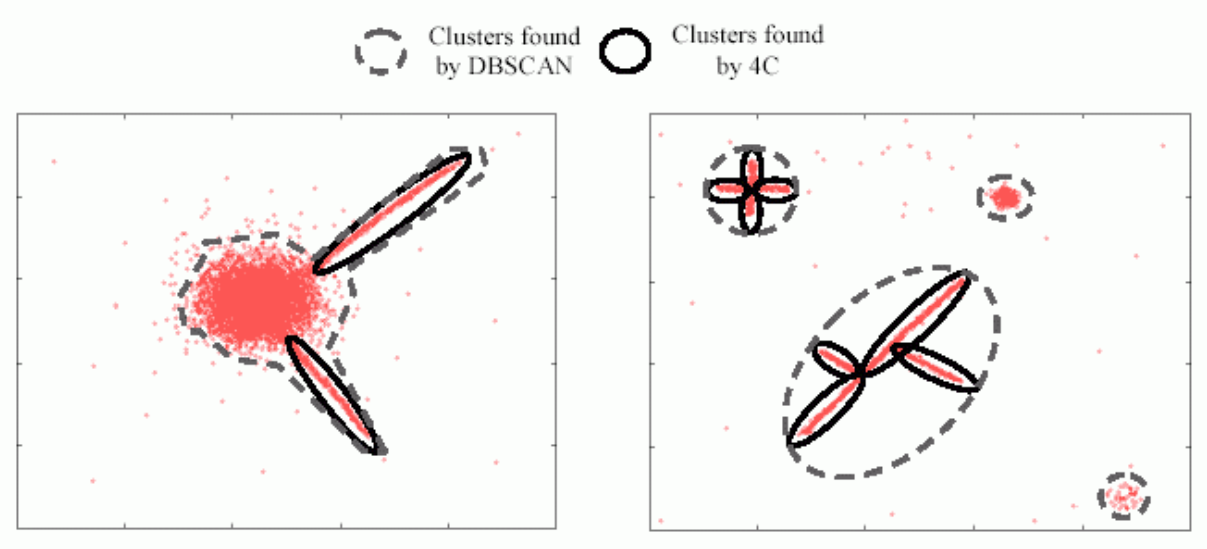
berechne ε-Umgebung $N_\varepsilon(o)$ von *o*;
Wenn $|N_\varepsilon(o)| \geq MinPts$
    berechne Σ;
    Wenn (d-λ) Eigenwerte ≈ 0
        berechne E´;
        berechne ε-Umgebung $N'_\varepsilon(o)$ von *o* bzgl. E´;
        teste $|N'_\varepsilon(o)| \geq MinPts$;

Schritt 2: Expandiere Cluster

berechne alle Punkte, die korrelations-dichte-erreichbar von *o* sind;

- ähnlich wie DBSCAN
- benutze dabei E´ als Distanzmaß
- achte auf Symmetrie

Ergebnisse: Vergleich mit DBSCAN



Diskussion von 4C:

Stärken

- Erstes Verfahren, das Teilmengen in einer Menge von Merkmalsvektoren ermittelt, die einheitliche Korrelation aufweisen (mit Ausnahme von ORCLUS, dessen „orientierte Cluster“ ähnlich funktionieren)
- Wesentlich bessere Ergebnisse als ORCLUS (*k*-Means)

Schwächen

- Mengen müssen zusätzlich zur Korrelation auch Dichte-verbunden sein (Parameter ε)
- (zurzeit noch) nicht hierarchisch → 4C-OPTICS?
- Dimensionalität λ der Korrelation muss vorgegeben werden
- Findet nur lineare Abhängigkeiten
- Punkte können nur einem Cluster zugeordnet sein

(Böhm 2003)