Histogramm

Histogramm-Funktion: $\boldsymbol{H(x) = \sum_{i=1}^{k-1} \left(\frac{n_i}{n} \frac{1}{t_{i+1} - t_i} \right) I_{[t_i, t_{i+1})}(x)}$

k… Klassenanzahl, t … Intervallgrenzen I … Indikatorfunktion, n… Umfang Stichprobe,n_i... # im Intervall

Intervalllänge nach Sturges: $\boldsymbol{h_n = t_{i+1} - t_i = \lceil \log_2(n) + 1 \rceil}$

Intervalllänge nach Scott: $\boldsymbol{h_n = \frac{3.5s}{\sqrt[3]{n}}}$ s … empirische Standardabweichung

Intervallänge nach Freedman: $\boldsymbol{h_n = \frac{(2IQR)}{\sqrt[3]{n}}}$

Verteilungsfunktion

von 0 bis 1, monoton wachsend, $\lim_{x \to \infty} F(x) = 1$, $\lim_{x \to -\infty} F(x) = 0$

theoretische Verteilungsfunktion: $\boldsymbol{F(x) = \int_{-\infty}^{x} f(t)dt}$

empirische Verteilungsfunktion: $\boldsymbol{F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{[x_i, \infty)}(x)}$

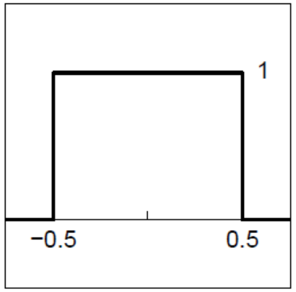
Dichteschätzung/Kernschätzung

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}W\left(\frac{x-x_i}{h}\right)$$

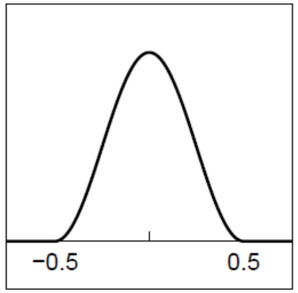
W(t) … Gewichtsfunktion $\int_{-\infty}^{\infty} W(t) = 1$,

$\left[x-\frac{h}{2}, x+\frac{h}{2}\right]$ … Fenster (h = Intervalllänge)

Rechtecks Gewichtsfunktion: $W(t) = \begin{cases} 1 & |t| \leq \frac{1}{2} \\ 0 & sonst \end{cases}$



Cosinus Gewichtsfunktion: $W(t) = \begin{cases} 1+\cos 2\pi t & |t| < \frac{1}{2} \\ 0 & sonst \end{cases}$



Wahrscheinlichkeitsnetz

=Verteilungsfunktion mit verzerrter vertikaler Achse → zb.: Abstände proportional zur Inversen der Verteilungsfunktion von $N(0,1) = \Phi$. y-Achse = $\Phi^{-1}\left(F_n(x)\right)$ x-Achse = x.

Bei ungefähr Normalverteilung gilt: $\Phi^{-1}\left(F_n(x)\right) \sim \Phi^{-1}\left(\Phi\left(\frac{x-\mu}{\sigma}\right)\right) = \frac{x-\mu}{\sigma}$ → auf einer Geraden.

bei Wahrscheinlichkeit 50% Schätzung für μ bei 84% Schätzung für $\mu+\sigma$

Quantile-Quantile Plots

$$F_x(t) = P(X \leq t), q_x(p) = F_x^{-1}(p)$$

$$F_x = F_y \leftrightarrow \left(q_x(p_i), q_y(p_i)\right) \textit{ liegen auf der Geraden } y = x$$

univariate Schätzer

arithmetisches Mittel: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ empirische Standardabweichung $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

gestutztes Mittel: $m(\alpha) = \frac{1}{n-2g}\left(x_{(g+1)} + .. + x_{(n-g)}\right)$ $\alpha \ldots 0 \leq \alpha \leq 0.5$ $g = \lfloor n\alpha \rfloor$ (auf Ganzzahl abgerundet)

gestutzte Streuung $S(\alpha) = \sqrt{\frac{1}{n-2g-1}\sum_{i=g+1}^{n-g}\left(x_{(i)} - m(\alpha)\right)^2}$

gestutzte SD = $\frac{S(\alpha)}{c_\alpha}$ c_α .. abhängig von α

Median $median(x_1, \dots, x_n) = \begin{cases} \boldsymbol{x}_{\left(\frac{n+1}{2}\right)} & n\ ungerade \\ \frac{\boldsymbol{x}_{\left(\frac{n}{2}\right)} + \boldsymbol{x}_{\left(\frac{n}{2}+1\right)}}{2} & n\ gerade \end{cases}$

MAD (Median Absolute Deviation) = $\boldsymbol{median_{1 \le i \le n}\left(\left|x_i - median_{1 \le j \le n}(x_j)\right|\right)}$

$$s_{MAD} = \frac{MAD}{0.675} = 1.483 \cdot MAD$$

25 Quartil $\boldsymbol{Q_{0.25} = median\left(x_{(1)}, \dots, x_{\left(\left[\frac{n+1}{2}\right]\right)}\right)}$

75 Quartil $\boldsymbol{Q_{0.75} = median\left(x_{\left(\left[\frac{n}{2}+1\right]\right)}, \dots, x_{(n)}\right)}$

Interquartilabstand **IQR =** $\boldsymbol{Q_{0.75} - Q_{0.25}}$ $\boldsymbol{s_{IQR} = \frac{IQR}{1.35}}$

Q_n Streuungsschätzer $\boldsymbol{Q_n = \left\{\left|x_i - x_j\right|; i < j\right\}_k}$ $\boldsymbol{s_{Q_n} = 2.219 \cdot Q_n}$

Dichteschätzung in zwei Dimensionen

Boxcar Funktion $W(u,v) = \begin{cases} \frac{1}{\pi}\ u^2 + v^2 \leq 1 \\ 0\ sonst \end{cases}$

Cosinus Gewichtsfunktion $W(u,v) = \begin{cases} \frac{1+\cos\left(\pi\sqrt{u^2+v^2}\right)}{\pi}\ \ wenn\ u^2 + v^2 \leq 1 \\ 0\ sonst \end{cases}$

Dichtefunktion $\hat{\boldsymbol{f}}(\boldsymbol{x},\boldsymbol{y}) = \frac{\mathbf{1}}{\boldsymbol{h}^2\boldsymbol{n}}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{W}\left(\frac{\boldsymbol{x}-\boldsymbol{x}_{\boldsymbol{i}}}{\boldsymbol{h}}, \frac{\boldsymbol{y}-\boldsymbol{y}_{\boldsymbol{i}}}{\boldsymbol{h}}\right)$ h … Fensterbreite

Robuste Schätzung linearer Trends

$\boldsymbol{y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\boldsymbol{x} + \boldsymbol{\varepsilon}$ $\alpha \ldots Abszissenabstand\ \beta \ldots Steigung\ \ \varepsilon \ldots Fehler$

$\boldsymbol{r}_{\boldsymbol{i}} = \boldsymbol{y}_{\boldsymbol{i}} - \hat{\boldsymbol{y}}_{\boldsymbol{\imath}} = \boldsymbol{y}_{\boldsymbol{i}} - \hat{\boldsymbol{\alpha}} - \hat{\boldsymbol{\beta}}\boldsymbol{x}_{\boldsymbol{i}}$ $\quad \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{n}$

Least Squares

$$\left(\hat{\boldsymbol{\alpha}}_{\boldsymbol{LS}}, \hat{\boldsymbol{\beta}}_{\boldsymbol{LS}}\right) = \boldsymbol{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} (\boldsymbol{y}_{\boldsymbol{i}} - \boldsymbol{\alpha} - \boldsymbol{\beta}\boldsymbol{x}_{\boldsymbol{i}})^{\mathbf{2}} = \boldsymbol{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{r}_{\boldsymbol{i}}^{\mathbf{2}}$$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{LS}} = \frac{\sum(\boldsymbol{x}_{\boldsymbol{i}}-\overline{\boldsymbol{x}})(\boldsymbol{y}_{\boldsymbol{i}}-\overline{\boldsymbol{y}})}{\sum(\boldsymbol{x}_{\boldsymbol{i}}-\overline{\boldsymbol{x}})^{\mathbf{2}}}\ \hat{\boldsymbol{\alpha}}_{\boldsymbol{LS}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{\beta}}_{\boldsymbol{LS}}\overline{\boldsymbol{x}}$$

Bruchpunkt: $\frac{\boldsymbol{k}}{\boldsymbol{n}}$ n … Stichprobenumfang, k … maximale Anzahl von Ausreißern

Tukey

Datenpaare in 3 Gruppen nach x-Werte: $\boldsymbol{n_L + n_M + n_R = n}$

Median pro Gruppe $\boldsymbol{x_L = median_{(x_i,y_i)\in L} x_i \quad y_L = median_{(x_i,y_i)\in L} y_i}$

Schätzwerte $\boldsymbol{\hat{\beta}_0 = \frac{y_R - y_L}{x_R - x_L}}$, Residuen $r_i^0 = y_i - (\hat{\alpha}_0^{(*)} + \hat{\beta}_0(x_i - x_M))$ Bruchpunkt = $\frac{1}{6}$ (Median eines $\frac{1}{3}$)

Theil

$$\boldsymbol{\hat{\beta}_{ij} = \frac{y_j - y_i}{x_j - x_i}} \qquad 1 \le i < j \le n \quad \boldsymbol{\hat{\beta}_T = median_{1\le i<j\le n}\left(\hat{\beta}_{ij}\right)}$$ Bruchpunkt = 0.29

Siegel (Repeated Median Line) (Bruchpunkt 0.5)

$$\boldsymbol{\hat{\beta}_{RM} = median_{1\le i\le n}\left(median_{1\le j\le n}\left(\hat{\beta}_{ij}\right)\right) \quad \hat{\alpha}_{RM} = median_{1\le i\le n}\left(y_i - \hat{\beta}_{RM} x_i\right)}$$

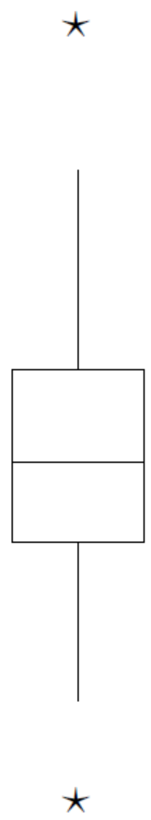
LMS (Least Median of Squares) Regression (Bruchpunkt 0.5)

$$\boldsymbol{\left(\hat{\alpha}_{LMS}, \hat{\beta}_{LMS}\right) = argmin_{\alpha,\beta} median_i r_i^2}$$

LTS (Least Trimmed Squares) Regression

$$\boldsymbol{\left(\hat{\alpha}_{LTS}, \hat{\beta}_{LTS}\right) = argmin_{\alpha,\beta} \sum_{i=1}^{h} r_{(i)}^2} \qquad \frac{n}{2} < h < n$$

Stichprobe: $x_1, \ldots, x_n$
$x_M := median(x_1, \ldots, x_n)$
$Q_{0.25}$ und $Q_{0.75}$ sind die Quantile 0.25 bzw. 0.75
$\mathrm{IQR} = Q_{0.75} - Q_{0.25}$ ist der Interquartil-Abstand



Nichtlineare Glättung

Lineare Filter $\boldsymbol{z_t = \sum_{i=-l_1}^{l_2} \alpha_i x_{t+i}}$ gewichtete Summe der x_i in einer Umgebung von t

Medianglättung $\boldsymbol{Gx_t = median(x_{t-s}, x_{t-s+1}, ..)}$ Wert an Stelle t ist Median aus einer 2s + 1 Umgebung

Repeated Median

lokale lineare Approximation $\boldsymbol{\mu_{t+1} \approx \mu_t + \beta_t \cdot i}$ $\mu_t \dots Level \quad \beta_t \dots Steigung, i = -s \; bis \; s$

$$\boldsymbol{r_t(t+i) = x_{t+i} - \hat{\mu}_t - \hat{\beta}_t \cdot i}$$

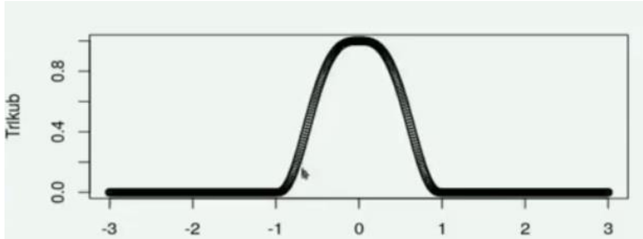
$$\boldsymbol{\widehat{\beta_t} = median_{-s \le i \le s}\left(median_{-s \le j \le s}\left(\frac{x_{t+i} - x_{t+j}}{i-j}\right)\right) \quad \widehat{\mu_t} = median_{-s \le i \le s}(x_{t+i} - \hat{\beta}_t \cdot i)}$$

Ausreißer $\boldsymbol{|\hat{r}_t| > 2 \cdot \hat{\sigma}_t}$ $\qquad \boldsymbol{\hat{r}_t = x_t - \hat{\mu}_t}$

LOWESS (LOcally WEighted regression Scatter plot Smoothing)

$\boldsymbol{q = \lfloor nf + 0.5 \rfloor}$ … Anzahl der Punkte die zur Glättung verwendet werden

$\boldsymbol{d_{ik} = |x_i - x_k|}$ $\boldsymbol{d_i = \left| x_i - x_{i_{max}} \right|}$ i_{max} … Index des am weitesten entfernen Punkt aus q



Trikubische Gewichtsfunktion $\boldsymbol{T(t) = \begin{cases} \left(1 - |t|^3\right)^3 & |t| < 1 \\ 0 & sonst \end{cases}}$

Gewicht von (x_k, y_k) $bezüglich$ (x_i, y_i) $\boldsymbol{t_i(x_k) = \begin{cases} T\left(\frac{|x_i - x_k|}{d_i}\right) = T\left(\frac{d_{ik}}{d_i}\right) & d_i \neq 0 \\ 1 & d_i = 0 \end{cases}}$

gewichtete Regression (Zeitpunkt i) $\boldsymbol{min \sum_{k=1}^{n} t_i(x_k)(y_k - a - bx_k)^2}$ Residuen $\boldsymbol{r_i = y_i - \hat{y}_i}$

Biweight Gewichtsfunktion $\boldsymbol{B(t) = \begin{cases} \left(1 - t^2\right)^2 & |t| < 1 \\ 0 \; sonst \end{cases}}$

$\boldsymbol{m = median_{1 \leq k \leq n} |r_k|}$ $\rightarrow 3m \approx 2\sigma$ Gewichte Residuen $\boldsymbol{w(x_k) = B\left(\frac{r_k}{3m}\right)}$

robust gewichtete Regression $\boldsymbol{min \sum_{k=1}^{n} w(x_k) t_i(x_k)(y_k - a - bx_k)^2}$

Upper and Lower Smoothing Aufteilung der Residuen in positive & negative Residuen $\boldsymbol{r_i = y_i - \hat{y}_i}$

$\boldsymbol{r_i^+ \; für \; (x_i^+, \hat{y}_i^+)}$

Zeitreihenanalyse

$\boldsymbol{x_t = \tau_t + \delta_t + e_t}$ τ_t ... *Trendkomponente,* δ_t ... *Saisonkomponente,* e_t ... *Restkomponente*

Lineares Modell $\boldsymbol{\ln(x_t) = \beta_0 + \beta_1 t + e_t}$

quadratisches Modell $\boldsymbol{\ln(x_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + e_t \rightarrow \hat{x}_t = exp(\hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2)}$

Fourierreihe 1. Ordnung $\boldsymbol{f(t) = a_0 + \sum_{j=1}^{J}(a_j \cos(\omega_j t) + b_j \sin(\omega_j t))}$ $\boldsymbol{w_j = \frac{2\pi j}{P}}$

$$\boldsymbol{\ln(x_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 \cos\left(\frac{2\pi}{P} \cdot t\right) + \beta_4 \sin\left(\frac{2\pi}{P} \cdot t\right) + e_t}$$

Exponentielles Glätten

$\boldsymbol{\tilde{x}_t = \alpha x_t + (1 - \alpha)\tilde{x}_{t-1}}$ α ... *Glättungsfaktor* $0 < \alpha < 1$ *Startwert* $\tilde{x}_m = mean(x_0 bis\ x_m)$

Prognose $\boldsymbol{\tilde{x}_{t+h|t} = \tilde{x}_t}$

Glättung noch Holt-Winters

$\boldsymbol{\tilde{x}_t = \alpha x_t + (1 - \alpha)(\tilde{x}_{t-1} + b_{t-1})}$ $\boldsymbol{b_t = \beta(\tilde{x}_t - \tilde{x}_{t-1}) + (1 - \beta)b_{t-1}}$

Prognose $\boldsymbol{\tilde{x}_{t+h|t} = \tilde{x}_t + hb_t}$

Autokovarianz $Cov(x_t, x_{t-k})$ Schätzer $c_k = \frac{1}{T}\sum_{t=k+1}^{T}(x_t - \overline{x})(x_{t-k} - \overline{x})$ k … lag (Abstand)

Autokorrelation der Ordnung k $\rho_k = Corr(x_t, x_{t-k}) = \frac{Cov(x_t, x_{t-k})}{Var(x_t)}$ $r_k = \frac{c_k}{c_0}$ $c_0 \ldots Varianz\ von\ x_t$

Zeitreihenmodelle

Moving Average (MA) Modell MA(1): $x_t = a + u_t - \Theta_1 u_{t-1}$ $u_t \ldots .white\ noise$

Autoregressives (AR) Modell AR(1): $x_t = a + \phi x_{t-1} + u_t$ $u_t \ldots .white\ noise$

ARMA (Autoregeressive-Moving-Average) Modell:

ARMA(1,1) = $x_t = a + \phi x_{t-1} + u_t - \Theta_1 u_{t-1}$

ARIMA Modell

Differenz Operator $\Delta x_t = x_t - x_{t-1}$

$$\Delta^2 x_t = \Delta(\Delta x_t) = \Delta(x_t - x_{t-1}) = x_t - 2x_{t-1} + x_{t-2}$$

ARIMA(1,1,1): $x_t = x_{t-1} + a + \phi(x_{t-1} - x_{t-2}) + u_t - \Theta_1 u_{t-1},$

Kovarianz

$$\boldsymbol{\sigma_{jk} = E\left[\left(x_j - E(x_j)\right)\left(x_k - E(x_k)\right)\right]} \quad , \quad (j,k \in \{1,\dots,p\}\ p \dots \#Variablen)$$

$$\boldsymbol{f}\text{ü}\boldsymbol{r\ j = k\text{: } \sigma_{jj} = E\left[\left(x_j - E(x_j)\right)^2\right] = Varianz}$$

Stichprobenkovarianz

$$\boldsymbol{s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})(x_{ik} - \overline{x_k})}$$

Korrelationskoeffizient

$$\boldsymbol{\rho_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}} = \frac{Kovarianz\ jk}{\sqrt{Varianz(j) \cdot Varianz(k)}}}$$

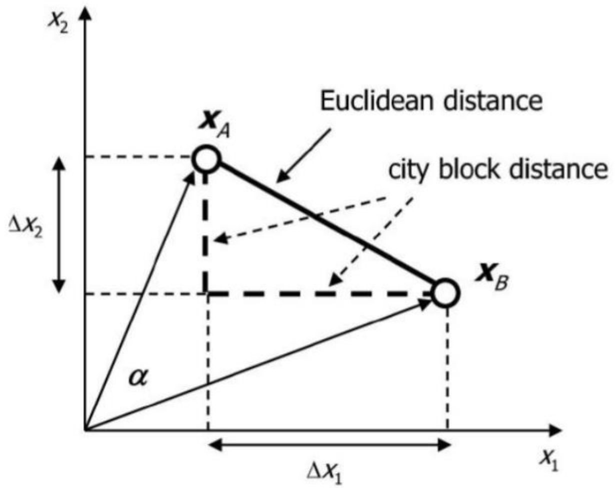
Stichprobenkorrelation

$$\boldsymbol{r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}s_{kk}}}}$$

Distanzmaß

Euklidische Distanz $\boldsymbol{d_E(x_A, x_B) = \left(\sum_{j=1}^{p}\left(x_{Bj} - x_{Aj}\right)^2\right)^{\frac{1}{2}} = ||x_B - x_A||}$

Manhattan Distanz $\boldsymbol{d_M(x_A, x_B) = \sum_{j=1}^{p}\left|x_{Bj} - X_{Aj}\right|}$



Minkowski Distanz $\boldsymbol{d_{Mink}(x_A, x_B) = \left(\sum_{j=1}^{p}\left(x_{Bj} - x_{Aj}\right)^m\right)^{\frac{1}{m}}}$

Kosinus des Winkels $\boldsymbol{\cos\alpha = \frac{x_A^T x_B}{\sqrt{(x_A^T x_A)(x_B^T x_B)}} = \frac{x_A^T x_B}{||x_A|| \cdot ||x_B||}}$

Mahalanobis Distanz $\boldsymbol{d_{Mahal}(x_A, x_B) = \left[(x_B - x_A)^T \sum^{-1}(x_B - x_A)\right]^{\frac{1}{2}}}$ Σ^{-1} inverse Kovarianzmatrix

$$\boldsymbol{d_{Mahal}(x_i, \mu) = \left[(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right]^{\frac{1}{2}}}$$ für i = 1 bis n, $\mu \ldots Zentraum\ der\ Verteilung$

Linearkombinationen

$$\boldsymbol{u = b_1 x_1 + b_2 x_2 + \cdots + b_p x_p} \qquad \boldsymbol{u = x^T b}$$

Hauptkomponeneten (PC)

$\boldsymbol{U = XB}$ X … projizierte Datenwerte → $(n \times p)$ B … (Spalten = Ladungen der PC → $(p \times p)$)

$\boldsymbol{b_j = \left(b_{1j}, \dots, b_{pj}\right)^T \rightarrow Max\left(Var\left(u_j\right)\right)}$ & $b_j^T b_l = 0$ & $b_j^T b_j = 0$ $\qquad shape\left(b_j\right) = p \times 1$

$$\boldsymbol{Var\left(u_j\right) = Var\left(x_1 b_{1j} + \cdots + x_p b_{pj}\right) = b_j^T Cov\left(x_1, \dots, x_p\right) b_j = b_j^T \Sigma b_j}$$

$\Sigma .. theo. Kovarianzmatrix$

Lagrange Problem $\boldsymbol{\phi_j = b_j^T \Sigma b_j - \lambda_j\left(b_j^T b_j - 1\right)}$ j = 1 bis p λ_j …$Langrange\ Parameter$

$$\boldsymbol{Var\left(u_j\right) = b_j^T \Sigma b_j = b_j^T \lambda_j b_j = \lambda_j b_j^T b_j = \lambda_j}$$

Anteil der erste k PCs $\boldsymbol{\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}}$

Clusteranalyse

k-means Zentroide $\overline{\boldsymbol{x}}_j = \frac{1}{n_j}\sum_{i\in I_j} \boldsymbol{x}_i$ j = 1 bis k Zielfunktion: $\sum_{j=1}^{k} n_j \sum_{i\in I_j} \left|\left|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_j\right|\right|^2 \to min$

Hierachische Clustermethoden

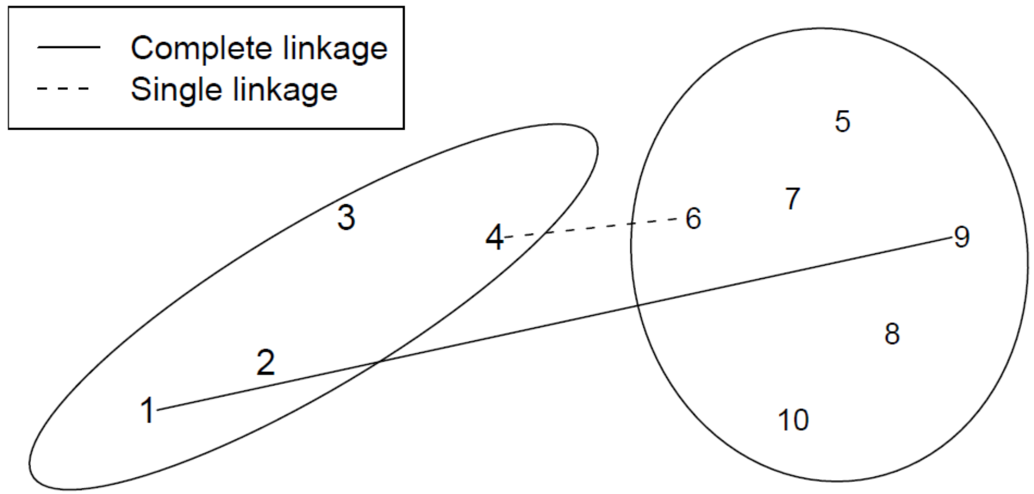
Complete linkage: $\max_{i\in I_j, i'\in I_j} d(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$

Single linkage: $\min_{i\in I_j, i'\in I_j} d(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$

Average linkage: $\operatorname*{average}_{i\in I_j, i'\in I_j} d(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$

Centroid Methode: $d\left(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}}_{j'}\right)$

Ward Methode: $d\left(\overline{\boldsymbol{x}}_j - \overline{\boldsymbol{x}}_{j'}\right) \frac{\sqrt{2n_j n_{j'}}}{\sqrt{n_j + n_{j'}}}$



Fuzzy Clustering

c-means Zentroide $\overline{\boldsymbol{x}}_j = \frac{\sum_{i=1}^{n_j} u_{ij}^2 x_i}{\sum_{i=1}^{n_j} u_{ij}^2}$ Zielfunktion: $\sum_{j=1}^{k} \sum_{i\in I_j} u_{ij}^2 \left|\left|\boldsymbol{x}_i - \overline{\boldsymbol{x}}_j\right|\right|^2 \to min$ $u_{ij}..Zugehörigkeit$

Gütemaße

Within-Cluster Sum-of-Squares (für Varianz)

$$\boldsymbol{W_k = \sum_{j=1}^{k} \sum_{i \in I_j} \left\|\bar{x}_j - \bar{x}\right\|^2}$$

Between-Cluster Sum-of-Squares (für Heterogenität)

$$\boldsymbol{B_k \sum_{j=1}^{k} \left\|\bar{x}_j - \bar{x}\right\|^2 mit}\ \bar{x} = \frac{1}{k}\sum_{j=1}^{k} \bar{x}_j$$

Calinski-Harabasz-Index

$$\boldsymbol{CH_k = \frac{B_k/(k-1)}{W_k(n-k)}}$$

Hartigan-Index

$$\boldsymbol{H_k = log\left(\frac{B_k}{W_k}\right)}$$

Diskriminanzanalyse

$$\boldsymbol{\phi_j(x) = \frac{1}{\sqrt{(2\pi)^p det(\Sigma_j)}} exp\left\{-\frac{(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}{2}\right\}}$$ p … Features, Σ ... Kovarianz, j ... Gruppe,

ϕ_j .. Dichtefunktion pro Gruppe

$$\boldsymbol{P(G=j|x) = \frac{\phi_j(x)p_j}{\sum_{l=1}^{k}\phi_l(x)p_l}}$$ pj ... a-priori Wahrscheinlichkeit pro Gruppe. G … Gruppenzugehörigkeit

x wird j zugeteilt wenn gilt: $\log\left(\frac{P(G=j|x)}{P(G=l|x)}\right) = \log\left(\frac{\phi_j(x)p_j}{\phi_l(x)p_l}\right) = \log\left(\frac{\phi_j(x)}{\phi_l(x)}\right) + \log\left(\frac{p_j}{p_l}\right) > 0$

Lineare Diskriminanzanalyse (LDA)

es gilt: $\boldsymbol{\Sigma_1 = \cdots = \Sigma_k = \Sigma}$

lineare Diskriminanzfunktion:

$$\boldsymbol{\delta_j(x) = x^T\,\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \log p_j}$$

Quadratische Diskriminanzanalyse (QDA)

$$\boldsymbol{\delta_j^{(q)}(x) = -\frac{1}{2}log\left(det(\Sigma_j)\right) - \frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j) + log(p_j)}$$