# Multivariate Statistics - Exam Questions

## TU Wien, 107.388 WS 2018

## v0.1

by ∼ondelette∼

27. Januar 2020

## Inhaltsverzeichnis

# 1 Mögliche Prüfungsfragen

Liste möglicher Prüfungsfragen:

1. Spektralzerlegungssatz (Eigenwerte, Eigenvektoren, symmetrische Matrizen)
2. Kovarianzmatrix, Korrelationsmatrix: Variablen und Daten (Stichproben). (Erwartungswert, Kovarianzmatrix sollte man berechnen können).
3. Multivariate Normalverteilung, Transformation dorthin. Zusammenhang mit Mahalanobis-Distanz.
4. Distanzmaße, was ist das, wie verwendet?
5. Clusteranalyse: Klassifikationstypen (4). Fokus auf Partitionen, Hierarchische Methoden
6. Maße für Homogenität und für Heterogenität von Clustern
7. Gütemaß einer Klassifikation
8. Partitionen: k-means, Funktionsweise
9. Hierarchisches Clustering: divisive, agglomerative Methoden. Linkage Methoden.
10. Fuzzy Clustering. Membership-Matrix, -Koeffizienten. Wie? Kostenfunktion (objective function)
11. Lineares Regressionsmodell: Definitionen, Modellannahmen.
12. Kleinste Quadratsummenschätzer (Least Squares-estimation): Lösung herleiten/vorrechnen bis $\hat{\boldsymbol{\beta}}$ Gl. (3.8) (Regressionskoeffizienten).
13. Gauß-Markov-Theorem. Warum Voraussetzung Normalverteilung und Unkorreliertheit des Fehlervektors, wo benötigt? Zusammenhang mit Inferenz, Konfidenzintervallen.
14. Maximum-Likelihood-Schätzer für Regressionskoeffizienten (nicht sehr wichtig).
15. Multivariate Erweiterung der Linearen Regression. Parameterschätzung im multivariaten Fall.
16. Grundlegende Ideen der robusten Statistik. Bedeutung von Robustheit, Bruchpunkt (robust: Bruchpunkt $> 0$).
17. Robuste Regression: $L_1$-Regression. Kostenfunktion (objective function). (Robust gegen y-Ausreißer, nicht robust gegen x-Ausreißer = Hebelpunkte).
18. LMS-, LTS-Regression. Kostenfunktionen, Bruchpunkte.
19. Scale (Skalierungsparameter, Streuung) der Residuen schätzen mit robusten Schätzern (robust estimators of scale), für Standardisierung der Residuen für Diagnostik.
20. M-Estimators: nicht im Skriptum, aber in VO behandelt. Bringt Extrapunkte wenn man das weiß.
21. Robuste multivariate Schätzer für Lokation und Kovarianz. Eigenschaft der affinen Äquivarianz. Definitionen MVE und MCD.

22. Hat-Matrix, was ist das? (LS-Estimator). Warum nicht robust? (Diagonalelemente $\sim$ klassische Mahalanobisdistanz, diese ist nicht robust!).

23. Diagnostik-Werkzeuge zur Erkennung von Ausreißern (bei Regression). Vertikale Ausreißer, gute vs. schlechte Hebelpunkte. Wie Ausreißer identifizieren? Schwellwerte für Erkennung? Warum Quantile der $\chi^2$-Verteilung benutzt? Robustifizierte Distanzen. Diagnostik-Plots für (Residuen der) Regression.

24. PCA Model/Transformation: Symbole/Größen ($\boldsymbol{z}, \boldsymbol{\Gamma}$). Formulierung für Population und Stichprobendaten (samples). Warum zentrierte Daten? Rechenweg zur 1. Hauptkomponente (Maximierung der Varianz unter Nebenbedingungen, Unkorreliertheit der Hauptkomponenten $z$). Zusammenhang zu Eigenwert-, Eigenvektorzerlegung der Kovarianzmatrix $\boldsymbol{\Sigma}$. Anzahl der relevanten Komponenten (es gibt dafür statistische Tests, Faustregeln). Zusammenhang zu anderen Methoden: SVD (wie funktioniert es, welche Matrizen kommen dabei vor? Warum zentrierte Daten, Zusammenhang zur Eigenwertzerlegung?)

25. Biplots: Was zeigen sie? Aufteilung der Matrix $\boldsymbol{D}$ (über $c$). Innere Produkte als Approximationen der Daten und von statistischen Größen (Gl. (5.43) - (5.45), S. 70).

26. Diagnostik von PCA: Score-, Orthogonale-Distanz. Was ist das? Es gibt Schwellwerte für Identifizierung von Ausreißern und Hebelpunkten.

27. Faktorenanalyse: Faktormodell, vorkommende Größen. Unterschied zu PCA-Modell? (in PCA keine direkte Dimensionreduktion, in FA schon). Modellannahmen in FA, Zusammenhang zu Faktoren; Fehlerterm, Kovarianzmatrix ($\boldsymbol{\Psi}$, Diagnonalstruktur angenommen) des Fehlerterms. FA ist nicht eindeutig. Parameterschätzung für FA (S. 78ff). Faktorenrotation: Man zielt auf spezielle Struktur der Ladungsmatrix ab; verschiedene Varianten. Faktor-scores: Mit KQS (LS) oder Regresionsmethode.

28. Multiple Korrelation. Was wird minimiert, MSE zwischen Daten und linearer Vorhersage (S. 95).

29. Kanonische Korrelation: Worum geht es. Kostenfunktion (objective fuction). Lösungen für $\boldsymbol{a}$, $\boldsymbol{b}$. Warum landet man wieder bei einem Eigenvektorproblem? (erweiterte Cauchy-Schwarz-Ungleichung!). Es gibt statistische Tests für Ergebnisse der Korrelationsanalyse, z.B. ob 2 Vektoren $\boldsymbol{x}, \boldsymbol{y}$ unkorreliert sind. Was bedeutet eine kanonische Korrelation von 1, z.b. für den 1. kanonische Korrelationskoeffizienten? Bereits starke Überlappung von Datensets, Was wenn nur 1 Variable in beiden Datensets?

30. Diskriminanzanalyse. 2 Herangehensweisen (Bayes, Fisher). Bayes'scher Ansatz: Wie kommt man zur Klassifikationsregel? EKM (erwartete Kosten bei Missklassifikation), was wird minimiert. Wie dann anwenden, z.B. für Daten aus multivariater Normalverteilung. LDA (Linear Discriminant Analysis): welche zusätzlichen Annahmen, welche Parameter kommen vor, was braucht man/muss man schätzen aus den Daten? QDA (Quadratic DA), Unterschied zu LDA.

31. Diskriminanzanalyse: Ansatz nach Fisher: Vektor $\boldsymbol{a}$, Projektion auf 1 Dimension. Verteilungsannahmen oder Annahmen über Kovarianzstruktur bei Fisher (macht keine!)? Warum kommt man zum gleichen Ergebnis? Was passiert in beiden Fällen wenn man keine (multivariat) normalverteilten Daten mehr hat? (Tatsächlich braucht auch Fisher die Verteilungsannahme für optimale Ergebnisse im Sinn der Minimierung der EKM.)
32. Diskriminanzanalyse: Mehrgruppenfall (Bayes). Annahmen, wie benutzt man die Klassifikationsregeln. Formeln nicht wichtig, aber man benötigt idR die inverse der Kovarianzmatrix $\boldsymbol{\Sigma}$ bzw. der geschätzten Variante $\boldsymbol{S}$ (das kann problematisch sein).
33. Diskriminanzanalse: Mehrgruppenfall (Fisher). Zusammenhang der Gruppenmittel, Gesamtmittel (S. 120). Within-, Between-Group-Variations. Wieder landet man bei einem Eigenvektorproblem (Spezial: Beweis dafür). Fisher Diskriminanzfunktion für Klassfizierung (neuer) Beobachtungen.
34. Projection Pursuit: Grundlegende Idee.

## 2 Antworten

Alle Seitenangaben, etc., beziehen sich auf die Version des Skriptums vom WS1819.

### 1: Spektralsatz

[p. 8ff]

Jede symmetrische Matrix $\boldsymbol{\Sigma}$ $(p \times p)$ kann zerlegt werden in:

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^{\mathsf{T}} = \sum_{i=1}^{p} a_i \boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^{\mathsf{T}}. \tag{1}$$

Dabei ist $\boldsymbol{A}$ eine Diagonalmatrix, deren Einträge $a_1, a_2, \ldots, a_p$ die Eigenwerte von $\boldsymbol{\Sigma}$ sind. Die Matrix $\boldsymbol{\Gamma}$ ist eine orthogonale Matrix, d.h. $\boldsymbol{\Gamma}^{\mathsf{T}} = \boldsymbol{\Gamma}^{-1}$, und die Spalten (diese haben Norm 1) sind die normierten (standardisierten) Eigenvektoren $\boldsymbol{\gamma}_i$ von $\boldsymbol{\Sigma}$.

Es gilt daher:

$$\boldsymbol{\gamma}_i^{\mathsf{T}} \boldsymbol{\gamma}_j = \begin{cases} 0, & i \neq j \\ 1, & i = j. \end{cases} \tag{2}$$

Weiters haben $\boldsymbol{\Sigma}$ und $\boldsymbol{A}$ dieselben Eigenwerte $a_i$.

Bei einer symmetrischen Matrix sind die Eigenvektoren, die zu verschiedenen Eigenwerten gehören, orthogonal. Außerdem sind die Eigenwerte einer symmetrischen Matrix nicht-negativ, d.h. $\geq 0$.

### 2: Kovarianzmatrix

Erweiterung Kovarianz auf mehrdimensionalen Fall. Siehe S. 10.

Für Zufallsvektor mit Dimension $p$ hat die Kov.matrix Dimension $p \times p$.

Eigenschaften von $\boldsymbol{\Sigma}$: Symmetrisch, positiv semi-definit.
(Klassische) Schätzer für $\boldsymbol{\Sigma}$ (siehe z.B. S. 16):

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{i\cdot} - \bar{\boldsymbol{x}}_i)^{\mathsf{T}}, \tag{3}$$

mit $\boldsymbol{x}_{i\cdot}$ den als Spaltenvektoren geschriebenen Zeilen von $\boldsymbol{X}$.

Zentriert man die Spalten der Datenmatrix $\boldsymbol{X}$ jeweils, d.h. subtrahiert von jeder Spalte jeweils ihren empirischen Mittelwert $\bar{\boldsymbol{x}}$, erhält man die zentrierte Datenmatrix $\boldsymbol{X'}$ (Notation so nicht im Skriptum). Der Schätzer $\boldsymbol{S}$ lässt sich dann auch schreiben als (siehe S. 68):

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X'}^{\mathsf{T}}\boldsymbol{X'}. \tag{4}$$

### 3: Multivariate Normalverteilung

Zur Erinnerung: univariate Normalverteilung:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\frac{(x-\mu^2)}{\sigma^2}\right) \tag{5}$$

**Multivariate Normalverteilung**: Für $p$-dimensionale Zufallsvektoren $\boldsymbol{x}$ die $p$-dimensional normalverteilt sind hat die Dichtefunktion folgende Form:

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{6}$$

wobei $|\boldsymbol{\Sigma}|$ die Determinante der Kovarianzmatrix $\boldsymbol{\Sigma}$ (Dimension $p \times p$) ist. In der Verteilungsfunktion scheint die (quadrierte) Mahalanobisdistanz $\mathrm{MD}^2(\boldsymbol{x}, \boldsymbol{\mu})$ auf:

$$\mathrm{MD}^2(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}). \tag{7}$$

Zur Erinnerung: Die Mahalanobisdistanz ist (in der Statistik) ein Distanzmaß zwischen verschiedenen (multivariaten) Realisierungen $\boldsymbol{x}_i$, wobei auch die Kovarianzsstruktur mitberücksichtigt wird (über $\boldsymbol{\Sigma}$). Die MD hängt nicht von der Skalierung der Variablen ab. Ersetzt man $\boldsymbol{\Sigma}$ durch $\boldsymbol{I}$, erhält man die euklidische Distanz. (Siehe z.B. auch Skriptum Datenanalyse).

### 4: Distanzmaße

[S. 18f] Man führt Ähnlichkeit zwischen Objekten auf Distanzen zwischen (gemessenen/beobachteten) Ausprägungen gemeinsamer Eigenschaften (oder so ähnlich, bla bla) zurück.

Für die Messung von Abständen zwischen Punkten $\boldsymbol{x}_i$ (Datenpunkten, Vektoren) der Dimension $l \times 1$ (geänderte Notation, sonst verwenden wir $p$) in Vektorräumen. Man kann Distanzmaße mit Vektor-Normen formulieren, etwa den $p$-Normen(auch Minkowski--Normen) (Achtung, das ist nicht das $p$ dass wir sonst für die Dimension der Vektoren

benutzen!). (Es muss $p \geq 1$ gelten damit man eine Metrik erhält, sonst gilt nämlich die Dreiecksungleichung nicht):

$$d(i,j) = \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|_p = \left( \sum_{k=1}^{l} |x_{ik} - x_{jk}|^p \right)^{1/p} \tag{8}$$

Wir benutzen meistens die $l_1$-Norm ($p = 1$) (Manhatten-, Cityblock-, Betragsnorm) oder die euklidische Norm $l_2$ ($p = 2$).

Man kann mit einem gewählten Distanzmaß dann eine Distanzmatrix mit allen paarweisen Distanzen zwischen allen $n$ Objekten bzw. Datenpunkten (Zeilen der Datenmatrix $\boldsymbol{X}$) berechnen. Auf dieser Distanzmatrix basierend kann man Clusteranalysen durchführen.

Bei auf Distanzmaßen basierenden Clusteranalysen sollten die Daten vorher standardisiert werden (pro Dimension), sonst kann die (eventuell stark) unterschiedliche Skalierung das Ergebnis negativ beeinflussen (die Cluster stark verzerren?).

Kommentar: Normen müssen die 3 Normaxiome (Definitheit, absolute Homogenität, Dreiecksungleichung) erfüllen, Metriken 3 Axiome für Metriken (positive (semi-?) Definitheit, Symmetrie, Dreiecksungleichung). Da Metriken symmetrisch sein müssen, d.h. $d(i,j) = d(j,i)$, muss eine Distanzmatrix symmetrisch sein.

## 5: Clusteranalyse: Klassifikationstypen

[p. 18f] Es gibt vier Typen:

- Überdeckung: Klassen dürfen sich Überschneiden, aber nicht zusammenfallen.
- Partition: Alle Klassen disjunkt.
- Quasihierarchie: Folge von Überdeckungen. Baum von sich überschneidenden Klassen, von unten grob nach oben feinste Überdeckung.
- Hierarchie: Folge von Partitionen, d.h. disjunkten Klassen. Dendrogramm als mögliche Visualisierung.

Bei jedem dieser Typen kann man noch unterschieden zwischen

1. exhaustiver Klassifikation
2. nicht-exhaustiver Klassifikation

Fokus liegt auf Partition und hierarchischen Methoden.

## 6: Maße für Homogenität und Heterogenität von Clustern

### Homogenitätsmaße

Ziel: Maßzahl $\geq 0$ für Homogenität der Elemente einer Klasse. Je kleiner, desto homogener ist untersuchte Klasse $K_l$.

- Summe der Distanzen

- Maximum der Distanzen
- Minimum der Distanzen
- Summe der Varianzen der Variablen

**Heterogenitätsmaße**
Ziel: Maßzahl $\geq 0$ für Heterogenität der Elemente zweier unterschiedlicher Klassen $K_{l_1}$, $K_{l_2}$. Je kleiner, desto ähnlicher, je größere desto unähnlicher, d.h. heterogener sind die Klassen. Maß soll symmetrisch sein. Unterscheidung notwendig ob disjunkte Klassen (Partition) oder nicht (Überdeckung).
**Disjunkte Klassen (Partition)**:

- Complete Linkage
- Single Linkage
- Average Linkage
- Centroid Methode

**Möglicherweise überschneidende Klassen (Überdeckung)**: Obige Maße werden verwendet, aber die Elemente der Schnittmengen der Klassen werden jeweils vorher ausgeschlossen und gehen nicht in die Berechnung der Maße ein.

## 7. Gütemaße einer Klassifikation

Gütemaß: je kleiner die Maßzahl, desto besser die Klassifikation. Hängt vom Klassifikationstyp ab.
**Partition**, mögliche Gütemaße:

- Summe der Klassenhomogenitäten
- normierter Kehrwert der Heterogenitäten
- Quotient Homogenitäten und Hetereogenitäten

**Hierarchie**: Gütemaße wie oben, werden aber für jede Stufe der Hierarchie berechnet.

## 8: Iterative Partitionierung, k-means

Es gibt Partitionierungsmethoden, die einen iterativen Algorithmus benutzen. Es muss die (bekannte oder irgendwie geschätzte) Anzahl $k$ von Klassen vorgegeben werden. Man kann natürlich mit verschiedenen $k$ testen und nach einem bestimmten Gütemaß für das beste $k$ entscheiden.

Es muss eine Anfangspartition vorgegeben werden, d.h. $k$ Punkte (Clusterzentren) ausgewählt werden, die dann die jeweiligen Klassen (vorläufig) repräsentieren. Jeder der übrigen $n - k$ Punkte wird dann jeweils dem nächsten Zentrum zugeordnet, hierbei kommt ein Distanzmaß (siehe Frage weiter oben) zum Einsatz. Für jeden der $n - k$ Punkte wird dann geprüft ob sich die Güte verbessert (das Gütemaß verringert wird!) wenn er in einen anderen Cluster verschoben wird. Kann keine wesentliche Verbesserung mehr erreicht werden wird abgebrochen und die Clusterzentren aus den neu gefundenen Clustern berechnet. Dann startet die Prozedur von Vorne.

### k-means

[S. 22] Ähnlich wie oben, aber es gibt speziell Fehlermaße, die durch verschieben von Beobachtungen in neue Cluster reduziert werden. Ein mögliches Maß ist die Summe über die quadrierten euklidischen Distanzen jedes zu dem ihm zuvor (im vorigen Schritt des Algorithmus) zugeordneten Clusterzentrum. Man kann jeweils mehrmals mit unterschiedlichen Initialisierungen laufen lassen und das beste Ergebnis benutzen. k-means erzeugt sphärische (kreisförmige) Cluster, wenn man eine Metrik benutzt die alle Richtungen gleich gewichtet (anders wenn man z.B. die Mahalanobisdistanz benutzen würde?).

## 9: Hierarchisches Clustering

Folgen von Partitionen (disjunkte Klassen). Divisive (sehr rechenaufwändig) oder aggleromative Methoden (diese haben wir weiter behandelt).

Startpunkt ist eine Partitionierung der $n$ Datenpunkte in ebensoviele Klassen. In jedem Schritt Zusammenfassung jener beiden Klassen die minimal verschieden sind, bis nur noch eine Klasse über ist. Das ist eine Folge von Partitionierungen (jeweils disjunkte Klassen).

„Minimal verschieden“ wird über ein Distanzmaß (Maß für Homogenität disjunkter Klassen) bestimmt.

- Complete Linkage: bei inhomogenen Klassen schlechte Gruppierungen. Wenige, relativ große Klassen.
- Single Linkage: viele kleine Klassen. Große Klassen werden früh vereint.
- Average Linkage
- Centroid Methode

Es ist möglich die Hierarchie in einem Dendrogramm zu visualisieren (da die Distanzen zwischen den Gruppen (?) in jedem Schritt größer werden).

## 10: Fuzzy Clustering

Keine feste Zuordnung der Objekte zu Clustern, es wird jede Beobachtung jedem Cluster mit einem gewissen Prozentsatz (oder einem Wert $u_{iv}$ in [0, 1], wobei $i = 1, \ldots n$ Beobachtungen und $v = 1, \ldots, k$ Cluster) zugeordnet. Dies wird als Zugehörigkeitskoeffizient bezeichnet. die Koeffizienten summieren für feste Beobachtung zu 1. Die Anzahl $k$ der Cluster muss vorgegeben werden.

Vorteil: man erhält detaillierte Information über die Struktur der Daten, z.B. auch „wie sicher“ eine Beobachtung welchen Clustern zugeordnet werden kann. Nachteil: kann für $n$ groß unüberschaubar werden (?)

Es gibt mehrere mögliche Methoden bzw. zugehörige Kostenfunktionen. Z.B. fuzzy k-means oder MND-2 Algorithmus (glaub nicht dass man das kennen/können muss).

## 11: Lineares Regressionsmodell

Mit einer unabhängigen Variable (input) $x_1$ und einer abhängigen Variable $y$ (output):

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \tag{9}$$

**Lineare multiple Regression**: mit $q$ unabhängigen Variablen $x_1, x_2, \ldots, x_q$ und einer abhängigen Variable $Y$:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon \tag{10}$$

Zusammenfassung in Vektor: $(\beta_0, \beta_1, \ldots, \beta_q)^\mathsf{T} = \boldsymbol{\beta}$, Dimension $((q+1) \times 1)$, führt auf:

$$Y = \boldsymbol{\beta}^\mathsf{T} \boldsymbol{x}^* + \varepsilon \tag{11}$$

wobei der $((q+1)\times 1)$-dimensionale Vektor $\boldsymbol{x}^* = (1, \boldsymbol{x}^\mathsf{T})^\mathsf{T}$ der um 1 erweiterte Datenvektor (Datenpunkt) $\boldsymbol{x}$ ist (Die Notation mit * ist nicht im Skriptum, wird hier eingeführt um das zu verdeutlichen).

Man hat $n$ Beobachtungen von $Y$ und von $x_1, \ldots, x_q$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_q x_{iq} + \varepsilon_i. \tag{12}$$

Zusammengefasst wieder:

$$Y_i = \boldsymbol{\beta}^\mathsf{T} \boldsymbol{x}_i^* + \varepsilon_i. \tag{13}$$

Die $i = 1, \ldots, n$ Gleichungen (13) können in Matrixschreibweise zusammengefasst werden als:

$$\boldsymbol{Y} = \boldsymbol{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{14}$$

Dimensionen: $\boldsymbol{Y}$ hat $n \times 1$, $\boldsymbol{\beta}$ hat $(q+1) \times 1$, $\boldsymbol{\varepsilon}$ hat $n \times 1$. $\boldsymbol{X}^*$ ist die sogenannte Design-Matrix (Dimension: $n \times (q+1)$), diese ist hier aus den Vektoren $\boldsymbol{x}_i^{*\mathsf{T}}$ als ihre Zeilen aufgebaut, und hat daher als erste Spalte einen Einsvektor:

$$\boldsymbol{X}^* = \begin{pmatrix} 1, x_{11}, x_{12}, \ldots, x_{1q} \\ 1, x_{21}, x_{22}, \ldots, x_{2q} \\ \vdots \\ 1, x_{i1}, x_{i2}, \ldots, x_{iq} \\ \vdots \\ 1, x_{n1}, x_{n2}, \ldots, x_{nq} \end{pmatrix} \tag{15}$$

Es handelt sich also hier um die mit einer Spalte aus 1ern erweiterte Datenmatrix $\boldsymbol{X}$.

**Modellannahmen**:

- (Man nimmt and/hofft dass sich $y$ bzw. $Y$ „einigermaßen gut“ durch ein lineares Modell beschreiben lässt. Es gibt dafür Maße, etwa das Bestimmtheitsmaß $R^2$)
- Fehlerterm: Mittelwert 0, $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$

- Fehlerterm: Varianz $\sigma^2$ für alle Komponenten von $\boldsymbol{\varepsilon}$ gleich (Homoskedastizität), außerdem sind die Komponenten unkorreliert. Die Kovrianzmatrix von $\boldsymbol{\varepsilon}$ hat daher die Form: $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}) = \sigma^2 \boldsymbol{I}_n$.
- Zum Testen von Hypothesen (über die Koeffizienten $\boldsymbol{\beta}$?) bzw. Konstruktion von Konfidenzintervallen wird man noch gemeinsame Normalverteilung voraussetzen müssen, $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$.

### 12: Regression: KQS-Schätzer (LS-Estimator) Regressionskoeffizienten

Man geht, für jeden der $n$ Datenpunkte, von der Differenz der abhängigen (modellierten) Variable $y_i$ zu der Vorhersage des Modells $\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}$ aus (Anmerkung: bei $\boldsymbol{x}$ und $\boldsymbol{X}$ handelt es sich um die mit 1 bzw. einer 1er Spalte erweiterten Versionen, um den Intercept $\beta_0$ nicht als extra Term mitschleppen zu müssen; ist hier im Gegensatz zu oben nicht mehr gekennzeichnet, damit die Notation der im Skriptum entspricht):

$$r_i = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq} = y_i - \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}_i. \tag{16}$$

Die Methode der kleinsten Quadrate minimiert die Summe der quadriertern Residuen (Residual Sum of Squares, RSS):

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq})^2 = \tag{17}$$

$$= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2. \tag{18}$$

Vorgehen:

1. Gleichung für RSS ausmultiplizieren, zusammenfassen
2. Gleichung für RSS nach $\boldsymbol{\beta}$ partiell ableiten
3. Ergebnis nullsetzen
4. Lösung ist LS-Schätzer für $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{y}. \tag{19}$$

TO DO: ?

### 13: Regression: Gauß-Markov-Theorem

Unter den Annahmen für das multiple Regressionsmodell (Unkorreliertheit der Element des Fehlervektors $\boldsymbol{\varepsilon}$, und Homoskedastizität, d.h. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$) gilt:

1. $\hat{\boldsymbol{\beta}}$ ist ein eindeutig bestimmter, effizienter, linearer Schätzer für $\boldsymbol{\beta}$.
2. $s^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\mathsf{T}\hat{\boldsymbol{\varepsilon}}}{n-q-1}$ ist ein unverzerrter Schätzer für die Residuenvarianz.

„Ein effizienter Schätzer hat eine Kovarianzmatrix, die kleiner ist als die jedes anderen linearen unverzerrten Schätzers. Beim Kleinsten-QS-Schätzer spricht man auch vom besten linearen unverzerrten Schätzer (BLUE)."

Anmerkung: BLUE = Best Linear Unbiased Estimator. „Best" bezieht sich hier auf die geringste Varianz des Schätzwertes. Es gibt also unter den genannten Voraussetzungen keinen anderen (linearen unverzerrten!) Schätzer (Schätzfunktion) der eine geringere Varianz liefert. (Es kann sein dass es Schätzer mit geringerer Varianz gibt die nichtlinear, verzerrt, etc., sind. Laut Wikipedia ist z.B. Ridge-Regression solch ein Schätzer).

### 14: Regression: Maximum Likelihood-Schätzer für Regressionskoeffizienten

Unter den Annahmen dass $\boldsymbol{X}$ vollen Rang hat (erweiterte Datenmatrix mit 1er Spalte), und den üblichen Annahmen für die Residuen, ist der ML-Schätzer für $\boldsymbol{\beta}$ gleich dem LS-Schätzer.

TO DO: weitere Eigenschaften

### 15: Regression: Multivariate Erweiterung

Zuvor wurde eine abhängige Variable $y$ durch eine unabhängige Variable $x$, oder eine abhängige Variable $Y$ durch mehrere unabhängige Variablen $x_1, \ldots, x_q$ modelliert. Bei der multivariaten linearen Regression gibt es $Y_1, Y_2, \ldots, Y_m$ abhängige Variablen, die durch die unabhängigen Variablen $x_1, \ldots, x_q$ modelliert werden sollen.

Für jeweils $n$ Beobachtungen bekommt man also (andere Notation als im Skriptum!) $\boldsymbol{\mathcal{Y}} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$, mit Dimension $n \times m$. Fehlermatrix $\boldsymbol{\mathcal{E}} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m)$, ebenfalls $n \times m$. Koeffizientenmatrix $\boldsymbol{\mathcal{B}} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m)$ mit Dimension $(q+1) \times 1$ (es ist hier ja in jeder Spalte als erster Eintrag der Intercept-Koeffizient). Datenmatrix $\boldsymbol{X}$, mit Dimension $(q+1) \times 1$.

Modell:

$$\boldsymbol{\mathcal{Y}} = \boldsymbol{X}\boldsymbol{\mathcal{B}} + \boldsymbol{\mathcal{E}}. \tag{20}$$

Schätzer:

$$\hat{\boldsymbol{\mathcal{B}}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\mathcal{Y}}. \tag{21}$$

### 16: Robuste Statistik, Grundideen

Grundidee: Methoden zu finden, die gegenüber Abweichungen der Daten von Modellannahmen resistent sind. Kleine Abweichungen vom Modell sollen nur geringe Auswirkungen haben. Z.B. kann ein einzelner stark abweichender Wert (Ausreißer) in einer Stichprobe den empirischen Mittelwert stark verzerren. Ein robuster Lokationsschätzer würde (sollte) das verhindern bzw. die Verzerrung gering(er) halten.

Der Bruchpunkt ist der Punkt wo eine Methode völlig fehl schlägt. Es geht dabei darum, wie viele Datenpunkte aus einer Menge von $n$ Datenpunkten von der Mehrheit der Daten abweichen kann, bevor die Schätzwerte völlig sinnlos sind, also von den wahren Werten (z.B. Stichprobenmittel gegen Populationsmittel) sehr stark abweichen. Verzerrung der Ergebnisse sind bereits vorher möglich.

## 17: Robuste Regression: L1-Regression

### $L_1$ Regression

Statt den quadrierten Residuen (RSS) wird der Betrag der Residuen (also die $L_1$-Norm) minimiert, man erhält also eine andere Kostenfunktion:

$$\sum_{i=1}^{n} |r_i(\boldsymbol{\beta})| \tag{22}$$

Dafür gibt es keine geschlossene Lösung, man muss mit einem Algorithmus eine approximative Lösung finden.

$L_1$-Regression ist robust gegenüber $y$-Ausreißern, aber nicht gegenüber $x$-Ausreißern (Hebelpunkten). Gegenüber Hebelpunkten hat der $L_1$-Regression-Schätzer einen Bruchpunkt von 0%.

## 18: Robuste Regression: LMS und LTS Regression

Man wählt nicht mehr alle $n$ Datenpunkte für die Schätzung der Modellparameter (Anpassung des Modells), sondern nur eine Mehrheit.

### LMS Least Median of Squares

Kostenfunktion:

$$\underset{i}{\text{med}}\, r_i^2(\boldsymbol{\beta}) \tag{23}$$

soll minimiert werden. Keine geschlossene Lösung, nur approximative Algorithmen.

LMS hat einen Bruchpunkt von 50%.

### LTS Least Trimmed Sum of Squares

Kostenfunktion:

$$\sum_{i=1}^{h} (r^2(\boldsymbol{\beta}))_{(i)} \tag{24}$$

soll minimiert werden. $(r^2(\boldsymbol{\beta}))_{(i)}$ sind die der Größe nach aufsteigend sortierten quadrierten Residuen, der Sub-Index $_{(i)}$, $i = 1, \ldots, h$ kennzeichnet die Reihenfolge.

Bruchpunkt für $h \approx n/2$ ist etwa 50%, bei größerem $h$ sinkt er auf etwa $(n - h)/n$.

## 19: Robuste Schätzer für Skalierung der Residuen, für Diagnostik

TO DO

## 20: M-Estimators

(Nicht im Skriptum. Vorlesungsmitschrift). (M-Schätzer wurden eingeführt von Peter J. Huber, einem der Begründer der robusten Statistik.)

LS-Schätzer (bei Regression) sind sehr empfindlich gegenüber Ausreißern. Man arbeitet mit quadratischen Kostenfunktionen (führt auf den sample mean als Schätzer), also quadrierten Residuen (man minimiert ja RSS, Residual Sum of Squares). Deswegen

wirken sich große Residuen (potentielle Ausreißer in y-Richtung) noch stärker auf die Schätzung aus. (Wie ist das mit schlechten Hebelpunkten, d.h. Ausreißern in x-Richtung die das Ergebnis negativ beeinflussen?) Im Prinzip geht es darum, den Einfluss von Residuen, die im Gegensatz zu den anderen sehr groß sind (also eventuell von Ausreißern kommen, Datenpunkten die sehr stark von der Mehrheit der Datenpunkte abweichen), zu verringern.

Eine Kostenfunktion mit Absolutwerten führt auf den Median als Schätzer. Dieser ist zumindest robust gegenüber y-Ausreißeren, aber nicht gegenüber x-Ausreißern.

M-Schätzer minimieren folgende Art von Kostenfunktion:

$$\hat{\boldsymbol{\beta}}_M = \arg\min_{\beta} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}(\boldsymbol{\beta})}\right) \tag{25}$$

mit einer Funktion $\rho(\cdot)$ (diese muss differenzierbar sein?), für alle $n$ Datenpunkte, und den Residuen $r_i$, der Differenz zwischen Zielwert $y_i$ und dem durch das Regressionsmodell über $\boldsymbol{\beta}$ aus $\boldsymbol{x}_i$ vorhergesagten Wert. Man standardisiert die Residuen $r_i$ mit einer Schätzung $\hat{\sigma}$ der Residuen-Standardabweichung, damit die Methode Skaleninvariant wird. (Die skalierten Residuen sollte also Standard-Normalverteilt sein?)

Neben der Funktion $\rho(\cdot)$ spielt ihre Ableitung $\psi(\cdot)$ eine Rolle, sowie die sich daraus durch Skalierung mit dem aktuellen Residuenwert ($r_i$) ergebende Gewichtungsfunktion $w_i$, welche den Einfluss größerer Residuen auf die Schätzung von $\boldsymbol{\beta}$ reduziert. Es gibt mehrere verschiedene Möglichkeiten, diese Funktionen zu wählen; bei LS werden alle Residuen gleich gewichtet, bei $L_1$-Regression werden sie mit $\sim \frac{1}{(r_i/\hat{\sigma})}$ gewichtet. Weiters gibt es z.B. Huber und Tukey-biweight als Gewichtungsfunktionen, die (symmetrisch) größere Residuen stärker niedergewichten, bis hin zum Nullsetzen.

Die Methode führt letztlich auf einen iterativen Algorithmus:

- Finde Initialisierung $\hat{\boldsymbol{\beta}}_0$. Es sollte eine robuste Methode benutzt werden wie LMS oder LTS. (Es gibt auch robuste Methoden mit höherer Effizienz als diese).
- Für jeden Schritt $m$ müssen zunächst die vorhergesagten Werte über das Modell berechnet werden (mit $\boldsymbol{\beta}$ aus dem vorherigen Schritt), daraus die Residuen. Daraus wird mit einem robusten Schätzer (etwa MAD) ein Schätzwert $\hat{\sigma}$ für die Residuen-Standardabweichung berechnet, um die Residuen zu standardisieren. Daraus berechnet sich die Gewichtungsfunktion $w_i$. Dann wieder eine Schätzung für $\boldsymbol{\beta}$, womit ein neuer Zyklus beginnt.

## 21. Robuste Schätzer für multivariate Lokation und Skalierung: MVE, MCD

Die klassischen Schätzer $\bar{\boldsymbol{x}}$ und $\boldsymbol{S}$ sind nicht robust, jeweils Bruchpunkt 0%.

2 robuste Schätzer für Mittel und Kovarianz multivariater Daten $\boldsymbol{X}$: MVE, MCD.

### MVE: Minimum Volume Ellipsoid

Es gilt jenes Ellipsoid zu finden, das minimales Volumen hat und dabei mindestens die Hälfte der Datenpunkte von $\boldsymbol{X}$ enthält. Der Schätzer $\boldsymbol{C}(\boldsymbol{X})$ muss für Konsistenz mit Normalverteilung noch skaliert werden. Die Gleichung betrachtet die quadrierte

Mahalanobisdistanz von $\boldsymbol{x}_i$ zum geschätzten Zentrum $\boldsymbol{T}(\boldsymbol{X})$ kleiner oder gleich einer quadrierten Konstanten $c$. Diese wird als das 50% Quantil $\chi^2_{p,0.5}$ gewählt.

MVE ist komplexer zu berechnen als MCD.

### MCD: Minimum Covariance Determinant

Es wird nach einer Anzahl $h$ (aus den gesamten $n$) deren empirische Kovarianzmatrix kleinstmögliche Determinante $|\boldsymbol{S}|$ hat. Der Bruchpunkt wird maximal für $h \approx \frac{n}{2}$. Für größeres $h$ sinkt der Bruchpunkt auf $(n-h)/n$.

Kommentare:

- Die Lösungen von MCD und MVE sind i.A. verschieden.
- Robuste Schätzer sind idR weniger effizient als klassische Schätzer. Effizienz ist ein Gütekriterium für Punktschätzer (z.B. für Lokation). Geringere relative Effizienz bedeutet, dass man mehr Datenpunkte (Beobachtungen) braucht, um z.B. die gleiche Varianz (als ein mögliches Gütekriterium für unverzerrte Schätzer) für den Schätzwert zu erhalten wie bei einem relativ effizienteren Schätzer. Andersrum, für eine feste Anzahl verfügbarer Beobachtungen ergibt der weniger relativ effiziente Schätzer z.B. dann größere Varianz des Schätzwertes.
- Weiters können robuste Schätzer zu numerischen Problemen führen.

### Robustifizierte Mahalanobistdistanz = Robuste Distanz

Mit den robusten Schätzungen von Mittel und Kovarianz lässt sich eine robustifizierte Variante der Mahalanobisdistanz (dies benutzt ja klassische Schätzer) gewinnen (S. 48).

### Affine Äquivarianz

Ist eine Eigenschaft von Schätzern (Schätzfunktionen). Die klassischen Schätzer $\bar{\boldsymbol{x}}$ und $\boldsymbol{S}$ sind affin äquivariant (PRÜFEN). Bsp. im Skriptum: Lokationsschätzer $\boldsymbol{T}(\cdot)$, d.h. Schätzer für $\bar{\boldsymbol{x}}$ aus Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x_n}$ (also z.B. den Zeilen einer Datenmatrix). Affine Transformation der Daten, $\boldsymbol{A}\boldsymbol{x}_i + \boldsymbol{b}$, soll sich so auf den Schätzer auswirken:

$$\boldsymbol{T}(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}) = \boldsymbol{A}\,\boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) + \boldsymbol{b}. \tag{26}$$

Das Ergebnis der Schätzung aus affin transformierten Daten soll gleich sein der affinen Transformation des Ergebnisses des Schätzers selbst ($\boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \bar{\boldsymbol{x}}$, Dimension $p \times 1$). Solch eine Transformation kann z.B. Standardisierung der Variablen sein. Die Matrix $\boldsymbol{A}$ soll regulär ($p \times p$) sein, $\boldsymbol{b}$ ($p \times 1$).

Für einen Schätzer $\boldsymbol{C}(\cdot)$ für die Kovarianzmatrix gilt:

$$\boldsymbol{C}(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}) = \boldsymbol{A}\,\boldsymbol{C}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\boldsymbol{A}^\mathsf{T}. \tag{27}$$

Literatur z.B. Schlittgen (2009): Multivariate Statistik (Oldenbourg Verlag).

## 22: Hat-Matrix

Die Hat-Matrix $\boldsymbol{H}$ (Dimension $n \times n$ für Datenmatrix $\boldsymbol{X}$ mit $n \times p$) tritt (in unserem Kontext) beim KQ-Schätzer (LS) für Regression auf. Für eine Datenmatrix $\boldsymbol{X}$ ist die Hat-Matrix gegeben als:

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}. \tag{28}$$

Es gilt $\hat{\boldsymbol{y}}_{\text{LS}} = \boldsymbol{H}\boldsymbol{y}$, $\hat{\boldsymbol{y}}_{\text{LS}}$ sind die „fitted values“. Die Hat-Matrix ist symmetrisch und idempotent.

Sind Diagnonalelemente $h_{ii}$ der Hat-Matrix nahe 1, kann man von einem Hebelpunkt ausgehen, da $\hat{y}_i = y_i$, d.h. für diesen Datenpunkt eine exakte Anpassung des Modells an die anhängige Variable erfolgt, und andere Datenpunkte („lokal“?) keinen Einfluss haben. Ist aber nicht robust, da die $h_{ii}$ zur quadrierten Mahalanobisdistanz proportional sind, diese ist ja nicht robust. Die Methode kann nicht zwischen guten und schlechten Hebelpunkten unterscheiden, wenn es mehere solche $x$-Ausreißer gibt.

### 23: Diagnostik-Werkzeuge zur Erkennung von Ausreißern bei Regression

4 Typen von Datenpunkten:

- Rguläre Beobachtungen
- Vertikale Ausreißer ($y$-Richtung)
- Gute Hebelpunkte
- Schlechte Hebelpunkte

Dafür gibt es den Regression-Diagnostic-Plot. Dieser kombiniert Distance-Distance-Plot (Robuste Distanz vs. Mahalanobis-Distanz) und den Plot standardisierter robuster Residuen (aus robuster Regression, etwa LTS oder LMS) über dem Beobachtungsindex. Der Regression-Diagnostic-Plot zeigt robuste Residuen vs. robuste Distanz (robustifizierte Mahalanobisdistanz). Erlaubt Identifizierung von $y$-Ausreißern, guten und schlechten Hebelpunkten. Grenzewerte sind $\pm 2.5$ für die robusten Residuen bzw. $\sqrt{\chi^2_{q;0.975}}$ ($q$ ist die Dimension der Daten?).

### 24: Hauptkomponentenanalyse (PCA)

#### Herleitung

Wir haben einen $p$-dimensionalen Zufallsvektor $\boldsymbol{x}$ (diesen können wir später mit einer Beobachtung, i.e. mit einer als Spaltenvektor $\boldsymbol{x}_{j.}$ $(p \times 1)$ $(j = 1, \ldots, n)$ geschriebenen Zeile der Datenmatrix $\boldsymbol{X}$ identifizieren). Mittelvektor $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$ werden zunächst als gegeben angenommen.

Wir transformieren $\boldsymbol{x}$ mit einer orthogonalen Matrix $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$ ($p \times p$, es gilt $\boldsymbol{\Gamma}^{\mathsf{T}} = \boldsymbol{\Gamma}^{-1}$), deren Spalten $\boldsymbol{\gamma}_i$ sind paarweise orthogonal und haben jeweils Norm 1 (das wird für die Herleitung zunächst so gefordert):

$$\boldsymbol{z} = \boldsymbol{\Gamma}^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{\mu}) \tag{29}$$

in einen neuen Vektor $\boldsymbol{z}$ $(p \times 1)$ (Kommentar: Dies ist ebenfalls eine als Spaltenvektor geschriebene Zeile $\boldsymbol{z}_{j.}$, für „input“ $\boldsymbol{x}_{j.}$).

Die Herleitung im Skriptum nimmt folgenden Ansatz: man betrachtet die Komponenten $z_i$ $(i = 1, \ldots, p)$ des Vektors $\boldsymbol{z}$. Die Varianz der einzelnen Komponenten, $\text{Var}(z_i) = \boldsymbol{\gamma}_i \boldsymbol{\Sigma} \boldsymbol{\gamma}_i^{\mathsf{T}}$, soll maximiert werden. Dies wird als Lagrangsches Optimierungsproblem formuliert, mit der Nebenbedingung normierter Vektoren $\boldsymbol{\gamma}_i$. Partielle Ableitung nach $\boldsymbol{\gamma}$ und

Nullsetzen führt auf ein Eigenwertproblem für die Kovarianzmatrix $\boldsymbol{\Sigma}$ wobei sich die Vektoren $\boldsymbol{\gamma}_i$ als die Eigenvektoren von $\boldsymbol{\Sigma}$, die zugehörigen Eigenwerte $a_i$ gleichzeitig als die Varianzen der Komponenten $z_i$ herausstellen. Da man die Varianz $a_i$ der Hauptkomponenten $z_i$ maximieren will, sucht man für die 1. HK eben denjenigen Eigenvektor zum größten Eigenwert $a_i$ (Kommentar: eine symmetrische Matrix, was eine Kovarianzmatrix ja ist, hat Eigenwerte $\geq 0$).

Kommentar zu den Begriffen: Jeder Datenpunkt $\boldsymbol{x}_{j\cdot}$ hat einen zugehörigen Vektor $\boldsymbol{z}_{j\cdot}$ von Hauptkomponenten $z_i$ $(i = 1, \ldots, p)$. Die zugehörige Richtung jeder Komponente $z_i$ ist der Vektor $\boldsymbol{\gamma}_i$.

Zusammengefasst:

- Maximierung der Varianz der einzelnen Hauptkomponenten $z_i = (\boldsymbol{z})_i$, unter Nebenbedingung orthogonaler Spaltenvektoren der Transfomationsmatrix $\boldsymbol{\Gamma}$, führt auf ein Eigenwertproblem für die Kovarianzmatrix $\boldsymbol{\Sigma}$.
- Die Eigenvektoren von $\boldsymbol{\Sigma}$ sind genau die Spaltenvektoren von $\boldsymbol{\Gamma}$. Die Eigenwerte $a_i$ von $\boldsymbol{\Sigma}$ sind die Varianzen der Hauptkomponenten $z_i$. Für die Maximierung der Varianz der einzelnen HK werden die Spalten $\boldsymbol{\gamma}_i$ nach absteigender Größe der zugehörigen Eigenwerte $a_i$ in $\boldsymbol{\Gamma}$ angeordnet. Die erste HK hat dann die größte Varianz, die zweite HK die zweitgrößte, usw.
- Man fordert dass die HK (Variablen) $z_i$ unkorreliert sein sollen. Die (paarweise) Orthogonalität verschiedener Richtungen $\boldsymbol{\gamma}_1$, $\boldsymbol{\gamma}_2$, etc., ergibt sich aus der Forderung der Unkorreliertheit! (siehe S. 58).

Bei der praktischen Verwendung der PCA sollten die Daten auf Mittelwert 0 und Varianz 1 standardisiert werden, da die Methode nicht skaleninvariant ist (siehe S. 60).

### PCA von Stichproben (Daten)

Aus Daten $\boldsymbol{X}$ Schätzung von Mittel $\bar{\boldsymbol{x}}$ und Kovarianz $\boldsymbol{S}$ (dies kann auch mittels robuster Schätzer erfolgen, siehe Frage zur Diagnostik mittels PCA!).

Aus der geschätzten Kovarianzmatrix $\boldsymbol{S}$ erlangt man über Eigenwertzerlegung die Eigenvektoren $\hat{\boldsymbol{\gamma}}_j$ $(j = 1, \ldots, p)$ und zugehörige Eigenwerte $\hat{a}_j$. Die Eigenvektoren werden wieder nach absteigender Größe der zugehörigen Eigenwerte sortiert, das ergibt die Matrix $\hat{\boldsymbol{\Gamma}}$. Die Matrix der Hauptkomponenten ergibt sich zu:

$$\boldsymbol{Z} = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^{\mathsf{T}})\hat{\boldsymbol{\Gamma}}. \tag{30}$$

### Singulärwertzerlegung (SVD)

Singulärwertzerlegung (Singular Value Decomposition) kann alternativ zur Eigenwertzerlegung der Kovarianzmatrix verwendet werden. Man geht direkt von der Datenmatrix $\boldsymbol{X}$ aus:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}} \tag{31}$$

Die Matrizen $\boldsymbol{U}$ $(n \times n)$ und $\boldsymbol{V}$ $(p \times p)$ sind orthogonale Matrizen (Spaltenvektoren paarweise orthogonal und Norm 1).

TO DO: Zusammenhang mit PCA.

### Anzahl relevanter Hauptkomponenten

Es gibt für $p$ Variablen ebensoviele Hauptkomponenten. Jede HK „erklärt“ einen gewissen Teil der Gesamtvarianz der Daten $\boldsymbol{X}$. Man kann durch die Wahl der ersten $k$ HK die Dimension der Daten auf $k$, d.h. um $p-k$ reduzieren. Dabei wird noch ein gewisser Teil der Gesamtvarianz erklärt; man kann z.B. so viele HK wählen, dass 80% der Gesamtvarianz erklärt wird.

Es gibt verschiedene Methoden $k$ zu ermitteln. In der Praxis werden eher „Faustregeln“ benutzt, wie der Scree-Plot, der für jede HK den Anteil der HK an der Gesamtvarianz der Daten zeigt. Hier sucht man einen Knickpunkt des Graphen. Oder wie oben beschrieben so dass z.B. 80% der Gesamtvarianz durch die ersten $k$ HK beschrieben wird.

## 25: Biplots

Stellen im Prinzip 2 Aspekte der PCA in einem Plot dar. Einerseits die Loadings, i.e. die Richtungen der Hautpkomponenten ($\boldsymbol{\gamma}$). Andererseits die in den niedrigdimensionalen (z.B. $k=2$ dimensionen (immer für Biplot?)) Raum projizierten Scores $\boldsymbol{z}_{i\cdot}^{(k)}$. Rank 2 Approximation der Daten als Ausgangsbasis, über SVD.

Winkel der Vektoren lässt auf Korrelation zwischen den Variablen schließen.

TO DO: CHECK!

## 26: Diagnostik bei PCA

[S. 71ff] PCA kann auch für Erkennung von multivariaten Ausreißern benutzt werden. Es geht darum zu erkennen, ob (einzelne) Datenpunkte stark von der Struktur der übrigen Daten abweichen. Dies ist aber nur sinnvoll wenn robuste Methoden zur Bestimmung der Hauptkomponenten benutzt werden (S. 72)! Solche sind z.B. in der R-Funktion `PcaHubert()` implementiert (S. 73).

Für die Ausreißererkennung kommen Zwei Distanzmaße zur Anwendung: Score-Distanz (SD) und orthogonale Distanz (OD). Man hat $n$ Beobachtungen der Dimension $p$. Die Daten wurden in $k < p$ Hauptkomponenten dargestellt, d.h. die zugehörige (aus den Daten geschätzte) Loadingsmatrix ist $\hat{\boldsymbol{\Gamma}}_k$ (Dimension $p \times k$, die volle Matrix $\hat{\boldsymbol{\Gamma}}$ hat Dimension $p \times p$). Für jeden der $n$ Datenvektoren $\boldsymbol{x}_{i\cdot}$ ($p \times 1$) gibt es einen Score-Vektor $\boldsymbol{z}_{i\cdot}$ ($k \times 1$) (reduzierte Dimension).

Score-Distanz (SD) für die $i$-te Beobachtung (S. 72):

$$\mathrm{SD}_i = \mathrm{SD}(\boldsymbol{z}_{i\cdot}) = \left( \sum_{j=1}^{k} \frac{z_{ij}^2}{\hat{a}_j} \right)^{1/2}. \tag{32}$$

Entspricht der Mahalanobisdistanz von $\boldsymbol{z}_{i\cdot}$ zum Zentrum der Hauptkomponenten (dieses ist Null, denn wir haben ja die Daten zentriert, siehe auch (5.12), S. 59). Die Gleichung oben kann auch so geschrieben werden:

$$\mathrm{SD}^2(\boldsymbol{z}_{i\cdot}) = \boldsymbol{z}_{i\cdot}^{\mathsf{T}} \boldsymbol{A}_k^{-1} \boldsymbol{z}_{i\cdot}. \tag{33}$$

$\boldsymbol{A}$ ist die Kovarianzmatrix der Hauptkompontenen (Dimension xxx) (S. 59, Gl. (5.13)), $\boldsymbol{A}_k$ $(k \times k)$ ist die auf die ersten $k$ HK reduzierte Kovarianzmatrix der $\boldsymbol{z}_{i.}$ (diese sind ja $k \times 1$). Es ist leicht zu sehen dass $\mathrm{SD}^2(\boldsymbol{z}_{i.})$ von der Form her $\mathrm{MD}^2(\boldsymbol{z}_{i.})$ entspricht.

Orthogonale-Distanz (OD) für die $i$-te Beobachtung (S. 73):

$$\mathrm{OD}_i = \|\boldsymbol{x}_{i.} - \hat{\boldsymbol{\Gamma}}_k \boldsymbol{z}_{i.}\|_2 \tag{34}$$

## 27: Faktorenanalyse

Wir haben einen Zufallsvektor $\boldsymbol{x}$ $(p \times 1)$, diesen zentrieren und standardisieren wir komponentenweise und erhalten daraus den (ebenfalls $(p \times 1)$-dimensionalen) Vektor $\boldsymbol{y}$.

**Modellannahme**: $\boldsymbol{y}$ lässt sich, unter in Kauf nehmen eines Fehlers $\boldsymbol{e}$ $(p \times 1)$, durch eine kleinere Anzahl $k < p$ von Variablen, zusammengefasst in Vektor $\boldsymbol{f}$, darstellen ($k$-Faktormodell):

$$\boldsymbol{y} = \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{e}. \tag{35}$$

$\boldsymbol{\Lambda}$ hat Dimension $(p \times k)$, $\boldsymbol{f}$ hat $(k \times 1)$.

$f_j$ ist ein allgemeiner Faktor wenn fast alle $\lambda_{ij}$ deutlich von 0 verschieden sind. Wenn mindestens 2 Ladungen deutlich von Null verschieden sind ist es ein gemeinsamer Faktor.

Kommentar: Im Gegensatz zu PCA hat das Faktorenmodell also eine Reduktion der Variablen (von $p$ auf $k < p$) „fest eingebaut“. (PCA zerlegt ja zunächst nur die Daten in gleich viele, andere Variablen. Datenreduktion erfolgt dann erst durch weglassen einzelner Hauptkomponenten).

Annahmen:

- $\mathrm{E}(\boldsymbol{f}) = \boldsymbol{0}$
- $\mathrm{E}(\boldsymbol{e}) = \boldsymbol{0}$
- ... TO DO ...
- $\mathrm{Cov}(\boldsymbol{e}) = \boldsymbol{\Psi} = \mathrm{diag}(\psi_{11}, \ldots, \psi_{pp})$

$$\boldsymbol{\rho}_{red} = \boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} = \boldsymbol{\rho} - \boldsymbol{\Psi}. \tag{36}$$

Schätzung: $\hat{\boldsymbol{\rho}}$, $\hat{\boldsymbol{\Lambda}}$, $\hat{\boldsymbol{\Psi}}$.

### Faktorenrotation

Man möchte die Komponentenachsen so rotieren, dass sie z.B. mit möglichst vielen der Variablenachsen übereinstimmen. Wenige Variablen sollten jeweils große Ladungen auf einen Faktor haben, für bessere Interpretierbarkeit. Faktoren-loadings $\boldsymbol{f}$ (?) sind nicht eindeutig. Man kann sie mit orthogonalen Matrizen rotieren und das Modell bleibt gültig. Man nimmt Diagonalität von $\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}$ oder $\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}$ an.

## 28: Multiple Korrelation

Maße für linearen Zusammenhang zwischen Merkmal (Variable) $x$ und $p$-dimensionalem Merkmal $\boldsymbol{y}$. Dies haben eine gemeinsame $p+1$-dimensionale multivariate Verteilung mit $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$.

## 29: Kanonische Korrelation

[S. 97f] Zur Berechnung des linearen Zusammenhanges zwischen Gruppen von Variablen, also multivariaten Zufallsvariablen. Kanonische Korrelationsanalyse kann zur Entdeckung von Strukturen in Daten verwendet werden, sowie zur Dimensionsreduktion, wenn man z.B. nur kanonische Variablen mit besonders großer kanonischer Korrelation behält. Die kanonischen Variablen sollten dabei Interpretierbar sein.

Die multivariaten Zufallsvariablen sind $\boldsymbol{x}$ mit Dimension $p < q$, $\boldsymbol{y}$ mit Dimension $q$. Diese haben $\boldsymbol{\mu}_1$ bzw. $\boldsymbol{\mu}_2$ und $\boldsymbol{\Sigma}_{11}$ bzw. $\boldsymbol{\Sigma}_{22}$, außerdem gibt es die Kreuz-Kovarianzmatrix $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\mathsf{T}$. Diese werden alle mit vollem Rang angenommen.

Über Linearkombinationen mit Vektoren $\boldsymbol{a}_k$ und $\boldsymbol{b}_k$ gelangt man zu den zugehörigen kanonischen Variablen $\varphi_k$ und $\eta_k$ $(k = 1, \ldots, p)$:

$$\varphi_k = \boldsymbol{a}_k^\mathsf{T} \boldsymbol{x}, \tag{37}$$

$$\eta_k = \boldsymbol{b}_k^\mathsf{T} \boldsymbol{y} \tag{38}$$

Man möchte Vektoren $\boldsymbol{a}_k$, $\boldsymbol{b}_k$ finden, die die Korrelation (den sog. kanonischen Korrelationskoeffizienten) $\rho_k$ zwischen $\varphi_k$ und $\eta_k$ maximieren:

$$\max_{\boldsymbol{a}_k, \boldsymbol{b}_k} \mathrm{Cor}(\varphi_k, \eta_k) = \rho_k. \tag{39}$$

Es lassen sich mehrere Paare von Vektoren und damit kanonischer Variablen finden, die jeweils die Korrelation zwischen einander maximieren (es gilt dabei: $\rho_1^2 \geq \rho_2^2 \geq, \ldots, \geq \rho_p^2$) und dabei gleichzeitig unkorreliert mit allen übrigen kanonischen Variablen sind. (Wikiepdia: Man kann dadurch die Kovarianz zwischen Variablen erklären, ähnlich wie man Varianz von Variablen bei der PCA erklärt. Die kanonischen Variablen erklären in absteigender Reihenfolge Kovarianz in den Daten. So ungefähr?).

Es stellt sich heraus dass das Maximum jeweils erreicht wird durch:

$$\varphi_k = \boldsymbol{e}_k^\mathsf{T} \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{x}, \quad \text{mit} \quad \boldsymbol{a}_k^\mathsf{T} = \boldsymbol{e}_k^\mathsf{T} \boldsymbol{\Sigma}_{11}^{-1/2} \tag{40}$$

$$\eta_k = \boldsymbol{f}_k^\mathsf{T} \boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{y}, \quad \text{mit} \quad \boldsymbol{b}_k^\mathsf{T} = \boldsymbol{f}_k^\mathsf{T} \boldsymbol{\Sigma}_{22}^{-1/2}. \tag{41}$$

Dabei sind $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_p$ Eigenvektoren von $\boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1/2}$ zu den Eigenwerten $\rho_1^2, \rho_2^2, \ldots, \rho_p^2$. Zur Festlegung der Reihenfolge sortiert man dabei die Eigenwerte $\rho_k^2$ in absteigender Größe: $\rho_1^2 \geq \rho_2^2 \geq, \ldots, \geq \rho_p^2$ (siehe auch oben).

Was bedeutet ein kanonischer Korrelationskoeffizient $\rho_1$ von 1? Es reicht bereits, dass 1 Variable $x_i$ aus $\boldsymbol{x}$ gleich einer Variable $y_j$ aus $\boldsymbol{y}$ ist; ist (oBdA?) z.B. $x_1 = y_1$, (und die übrigen Einträge in $\boldsymbol{x}$ und $\boldsymbol{y}$ beliebig?) kann man mit Vektoren $\boldsymbol{a}_1 = (1, 0, \ldots, 0)^\mathsf{T}$ und $\boldsymbol{b}_1 = (1, 0, \ldots, 0)^\mathsf{T}$ linear kombinieren, so dass sich ein $\rho_1 = 1$ ergibt. Man kann dann nichts über die anderen Variablen aus $\boldsymbol{x}$ und $\boldsymbol{y}$ aussagen. (Kommentar: habe das so in etwa bei der Prüfung erklärt bekommen).

## 30: Diskriminanzanalyse: Grundlagen, Bayes

### Herleitung

Man hat Daten gegeben, deren Zugehörigkeit zu verschiedenen Klassen man kennt. Man möchte Regeln finden, die diese Klassenzugehörigkeiten widerspiegeln. Damit kann man

auch neue Beobachtungen, deren Klassenzugehörigkeit unbekannt ist, klassifizieren. Diskriminanzfunktionen sind (mathematische) Beschreibungen solcher Regeln.

Wir gehen von zwei Klassen $\pi_1$, $\pi_2$ aus. Es gibt a-priori Wahrscheinlichkeiten $p_1$ bzw. $p_2$ für die Zugehörigkeit (unbekannter) Objekte $\boldsymbol{x}$ zu $\pi_1$ bzw. $\pi_2$. Es gilt $p_1 + p_2 = 1$.

Alle möglichen Beobachtungen sind in $\Omega$ enthalten (Strichprobenraum), dieser ist partitioniert in die Teilräume $R_1$ bzw. $R_2$. Eine festgelegte Diskriminanzfunktion soll dabei jede neue Beobachtung $\boldsymbol{x}$ einem der beiden Teilräume $R_1$ oder $R_2$ zuordnen.

Beobachtungen können falsch zugeordnet, d.h. missklassifiziert, werden. Man kann die Wahrscheinlichkeiten dafür berechnen, als bedingte Wahrscheinlichkeiten (Bayes'scher Ansatz):

$$P(2|1) = P(\boldsymbol{x} \in R_2 | \boldsymbol{x} \in \pi_1) = \int_{R_2} f_1(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{42}$$

(Bedingte Wahrscheinlichkeit dafür, dass die Diskriminanzfunktion eine neue Beobachtung in den Teilraum $R_2$ einordnet (klassifiziert), wenn die Beobachtung aber tatsächlich aus der Klasse $\pi_1$ stammt.) Gleiches gilt für die „andere Richtung“:

$$P(1|2) = P(\boldsymbol{x} \in R_1 | \boldsymbol{x} \in \pi_2) = \int_{R_1} f_2(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{43}$$

Man kann 4 Fälle unterscheiden: Beobachtung jeweils richtig oder falsch als $\pi_1$ oder $\pi_2$ klassifiziert. Diese 4 gemeinsamen Wahrscheinlichkeiten (Verbundwahrscheinlichkeiten) kann man über die bedingten Wahrscheinlichkeiten und die a-priori Wahrscheinlichkeiten berechnen (Multiplikationssatz):

$$\begin{aligned} &P(\boldsymbol{x} \text{ richtig in } \pi_1 \text{ klassifiziert}) = \\ &\quad = P(\boldsymbol{x} \in R_1, \boldsymbol{x} \in \pi_1) = P(\boldsymbol{x} \in R_1 | \boldsymbol{x} \in \pi_1) P(\pi_1) = P(1|1)\, p_1 \end{aligned} \tag{44}$$

$$\begin{aligned} &P(\boldsymbol{x} \text{ falsch in } \pi_1 \text{ klassifiziert}) = \\ &\quad = P(\boldsymbol{x} \in R_1, \boldsymbol{x} \in \pi_2) = P(\boldsymbol{x} \in R_1 | \boldsymbol{x} \in \pi_2) P(\pi_2) = P(1|2)\, p_2 \end{aligned} \tag{45}$$

Analoges gilt wieder für die „andere Richtung“:

$$P(\boldsymbol{x} \text{ richtig in } \pi_2 \text{ klassifiziert}) = P(2|2)\, p_2 \tag{46}$$

$$P(\boldsymbol{x} \text{ falsch in } \pi_2 \text{ klassifiziert}) = P(2|1)\, p_1 \tag{47}$$

Man kann der Missklassifikation Kosten zuordnen. Damit kann man z.B. abbilden, dass die falsche Zuordnung zu einer Klasse „schlimmer“ oder „teurer“ ist als für eine andere, etc. Diese Kosten sind formuliert als $c(2|1) = c(\boldsymbol{x} \in R_2 | \boldsymbol{x} \in \pi_1)$ falls eine Beobachtung $\boldsymbol{x}$ durch die Diskriminanzfunktion dem Raum $R_2$ zugeordnet wird, wenn sie tatsächlich aus Klasse $\pi_1$ stammt. Analoges gilt für die „andere Richtung“.

Eine geeignete Klassifikationsregel (Diskriminanzfunktion) soll die erwarteten Kosten bei Missklassifikation (EKM) minimieren. Die EKM sind gegeben durch:

$$\mathrm{EKM} = c(2|1) P(2|1) p_1 + c(1|2) P(1|2) p_2. \tag{48}$$

Mögliche Klassifikationsregel zur Minimierung der EKM:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq () () \tag{49}$$

TO DO!

**Zwei Gruppen, MV Normalverteilung**
Multivariate Normalverteilung, 2 verschiedene Populationen bzw. Klassen. Parameter $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$.

Wir behandeln 2 Spezialfälle (CHECK: sind das LDA, QDA „mit Kosten“?):

1. $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (führt zu LDA)
2. $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ (führt zu QDA)

**LDA**
Annahme dass beide Gruppen gleiche Kovarianzsstruktur. Schätzen einer gepoolten Kovarianzmatrix (aus den gepoolten Daten beider Gruppen), sowie der beiden Gruppenmittel. Hier kann man die Daten beider Gruppen jeweils auf Mittel 0 zentrieren, dann aus den so gepoolten Daten die Kovarianzmatrix schätzen. Diskriminanzfunktion ist linear in den Daten $\boldsymbol{x}$. Die Diskriminanzfunktion benötigt die inverse der Kovarianzmatrix.

**QDA**
Die Gruppen haben unterschiedliche Kovarianzsstruktur. Schätzen der Kovarianzmatrix sowie Mittelwert per Gruppe, es müssen also mehr Parameter geschätzt werden als bei LDA. Neigt mehr zum overfitten (da mehr Parameter zu schätzen, im Vergleich z.B. zu LDA bei gleich vielen Daten). Diskriminanzfunktion ist quadratisch in den Daten $\boldsymbol{x}$. Die Diskriminanzfunktion benötigt die inverse der Kovarianzmatrix.

### 31: Diskriminanzanalyse: Fisher

Grundidee: multivariate Beobachtunten auf univariate transformieren, so dass Gruppen möglichst stark getrennt sind. D.h. dass die Mittel der resultierenden univariaten Verteilungen möglichst weit entfernt sein sollen.

Man sucht einen Vektor $\hat{\boldsymbol{a}}$, der die Beobachtungen $\boldsymbol{x}$ so linear kombiniert, dass die Trennung der resultierenden Gruppenmittel $\bar{y}_1$ bzw. $\bar{y}_2$ maximal ist.

Die Wahl von $\hat{\boldsymbol{a}}$

$$\hat{y} = \hat{\boldsymbol{a}}^\mathsf{T}\boldsymbol{x} = (\hat{\boldsymbol{x}}_1 - \hat{\boldsymbol{x}}_2)^\mathsf{T}\boldsymbol{S}_{\text{pooled}}^{-1}\boldsymbol{x}. \tag{50}$$

maximiert

$$\frac{(\hat{y}_1 - \hat{y}_2)^2}{s_y^2}. \tag{51}$$

### 32: Diskriminanzanalyse: Mehrgruppenfall, Bayes

### 33: Diskriminanzanalyse: Mehrgruppenfall Fisher

$\boldsymbol{W}$ within group variation, gewichtete Summe der Kovarianzmatrizen, mit den prior Probabilities gewichtet?

$\boldsymbol{B}$ between group variation

Die Matrizen $\boldsymbol{B}$, $\boldsymbol{W}$ sind wichtig, wie sehen diese aus? Diese führen wieder auf ein Eigenwertproblem.

### 34: Projection Pursuit

Explorative Analyse multivariater Daten. PCA ein Bsp. für PP, die Varianz ist dann der Projektionsindex.

Proejtkion von Datenpunkten auf niedrigdimensionaleren Raum (1D oder 2D, typischerweise).

$$\boldsymbol{y} = \boldsymbol{A}^\top \boldsymbol{x}. \tag{52}$$

Man sucht nach einer Projektionsmatrix $\boldsymbol{A}$ die einen Interessantheitsindex maximiert.

Man kann z.B. Projektion suchen die eine starke Struktur aufweisen, z.B: Abweichung von der Normalverteilung.