Multivariate Verfahren

Oliver Muthmann

31. Mai 2007

Gliederung

Komplexe Systeme werden meist durch mehrere Faktoren gleichzeitig beinflusst, und nicht immer ist es möglich, den Einfluss einzelner Faktoren bei konstanten anderen Faktoren zu messen. (Beispiel Gehirn)

- Wann sind die Mittelwerte zweier Stichproben (Datenmatrizen) verschieden?
- Welche Beziehung besteht zwischen zwei Messgrößen?
- Wie kann man die Datenmenge reduzieren ohne wesentliche Information zu verlieren?

Datenmatrix:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

n: # der Messungen
p: # der beobachteten Größen

Univariater Fall: ANOVA

- Test auf Gleichheit der Mittelwerte von g Stichproben bezüglich einer Dimension
- H_0: Mittelwerte sind gleich
- Schätze Varianz zwischen den Mittelwerten der Stichproben und innerhalb der Stichproben
- Bilde den Quotienten, dieser ist bei Gültigkeit der Nullhypothese F(g-1,(n-1)g)-verteilt

Multivariater Fall: MANOVA

- p-dimensionale Beobachtungsgrößen
- g Stichproben mit Erwartungswerten
  $\vec{\mu_k} = \vec{\mu} + \vec{\alpha_k} = (\ \mu_{1k} \quad \cdots \quad \mu_{pk}\ )^T; \quad k = 1, \cdots, g$
- $\vec{x_{ik}} = \vec{\mu} + \vec{\alpha_k} + \epsilon_{ik}$
- $H_0 : \ \vec{\alpha_1} = \vec{\alpha_2} = \ldots = \vec{\alpha_g} = 0$
- Annahme: Kovarianzmatrix für alle Stichproben gleich; Stichproben normalverteilt

- Matrix der Summe der quadrat. Abweichungen innerhalb der Stichproben:

$$\mathbf{W} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ik} - \bar{x_k})(x_{ik} - \bar{x_k})^T$$

- ... zwischen den Stichproben:

$$\mathbf{G} = \sum_{k=1}^{g} \sum_{i=1}^{n_k} (\bar{x_k} - \bar{x})(\bar{x_k} - \bar{x})^T$$

- $\Lambda = |\mathbf{W}|/|\mathbf{W} + \mathbf{G}|$
- Unter der Nullhypothese entspricht $\mathbf{W} + \mathbf{G}$ der Gesamtvarianz und $F = \frac{1-y}{y} \frac{m_2}{m_1} = \left(\frac{|G|}{|W|} \right)^{1/s} \frac{m_2}{m_1}$ ist näherungsweise F-verteilt mit m_1 und m_2 Freiheitsgraden, wobei

$y = \Lambda^{1/s}; \quad s = \sqrt{\frac{p^2(g-1)^2 - 4}{p^2 + (g-1)^2 - 5}}$

$m_1 = p(g-1); \quad m_2 = s[n - (p+g+2)/2] - \frac{p(g-1)}{2} + 1$

univariate (lineare) Regression

- n Realisierungen y_i einer Zufallsvariablen $Y(x_i)$
- Erklärung der Varianz von Y durch die Stichprobenwerte x_i und einen Rauschterm:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{mit } E_i = N(0, \sigma)$$

- Schätzung der Parameter mit der Maximum-Likelihood Methode

$$\rightsquigarrow \hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i y_i - \bar{x}\bar{y})}{\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1$$

multivariate (lineare) Regression

- Datenmatrix
  $$\mathbf{X} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{pmatrix} \quad \mathbf{x_1} \in M(q \times n) \quad \mathbf{x_2} \in M(s \times n)$$
- Umdefinition: $\mathbf{X} = \begin{pmatrix} \vec{1}^T \\ \mathbf{x_1} \end{pmatrix}^T \quad \mathbf{Y} = \mathbf{x_2}^T$
- Ziel: Erklärung von Y durch X

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}$$

- Schätzung von B:

$$\mathbf{B} = \begin{pmatrix} \vec{\beta_0}^T \\ \mathbf{B}^* \end{pmatrix} \quad \text{wobei}$$

$$E[\vec{x}_2|\vec{x}_1]^T = (\vec{\mu}_2 - \Sigma_{21}\Sigma_{11}^{-1}\vec{\mu}_1)^T + \vec{x}_1^T\Sigma_{11}^{-1}\Sigma_{12} = \vec{\beta_0}^T + \mathbf{B}^*$$

- Maximum-Likelihood Schätzer für $\vec{\beta_0}^T$; $\mathbf{B}^*$:

$$\hat{\vec{\beta}}_0^T = \vec{\bar{y}}^T - \vec{\bar{x}}_1^T \mathbf{S}_{x_1 x_1}^{-1} \mathbf{S}_{x_1 y}$$
$$\hat{\mathbf{B}}^* = \mathbf{S}_{x_1 x_1}^{-1} \mathbf{S}_{x_1 y}$$

- Schätzer ist äquivalent zu denen aus der univariaten multiplen Regression für jede Spalte von **y**:

$$\vec{y}_j = \mathbf{X}\vec{\beta}_j + \vec{u}_j \quad j = 1, \ldots, s$$

- Schätzung der Fehlerkovarianzmatrix:

$$\hat{\mathbf{U}^2} = \hat{\Sigma_{yy|x_1}} = \hat{\Gamma} = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})/(n - q - 1)$$

Nichtdiagonalelemente verschwinden nicht!

Faktorenanalyse

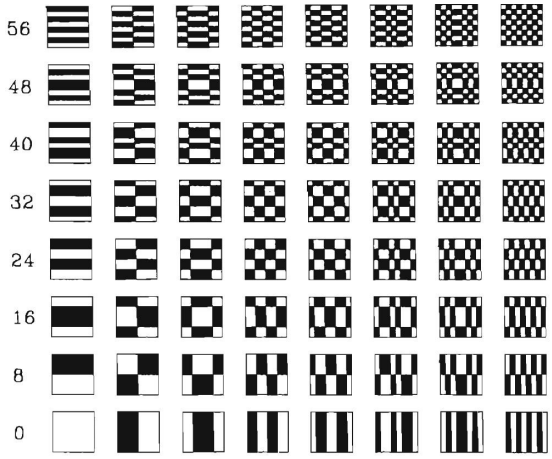
Ziel: Reduktion der Datenmenge/Dimension durch Einführung von neuen Variablen, den sogenannten *Faktoren*.
Extraktion der Faktoren kann durch verschiedene Verfahren erreicht werden, am gängigsten ist die Hauptkomponentenanalyse.

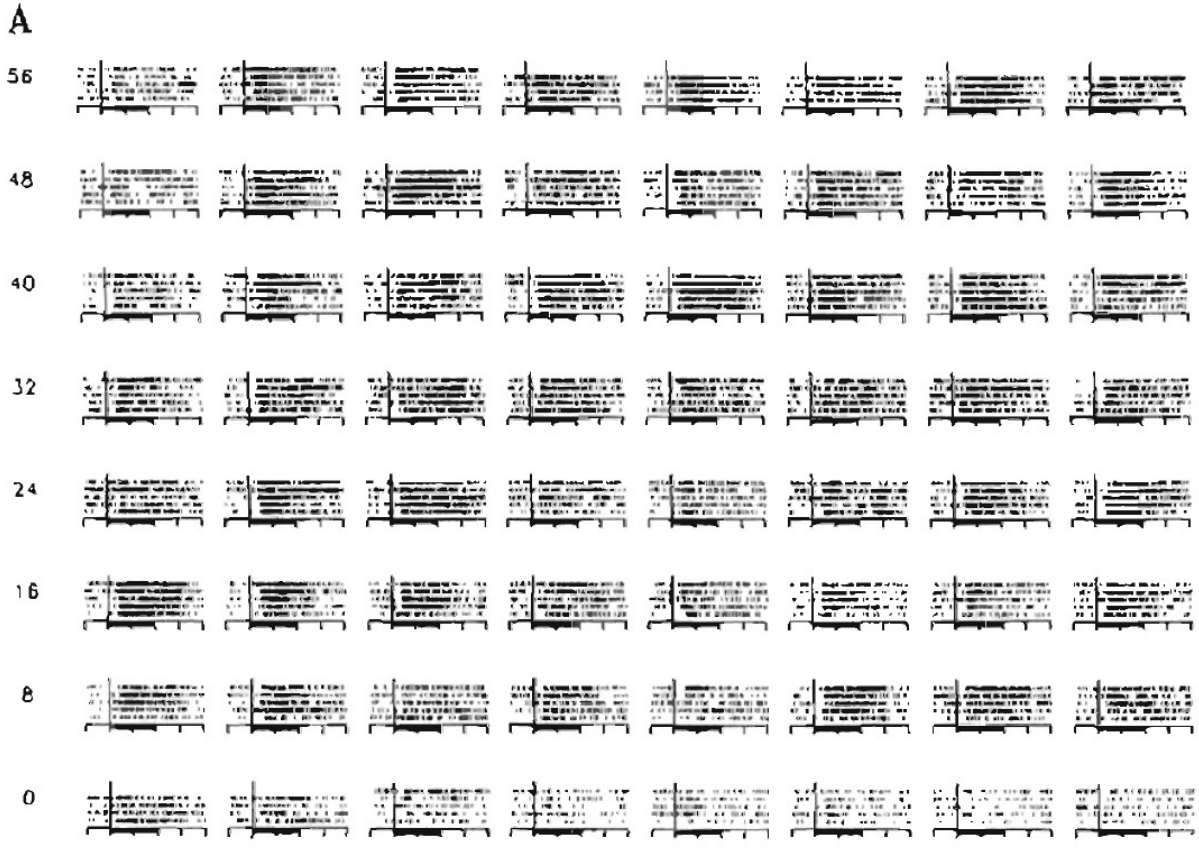
Hauptkomponentenanalyse (PCA, auch: Karhunen-Loeve-Transformation)

- Idee: Hauptachsentransformation der (geschätzten) Korrelationsmatrix
  $\rightsquigarrow$ orthogonale Eigenvektoren, sortiert nach deren Varianz.
- Die Varianz kann als Maß für den Informationsgehalt angesehen werden
  $\rightsquigarrow$ nehme nur die Eigenvektoren mit der größten Varianz!

Beispiel: Mustererkennung bei Rhesusaffen

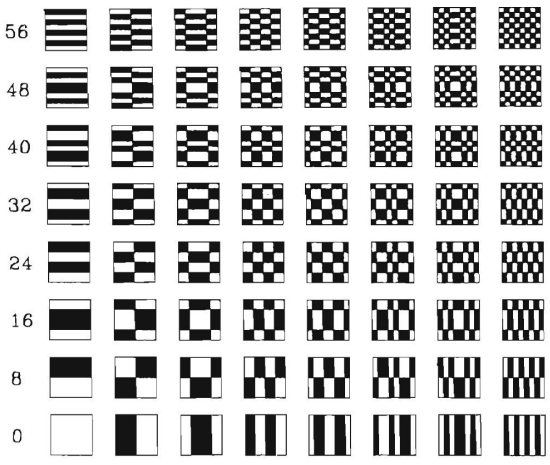
Richmond, Optican 1987

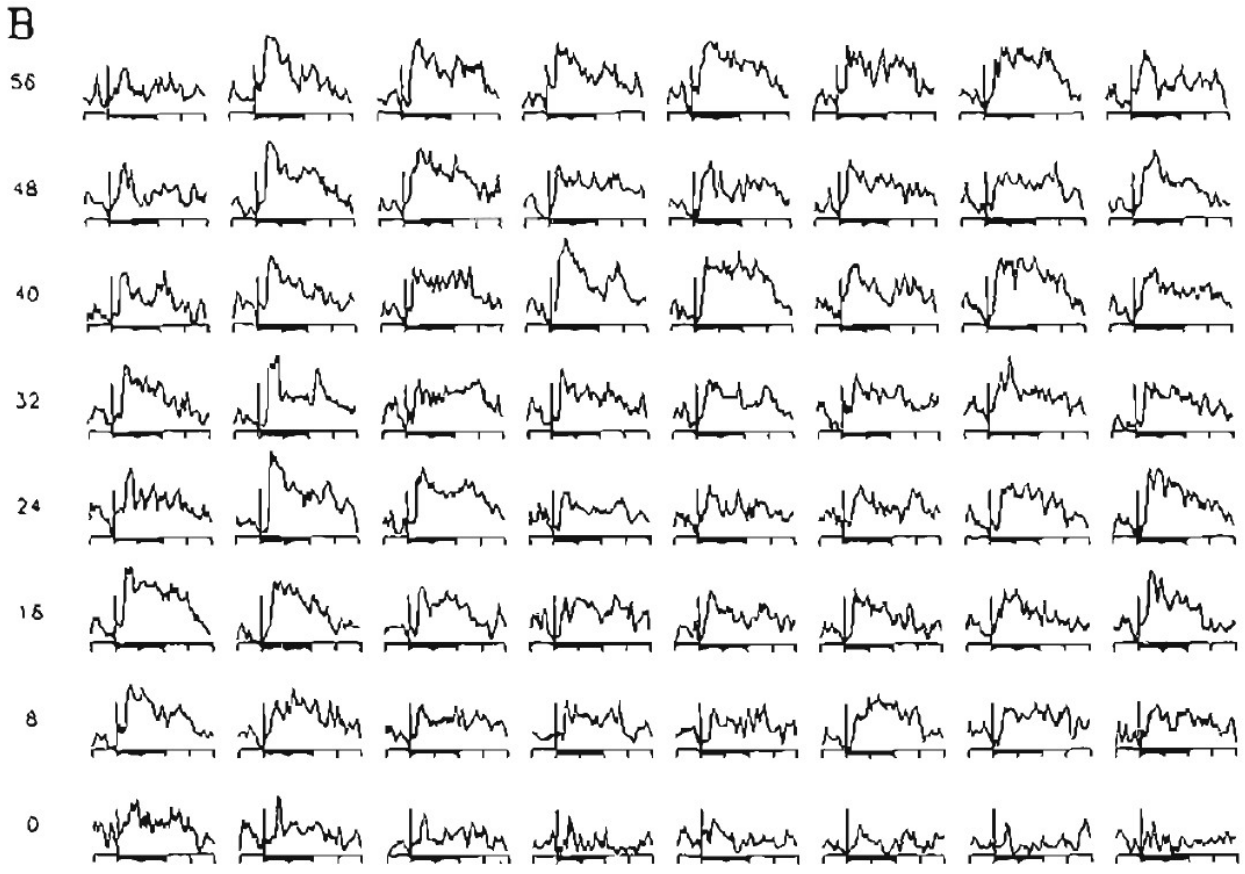




Spikes eines Neuron

Beispiel: Mustererkennung bei Rhesusaffen

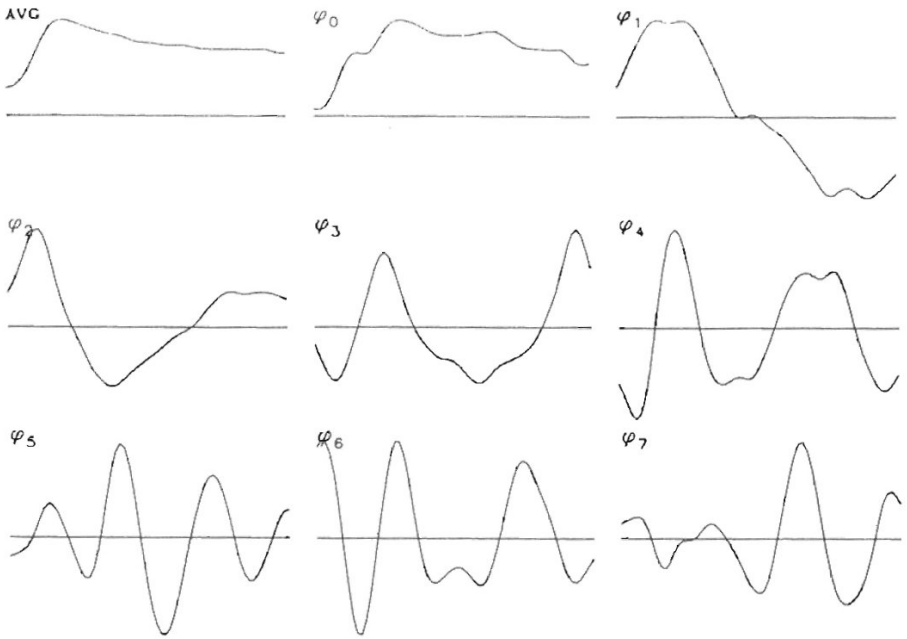




Aufgaben:

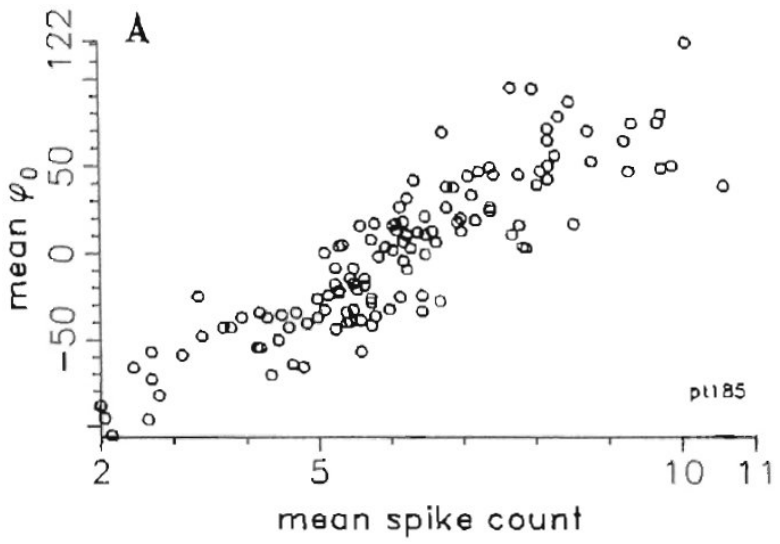
- Welche Hauptkomkonenten tragen keine (signifikante) Information?
  - Vergleich mit ursprünglicher Korrelation
  - Scree Test
  - Broken Stick Model
- Interpretation der Hauptkomponenten?
- Aufstellen von neuen Hypothesen für die Hauptkomponenten

gemittelte Spikedichte
und die ersten Hauptkomponenten



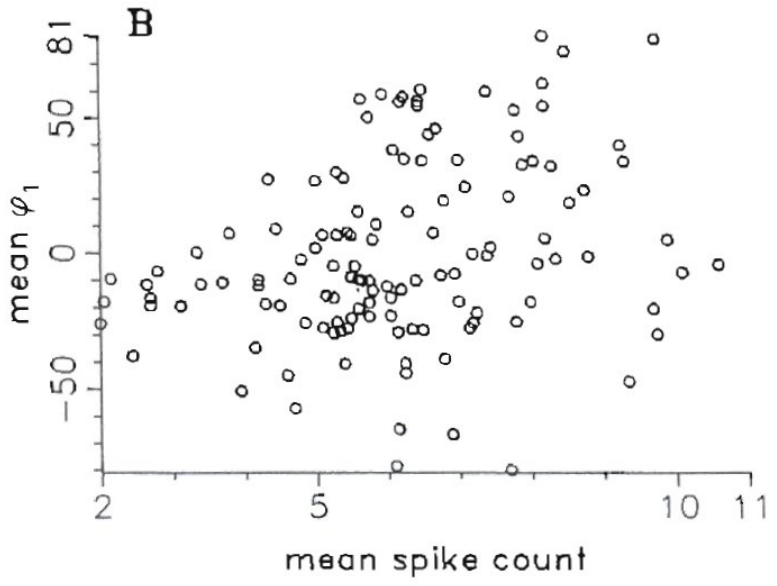
Korrelation mit der gemittelten Spikezahl

1. Hauptkomponente



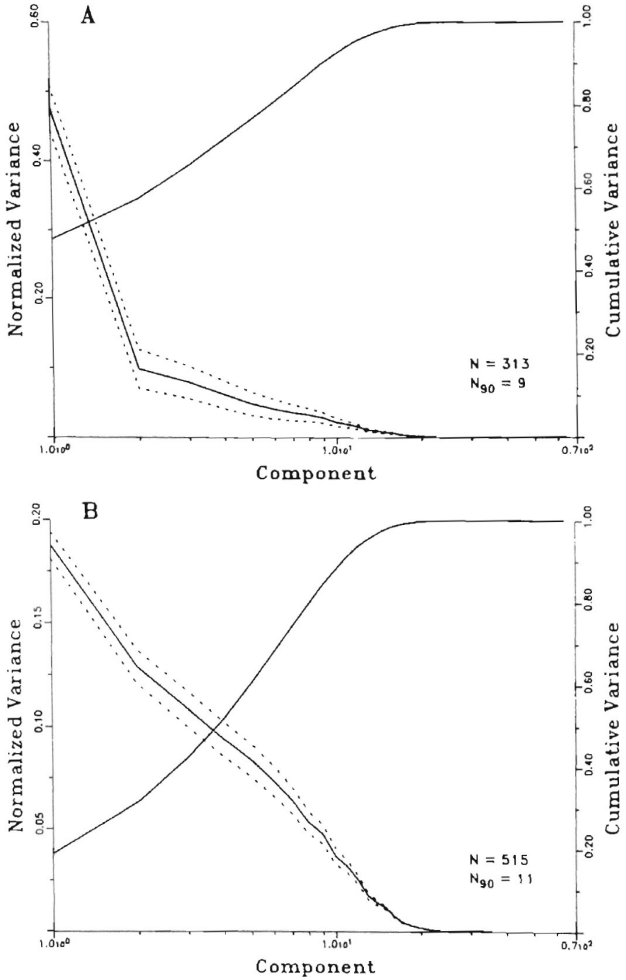
(r = 0.89)

2. Hauptkomponente

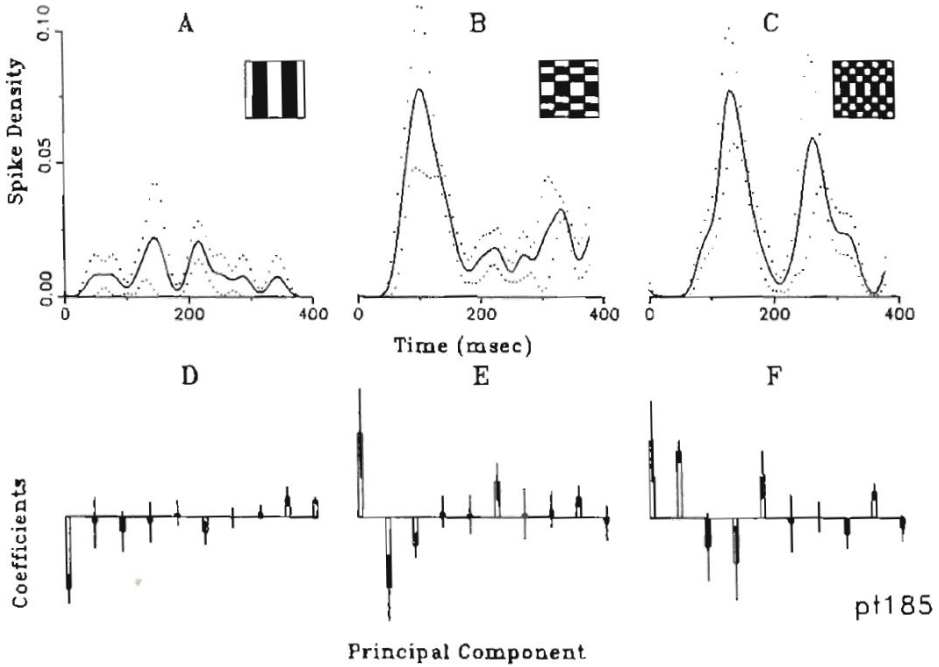


(r = 0.19)

Varianz der Hauptkomponenten



Beiträge der einzelnen Hauptkomponenten (=Faktorladungen)



Faktorenanalyse

- Definiere einen Satz gemeinsamer, orthonormierter Faktoren $F_1, \ldots, F_r$ so dass

$$(\vec{x} - \vec{\mu}) = \mathbf{A}\vec{f} + \vec{u}$$

- u_i unkorreliert mit $f_j \rightsquigarrow \Sigma = \mathbf{A}\mathbf{A}^T + \psi$
- Varianz von X_i:

$$\sigma_i^2 = \sum_{j=1}^{r} a_{ij}^2 + \sigma_{u_i}^2$$

- Kommunalität: $\sum_{j=1}^{r} a_{ij}^2$
- Die Korrelationsmatrix ist invariant unter orthogonalen Transformationen:

$$\Sigma = \mathbf{A}\mathbf{A}^T + \psi = \mathbf{A}\mathbf{T}\mathbf{T}^T\mathbf{A}^T + \psi \quad \rightsquigarrow \mathbf{B} = \mathbf{A}\mathbf{T}$$

Clusteranalyse

Betrachte $\mathbf{X}\mathbf{X}^T$: große Nichtdiagonalelemente bei ähnlichen Variablen (verschwindende bei orthogonalen Variablen)
Ziel: Reduktion der Datenmenge durch Zusammenfassen von Messgrößen zu Clustern.
Verfahren:

- hierarchisches Clustern
- k-means-Algorithmus
- fuzzy clustering

Hierarchisches Clustern

Abstandsmaße zwischen Variablen:

- standardisierter euklidische Metrik $(\vec{x_r} - \vec{x_s})^T S^{-1} (\vec{x_r} - \vec{x_s})$
- Mahalanobis Distance $(\vec{x_r} - \vec{x_s})^T S^{-1} (\vec{x_r} - \vec{x_s})$
- Minkowski Metrik $m_{rs} = \left[\sum_{j=1}^{p} |x_{rj} - x_{rs}|^{\lambda} \right]^{1/\lambda}$

Abstandsmaße zwischen Gruppen:

- Single Linkage (Nearest Neighbor)
- Complete Linkage (Furthest Neighbor)
- Average Linkage

- Aufstellung einer Abstandsmatrix
- Zusammenfassen der nächstliegenden Variablen zu einem Cluster
- Berechnung einer neuen Abstandsmatrix mit
  $p_{tu} = \alpha_r p_{ru} + \alpha_s p_{su} + \beta p_{rs} + \gamma |p_{ru} - p_{rs}|$
  Die Koeffizienten hängen von der verwendeten Metrik ab.

Ist die Aufteilung signifikant?

- externe Kriterien: Vergleich mit bekannten Clustern
- interne Kriterien: passt die Lösung zur Abstandsmatrix?
- Reproduzierbarkeit: mit halber Datenanzahl?
- Verwendung anderer Algorithmen

Zusammenfassung

- Im Multivariaten hat man große Datenmengen $\rightsquigarrow$ Herausfiltern der relevaten Daten durch
  - *Faktoranalyse*
  - *Clusteranalyse*
- Test ob 2 Datensätze sich in den Mittelwerten unterscheiden: *Varianzanalyse*
- Wenn ein linearer Zusammenhang zwischen den Datensätzen besteht: *Regressionsanalyse*

Literaturangaben

J.D. Jobson
Applied Multivariate Data Analysis; Volume 2
Springer-Verlag

Richmond, Podell, Optican, Spitzer
Temporal encoding of two-dimensional patterns by single units in inferior temporal cortex.
1.Response characteristics
2.Quantification of response waveform
3.Information theoretic analysis

Journal of Neuroscience 57:132-178