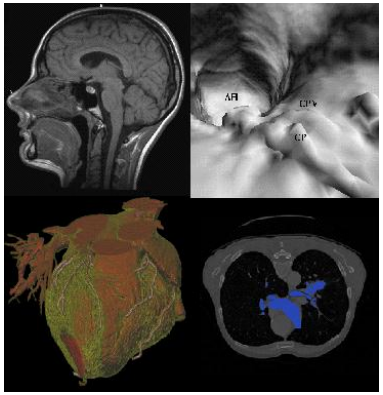
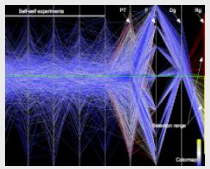
Informations-visualisierung

Thema:	3. Statistische Grundlagen
Dozent:	Prof. Dr. Gerik Scheuermann
	scheuermann@informatik.uni-leipzig.de
Sprechstunde:	nach Vorlesung
Umfang:	2
Prüfungsfach:	Modul Fortgeschrittene Computergrafik Medizininformatik, Angewandte Informatik

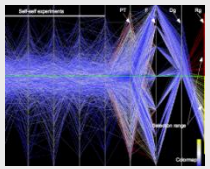


UNIVERSITÄT LEIPZIG



Übersicht

1. Einführung
2. Wahrnehmung von Grafik
3. **Statistische Grundlagen**
4. Darstellung von Tabellen
5. Darstellung von Graphen
6. Darstellung von Mediadaten und Prozessen
7. Interaktion
8. Visual Analytics

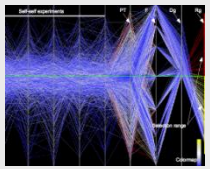


3.0 Multivariate Statistik

Informationsvisualisierung zielt auf das **Erkennen von Zusammenhängen** zwischen Variablen bei einer großen Menge von gegebenen Werten oder die **Überprüfung vermuteter Zusammenhänge**. Dazu erfolgt eine Abbildung der Variablen in graphischen Darstellungen, um das visuelle System des Menschen zur Zusammenhangssuche verwenden zu können.

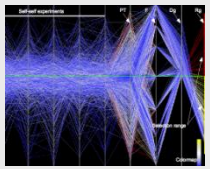
Die **klassischen Ansätze** zur Analyse von mehreren Variablen bei einer großen Menge von Werten gehören in den Bereich der **multivariaten Statistik**. Da häufig gerade das Erkennen unbekannter Zusammenhänge in der Informationsvisualisierung eine Rolle spielt, wird in diesem Kapitel ein wichtigstes **strukturerkennendes Verfahren** der multivariaten Statistik vorgestellt, da es zuweilen als Ergänzung oder Vorbereitung von Informationsvisualisierungen eingesetzt wird.

Es sei darauf hingewiesen, dass hier stets von **ganzzahligen oder reellen Variablen** ausgegangen wird. Es gibt durchaus auch statistische Ansätze für ordinale oder nominale Variablen!



- Bevor wir zu den einzelnen Verfahren kommen, benötigen wir eine geeignete Notation, die auf Matrizen beruht.
- Wenn wir m **Datensätze (Beobachtungen)** $d_1,...,d_m$ mit n **Komponenten (Variablen)** $x_1,...,x_n$ haben, so bilden wir die **Datenmatrix (Beobachtungsmatrix)** X:

$$X := \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ . & \cdots & . \\ . & \cdots & . \\ x_{m1} & \cdots & x_{mn} \end{pmatrix}$$



Mit Hilfe der Mittelwerte $\bar{x}_j := \frac{1}{m}\sum_{i=1}^{m} x_{ij}$

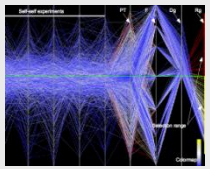
bestimmen wir die **Abweichungsmatrix** M

$$m_{ij} := x_{ij} - \bar{x}_j \qquad \text{mit} \quad \begin{matrix} i = 1,\ldots,m \\ j = 1,\ldots,n \end{matrix}$$

was sich mit Hilfe des Mittelwertvektors $\bar{x}$ und des m-Vektors 1

$$\bar{x} := (\bar{x}_1,\ldots,\bar{x}_n) \quad 1 := (1,\ldots,1)^T$$

als $$M = X - \frac{1}{m}\left(11^T X\right)$$ schreiben lässt.



Daraus leiten wir die **Abweichungsquadratematrix** T ab:

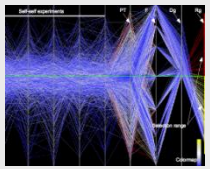
$$t_{jk} := \sum_i m_{ij} m_{ik} \qquad j,k = 1,\ldots,n$$

$$= \sum_i \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$$

also

$$T = M^T M = \begin{pmatrix} t_{11} & \ldots & t_{1n} \\ . & & . \\ t_{n1} & \ldots & t_{nn} \end{pmatrix}$$

die aufgrund ihrer Konstruktion symmetrisch ist.



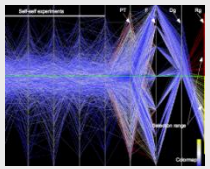
Es gilt:

$$T = M^T M = \left(X^T - \frac{1}{m}\left(X^T 11^T\right)\right)\left(X - \frac{1}{m}\left(11^T X\right)\right)$$

$$= X^T X - \frac{1}{m} X^T 11^T X = X^T \left(I - \frac{1}{m} 11^T \right) X = X^T H X$$

mit

$$H := I - \frac{1}{m} 11^T = \begin{pmatrix} 1-\frac{1}{m} & \frac{-1}{m} & \dots & \frac{-1}{m} \\ \frac{-1}{m} & 1-\frac{1}{m} & . & . \\ . & . & . & \frac{-1}{m} \\ \frac{-1}{m} & \dots & \frac{-1}{m} & 1-\frac{1}{m} \end{pmatrix}$$



Nun kann man die **Kovarianzmatrix** S durch

$$s_{jk} := \frac{1}{m} t_{jk} = \frac{1}{m} \sum \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right) \quad j,k = 1,\ldots,n$$

also

$$S := \frac{1}{m} T = \frac{1}{m} X^T H X$$

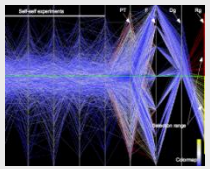
festlegen.

Einige Verfahren nutzen die **Korrelationsmatrix** R $r_{jk} := \frac{t_{jk}}{\sqrt{t_{jj}}\sqrt{t_{kk}}} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$

die mit Hilfe der Diagonalmatrix

$$diagT = \begin{pmatrix} t_{11} & & \\ & \cdot & \\ & & t_{nn} \end{pmatrix}$$

als $R = (diagT)^{-0.5} T (diagT)^{-0.5}$ geschrieben werden kann.



Häufig wird auch von der standardisierten Datenmatrix Z

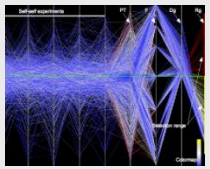
$$z_{ij} := \frac{x_{ij} - \bar{x}_j}{\sqrt{m-1}s_{jj}}$$

also

$$Z := \frac{1}{\sqrt{m-1}} M\left(diagT\right)^{0.5}$$

ausgegangen.

Es gilt $R = Z^T Z$

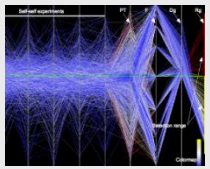


3.1 Hauptkomponentenanalyse

Mit der **Hauptkomponentenanalyse** (principle component analysis, PCA) kann man die n untereinander korrelierten beobachtbaren Variablen $x_1,...,x_n$ auf möglichst wenige p<n untereinander **unkorrelierte Variablen zurückführen**, die **möglichst viel Varianz** auf sich vereinigen.

Dazu werden die standardisierten Variablen z_i in neue Variablen f_l, die sogenannten **Hauptkomponenten**, linear transformiert, die **untereinander unkorreliert und nach fallender Varianz geordnet** sind. Die ersten p<n Hauptkomponenten vereinigen ein Maximum der Gesamtvarianz auf sich.

[Anmerkung: In der Bildverarbeitung gibt es eine entsprechende Bildtransformation. Die Transformation heißt nach ihren Erfindern Karhunen und Loeve auch KL-Transformation.]



3.1 Hauptkomponentenanalyse

Man sucht eine Darstellung der standardisierten Datenmatrix $Z \in R^{mn}$ in der Form

$$Z = FL^T$$

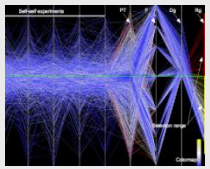
mit der **Faktormatrix** $F \in R^{mn}$ mit orthonormierten Spalten f_l und der **Ladematrix** $L \in R^{mn}$. Es gilt

$$R = Z^TZ = LF^TFL^T=LL^T \quad \text{(F orthogonale Matrix)}$$

Wählt man die p<n ersten normierten Hauptachsen $f_1,...,f_p$ aus, die die wichtigsten „Faktoren“ sind, so nähert man Z an:

$$Z \approx F^{(p)}L^{(p)T} \; mit \; Z \in R^{mn}, F^{(p)} \in R^{mp} orthogonal, L^{(p)} \in R^{np}$$

Dabei nimmt man an, dass die n Variablen $x_1,...,x_n$ linear unabhängig sind, was durch Weglassen von Spalten in X bzw. Z stets erreicht werden kann.



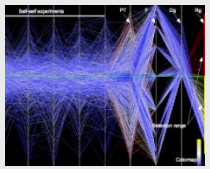
3.1 Hauptkomponentenanalyse

Im Wesentlichen erfordert dies eine **Hauptachsentransformation** der Korrelationsmatrix R. Da R symmetrisch und reell ist, gibt es n Eigenwerte λ_1 , ... , λ_n und orthogonale Eigenvektoren e_1,...,e_n mit

$$Re_j = \lambda_j e_j, \quad |e_j| = 1$$

also

$$R = E\Lambda E^T \quad mit \quad E = (e_1 \ldots e_n), \quad \Lambda := \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$



Die Matrix der Hauptachsen Y wird definiert durch

$$Y := Z\,E$$

und es gilt

$$Y^T Y = E^T Z^T Z E = E^T R E = \Lambda$$

Setzt man nun

$$F := Y\Lambda^{-0.5}, \quad L := E\Lambda^{-0.5}$$

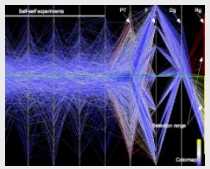
so hat man

$$F^T F = I, \quad Z = FL^T, \quad R = Z^T Z = LL^T$$

Dabei heißt die Zerlegung

$$Z = F^T \Lambda^{0.5} E^T$$

auch **Singulärwertzerlegung** von Z.



Geometrische Deutung

Seien z_1, z_2 gemeinsam normalverteilte Zufallsvariablen, also

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \approx N(0,\Sigma), \quad \Sigma = \begin{pmatrix} 1 & \sigma \\ \sigma & 1 \end{pmatrix}$$

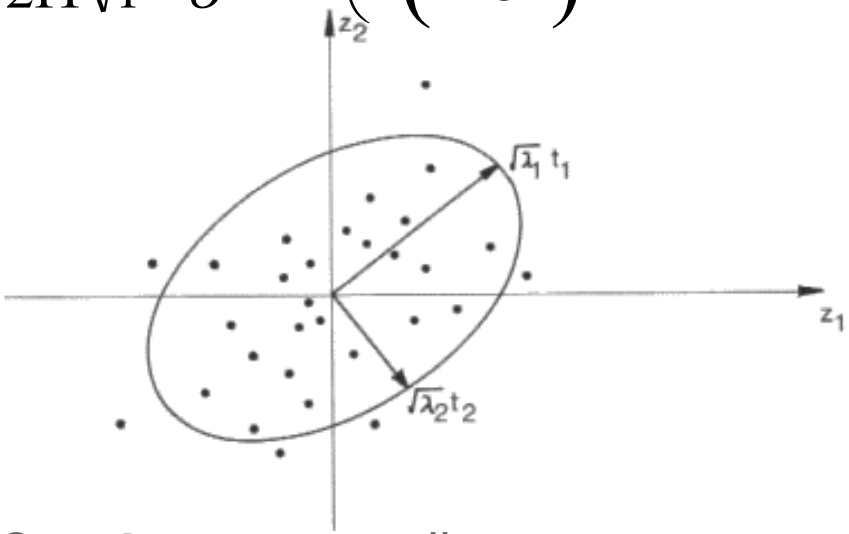
Dann liegen Punkte gleicher Dichte

$$f(z_1,z_2) = \frac{1}{2\Pi\sqrt{1-\sigma^2}} \exp\left(\frac{-1}{2(1-\sigma^2)} \left(z_1^2 - 2\sigma z_1 z_2 + z_2^2 \right) \right)$$

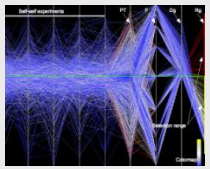
auf Ellipsen um den Nullpunkt.

M Beobachtungen (hier M=30)

geben die elliptische Form wieder:



Die Hauptachse t_1 gibt die Richtung größter Streuung, t_2 die Reststreuung an.



Komplette Analyse

Zunächst muss entschieden werden, ob überhaupt Abhängigkeiten in den Variablen vorliegen. Dazu stellt man die Hypothese H_0 auf:

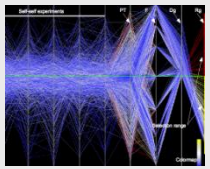
$$H_0 : R = I \qquad H_1 : R \neq I$$

Unter der Annahme, dass es sich bei Z um normalverteilte Zufallsvariablen handelt, kann man dann einen χ^2-Test durchführen.

Dazu vergleicht man

$$\chi^2_{err1} := -\left(m - \left(2n + 11\right)/6\right) \ln\left(detR\right)$$

mit den Werten der χ^2-Verteilung. Nach Festlegung der Irrtumswahrscheinlichkeit α wird also ermittelt, ob sich R von I signifikant unterscheidet.



Wenn R von I verschieden ist, erfolgt die Hauptachsentransformation und die Darstellung

$$\mathbf{Z = FL^T}$$

Um noch die Anzahl p<n der wichtigen Hauptkomponenten zu bestimmen, notiert man zunächst, dass der Anteil der Hauptkomponente f_l an der Gesamtvarianz n gleich λ_i/n ist.

Mit Hilfe der Testmaßzahl

$$\chi^2_{err2} := (m-1)(n-p)\ln detR$$

prüft man die Hypothese

H_2: Die n-p kleinsten Eigenwerte von R sind gleich

Als Faustregel können stattdessen einfach die Hauptkomponenten mit $\lambda_i>1$ benutzt werden. Beide Methoden sind nicht perfekt, da auch im ersten Fall χ^2_{err2} für $n \to \infty$ nicht gegen χ^2 konvergiert.