Formelsammlung zur multivariaten Statistik (Jens Mehrhoff)

R-Techniken und Q-Techniken

- Datenmatrix: $\mathbf{X}$, $[n \times m]$
  (n Zeilen: Objekte/Personen, m Spalten: Merkmale/Variablen)
- $\mathbf{R}$-Techniken: $\mathbf{R} = \frac{1}{n}\mathbf{Z'Z}$, $[m \times m]$, $z_{ij} = \frac{x_{ij} - \overline{x_i}}{s_{x_{ij}}}$
  (Matrix der Korrelationskoeffizienten zwischen Merkmalen/Variablen)
- $\mathbf{Q}$-Techniken: $\mathbf{Q} = \frac{1}{m}\mathbf{XX'}$, $[n \times n]$
  (Matrix der Korrelationskoeffizienten zwischen Objekten/Personen)

Hotelling's T^2: nominale Inputvariable, mehrere metrische Outputvariablen, Hypothesentest

- Einstichprobenfall: $T^2 = n\left(\overline{\mathbf{x}} - \boldsymbol{\mu}_0\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{x}} - \boldsymbol{\mu}_0\right)$
- Zweistichprobenfall: $T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2}\right)'\mathbf{S}^{-1}\left(\overline{\mathbf{x}_1} - \overline{\mathbf{x}_2}\right)$, $\mathbf{S} = \frac{n_1 - 1}{n-2}\mathbf{S}_1 + \frac{n_2 - 1}{n-2}\mathbf{S}_2$

Diskriminanzanalyse $(p = 2, m = 2)$: Q-Technik, Dependenzanalyse, Klasseneinteilung

- Elemente: $k = 1,2,\ldots,n_j$ $\left(x_{kji}\right)$
- Gruppen: $j = 1,2,\ldots,p$
- Merkmale: $i = 1,2,\ldots,m$
- Diskriminanzfunktion: $y = a_1 x_1 + a_2 x_2$ $\left(x_2 = \frac{y^*}{a_2} - \frac{a_1}{a_2}x_1\right)$
- Klassifikationsregel: $y \begin{cases} \geq y^* & \text{Objekt in Gruppe 1} \\ < y^* & \text{Objekt in Gruppe 2} \end{cases}$, $y^* = \frac{n_1}{n}\overline{y_1} + \frac{n_2}{n}\overline{y_2}$
  $\overline{y_1} = a_1\overline{x_{11}} + a_2\overline{x_{12}}$, $\overline{y_2} = a_1\overline{x_{21}} + a_2\overline{x_{22}}$
- Normalgleichungen: $a_1 = \frac{d_1 s_2^2 - d_2 s_{12}}{s_1^2 s_2^2 - s_{12}^2} = \frac{\frac{d_1}{s_1^2} - r_{12}\frac{d_2}{s_1 s_2}}{1 - r_{12}^2}$, $a_2 = \frac{d_2 s_1^2 - d_1 s_{12}}{s_1^2 s_2^2 - s_{12}^2} = \frac{\frac{d_2}{s_2^2} - r_{12}\frac{d_1}{s_1 s_2}}{1 - r_{12}^2}$
  $d_1 = \overline{x_{11}} - \overline{x_{21}}$, $d_2 = \overline{x_{12}} - \overline{x_{22}}$
  $s_1^2 = \frac{1}{n}\left[n_1\sum_{k_1=1}^{n_1}\left(x_{k_1 11} - \overline{x_{11}}\right)^2 + n_2\sum_{k_2=1}^{n_2}\left(x_{k_2 21} - \overline{x_{21}}\right)^2\right]$
  $s_2^2 = \frac{1}{n}\left[n_1\sum_{k_1=1}^{n_1}\left(x_{k_1 12} - \overline{x_{12}}\right)^2 + n_2\sum_{k_2=1}^{n_2}\left(x_{k_2 22} - \overline{x_{22}}\right)^2\right]$
  $s_{12} = \frac{1}{n}\left[n_1\sum_{k_1=1}^{n_1}\left(x_{k_1 11} - \overline{x_{11}}\right)\left(x_{k_1 12} - \overline{x_{12}}\right) + n_2\sum_{k_2=1}^{n_2}\left(x_{k_2 21} - \overline{x_{21}}\right)\left(x_{k_2 22} - \overline{x_{22}}\right)\right]$, $r_{12} = \frac{s_{12}}{s_1 s_2}$

Pfadanalyse: Analyse von Kausalstrukturen (nur Falsifikation)

- Gleichungen:

$$\begin{array}{ccccc} x & \rightarrow & y & \rightarrow & z \\ & & \uparrow & & \uparrow \\ & & u & & v \end{array}$$

$(1)\ \ y = p_{yx}x + u$

$(2)\ \ z = p_{zy}y + v$

- Korrelationskoeffizienten

$\mathrm{E}[(1)\cdot x]: r_{xy} = \mathrm{E}(xy) = p_{yx}\mathrm{E}(x^2) + \mathrm{E}(ux) = p_{yx}$

$\mathrm{E}[(2)\cdot y]: r_{yz} = \mathrm{E}(yz) = p_{zy}\mathrm{E}(y^2) + \mathrm{E}(vy) = p_{zy}$

$\mathrm{E}[(2)\cdot x]: r_{xz} = \mathrm{E}(xz) = p_{zy}\mathrm{E}(xy) + \mathrm{E}(vx) = r_{xy}r_{yz}$

- Bestimmtheit und Unbestimmtheit

$\mathrm{E}[(1)\cdot u]: \mathrm{E}(uy) = p_{yx}\mathrm{E}(ux) + \mathrm{E}(u^2) = \sigma_u^2 = 1 - r_{xy}^2$

$\mathrm{E}[(2)\cdot v]: \mathrm{E}(vz) = p_{zy}\mathrm{E}(vy) + \mathrm{E}(v^2) = \sigma_v^2 = 1 - r_{yz}^2$

Faktorenanalyse: R-Technik, Interdependenzanalyse, metrische Inputvariable, latente Outputvariable, Dimensionsreduktion

- Objekte/Personen: $k = 1,2,\ldots,n$
- Merkmale/Variablen: $i = 1,2,\ldots,m$
- Faktoren: $j = 1,2,\ldots,p$
- $\mathbf{Z = YF'}$

  $\mathbf{Y} = \mathbf{ZF}\left(\mathbf{F'F}\right)^{-1}$, $[n \times p]$: Faktorwerte (unkorrelierte Faktoren)

  $\mathbf{F}$, $[m \times p]$: Faktorladungen der Merkmale/Variablen in den Faktoren

- Fundamentaltheorem: $\mathbf{R} = \frac{1}{n}\mathbf{Z'Z} = \frac{1}{n}\mathbf{FY'YF'} = \mathbf{FF'}$ $\left(\frac{1}{n}\mathbf{Y'Y} = \mathbf{I}\right)$

  $\mathbf{R}$, $[m \times m]$: Korrelation zwischen Merkmalen/Variablen

- Extraktion der Faktoren: $(\mathbf{R} - \lambda_1\mathbf{I})\mathbf{f}_1 \overset{!}{=} 0$, $[(\mathbf{R} - \mathbf{f_1 f_1'}) - \lambda_2\mathbf{I}]\mathbf{f_2} = (\mathbf{R_{(1)}} - \lambda_2\mathbf{I})\mathbf{f_2} \overset{!}{=} 0$, …
- Eigenwertproblem: $|\mathbf{R} - \lambda_1\mathbf{I}| \overset{!}{=} 0$, $|\mathbf{R_{(1)}} - \lambda_2\mathbf{I}| \overset{!}{=} 0$, …

  (nicht-triviale Lösungen eines homogenen linearen Gleichungssystems)

  λ_1, λ_2, …: Eigenwerte

  $\mathbf{f_1}$, $\mathbf{f_2}$, …: Eigenvektoren

- Kommunalitätenproblem: $r_{ii} = 1 - u_{ii} = h_i^2$ (u_{ii}: Varianzanteil des Einzelrestfaktors, h_i^2: durch die gemeinsamen Faktoren erklärter Anteil der Gesamtvarianz eines Merkmals/einer Variablen)

  Hauptkomponentenanalyse: $\mathbf{R = FF'}$ (Zusammenfassung von auf einen Faktor hoch ladenden Merkmalen/Variablen)

  Faktorenanalyse: $\mathbf{R^* = R - U}$ (Ursachenbezeichnung für hohe Ladungen von Merkmalen/Variablen auf einen Faktor)

- Abbruchkriterium: Scree-Test (Anordnung der Eigenwerte der Größe nach)
- Rotationsproblem: $\mathbf{T} = \begin{bmatrix} \cos\delta & -\sin\delta \\ \sin\delta & \cos\delta \end{bmatrix}$ (Drehung gegen den Uhrzeigersinn)

$\mathbf{F^* = FT}$, $\mathbf{R^* = F^*F^{*\prime} = FTT'F' = FF'}$ $\left(\mathbf{TT' = I}\right)$

Latent Structure Analysis (Latent Dichotomy): R-Technik, dichotome Inputvariable, latente Outputvariable, Klasseneinteilung

- relative Klassengrößen: V_I, V_{II}
- bedingte Wahrscheinlichkeiten für positive Ausprägung von i, j: p_{Ii}, p_{IIi}, p_{Ij}, p_{IIj} (Latent Marginals)
- lokale Unabhängigkeit: $p_{Iij} = p_{Ii}p_{Ij}$, $p_{IIij} = p_{IIi}p_{IIj}$, $p_{Iijk} = p_{Ii}p_{Ij}p_{Ik}$, $p_{IIijk} = p_{IIi}p_{IIj}p_{IIk}$
- Accounting Equations
  $V_I + V_{II} = 1$, $p_i = V_I p_{Ii} + V_{II} p_{IIi}$, $p_{ij} = V_I p_{Iij} + V_{II} p_{IIij}$, $p_{ijk} = V_I p_{Iijk} + V_{II} p_{IIijk}$
- Kreuzprodukte (Kovarianz): $|ij| = p_{ij} - p_i p_j$
- geschichtete Kreuzprodukte (bedingte Kovarianz): $|ij,k| = p_k p_{ijk} - p_{ik} p_{jk}$
- Steigungen der Trace Lines (Regressionslinien): $b_i = \sqrt{\frac{|ij||ik|}{|jk|}}$
- Hilfsparameter: $\mu_k = p_{Ik} p_{IIk} = \frac{|ij,k|}{|ij|}$
- quadratische Gleichung: $t^2 + \phi_i t - 1 \overset{!}{=} 0$, $\phi_i := \frac{\mu_i}{p_i b_i} - \frac{p_i}{b_i} + \frac{b_i}{p_i}$ (Schiefe)
- Anteilswerte der latenten Klassen: $t = \sqrt{\frac{V_I}{V_{II}}} = \sqrt{\frac{V_I}{1 - V_I}}$, $V_I = \frac{t^2}{1 + t^2}$, $V_{II} = 1 - V_I$
- Position auf der latenten Variablen: $x_I = \frac{1}{t}$, $x_{II} = -t$
- Latent Marginals: $p_{Ii} = p_i + b_i x_I = p_i + \frac{b_i}{t}$, $p_{IIi} = p_i + b_i x_{II} = p_i - b_i t$
- Response Pattern (Reproduktion der Daten): $n_I p_{I1} p_{I2} (1 - p_{I3}) = n V_I p_{I1} p_{I2} (1 - p_{I3})$ (Beispiel: $++-$, Klasse I)

logistische Regression

- logistische Verteilung: $k = \mathrm{E}(y \mid x) = \boldsymbol{\beta}\mathbf{x_i}$
- Wahrscheinlichkeit für positive Ausprägung: $\mathrm{W}(X = 1) = \frac{\exp(K)}{1 + \exp(K)} = p$
- Wahrscheinlichkeit für negative Ausprägung: $\mathrm{W}(X = 0) = \frac{1}{1 + \exp(K)} = 1 - p$
- Odd Ratio: $\frac{\mathrm{W}(X = 1)}{\mathrm{W}(X = 0)} = \frac{p}{1 - p} = \exp(K)$
- Logit: $\ln \frac{p}{1 - p} = k$

Data Mining

- Minkowski r-Metriken: $d_r(\mathbf{x_i}, \mathbf{x_j}) = \left(\sum_{l=1}^{p} |x_{il} - x_{jl}|^r \right)^{\frac{1}{r}}$ (nicht skaleninvariant)
- Mahalanobis-Distanz: $d_M(\mathbf{x_i}, \mathbf{x_j}) = \sqrt{(\mathbf{x_i} - \mathbf{x_j})' \mathbf{K}^{-1} (\mathbf{x_i} - \mathbf{x_j})}$, $\mathbf{K} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})'$