Theoretische Grundlagen der partiellen kleinsten Quadrate

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

von B.Sc./D.H.S. Hayan Hasan

geb. am 12.04.1976 Safita-Syrien

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gutachter: Prof.Dr.Rainer Schwabe

Prof.Dr.Rainer Schlittgen

eingereicht am: 07.01.2014

Verteidigung am: 06.06.2014

Inhaltsverzeichnis

Danksagung

Mein Dank gilt an dieser Stelle Herrn Prof. Schwabe, der diese Arbeit betreut und nach Fertigstellung begutachtet hat. Ebenso danke ich Herrn Prof. Schlittgen für die Erstellung des Gutachtens.
Den Professoren und Mitarbeitern des Instituts für Mathematische Stochastik danke ich für die Möglichkeit, im Rahmen des Oberseminars zu meinem Thema vortragen zu können. Verschiedene Hinweise und Anregungen. in diesem Rahmen haben ebenfalls zur Entstehung der Arbeit beigetragen.
Für eine schöne, abwechslungsreiche freundschaftliche Arbeitsatmosphäre, viele wertvolle Anregungen und stete Hilfsbereitschaft, die wesentlich zum Gelingen dieser Arbeit beigetragen haben, möchte ich mich den Kolleginnen und Kollegen bedanken.
Bedanken möchte ich mich auch bei Frau Altenkirch für alles. Der besten Freundin Rita danke ich aus ganzem Herzen für alles.
Weiterhin möchte ich den Menschen danken, die mir das alles letztendlichmöglich gemacht haben. Für die aufopferungsvolle Unterstützung auf den vielen Wegen, die ich bis zu diesem Punkt gehen musste, bedanke ich mich bei meinen Eltern, meinen Brüdern, meinen Schwestern und meinen Freunden.
Bei der Aleppo-Universität, bedanke ich mich für die finanzielle Unterstützung meines Studiums, ohne die diese Arbeit so nicht möglich gewesen wäre.

Zusammenfassung

Die Methode der partiellen kleinsten Quadrate (PLS = partial least squares) ist ein aktuelles Verfahren zur Dimensionsreduktion und Vorhersage in hochdimensionalen Datensätzen, wie sie z.B. in den Neurowissenschaften bei MRT-Daten zur Analyse von Hirnaktivitäten oder bei der Bildanalyse vorliegen. Es gibt eine Reihe sehr unterschiedlicher Algorithmen, die unter dem Begriff partial least squares zusammengefasst werden. Die statistischen Eigenschaften der Methode der partiellen kleinsten Quadrate sind jedoch bisher wenig erforscht. Dies liegt zu einem nicht unwesentlichen Teil an der Problematik, dass diese Methode der partiellen kleinsten Quadrate bisher nur unzureichend mit wahrscheinlichkeitstheoretischen Mitteln beschrieben werden kann.
In der vorliegenden Arbeit befassen wir uns mit einem Ansatz zur Entwicklung theoretischer Grundlagen der Methode der partiellen kleinsten Quadrate und wenden diese zur Vorhersage und zur Dimensionsreduktion an. Hierzu führen wir zuerst eine etwas allgemeinere Theorie der Vorhersage ein, wie sie in der Literatur bisher wenig behandelt wurde. Dazu geben wir zuerst einen Abriss der für die Arbeit benötigten Grundlagen der mathematischen Statistik. Dies beinhaltet insbesondere die Eigenschaften des besten erwartungstreuen Schätzers, d.h. desjenigen mit gleichmäßig kleinster Varianz. Danach stellen wir die in dieser Arbeit behandelten Assoziationse wesentlichen Eigenschaften der besten erwartungstreuen Vorhersage. Hierzu wird ein allgemeines Kriterium zur Ermittlung der besten Vorhersage abgeleitet. Im Anschluss werden einige aus der Literatur bekannte mathematische Charakterisierungen vorgestellt und durch neue ergänzt sowie ein neues analytisches Verfahren zur Verbesserung der Vorhersage entwickelt und auf seine Tauglichkeit überprüft. Anschließend wird mit Hilfe einer Simulationsstudie unter der Verteilungsannahme einer multivariaten Normalverteilung mit autoregressiver Kovarianzstruktur am Beispiel des Vorhersageintervalls die Qualität des neu entwickelten modifizierten Verfahrens mit dem Standardansatz der partiellen kleinsten Quadrate sowie mit Methoden der Hauptkomponentenanalyse und der kanonischen Korrelationsanalyse verglichen. Als Ergebnis demonstriert die Simulationsstudie die Vorteile des neu vorgeschlagenen modifizierten Ansatzes gegenüber den bekannten Methoden. Die abschließende Diskussion fasst die Ergebnisse zusammen und gibt einen Ausblick auf mögliche Erweiterungen des vorgestellten Konzepts.

Summary

Partial least squares is a topic method for dimension reduction and for prediction in high-dimensional data sets, for example in neuroscience with fMRI data for the analysis of brain activity. There are a number of different algorithms using the PLS approach. However, until now the statistical properties of the PLS method are not well investigated. This may be caused by the problem that the method of partial least squares is difficult to be described in terms of probability theory.

The aim of the present thesis is to develop a theoretical background of the Partial Least Squares methodology and to apply this approach to prediction and dimension reduction. In this context we also address a general theory of prediction, which has been neglected in the literature. We first give a summary of the background of mathematical statistics required for our purposes. In particular, properties of the uniformly minimum variance unbiased estimator are introduced. After that we present an introduction to the relevant association structures. Based on this we discuss the theory of prediction and, in particular, the properties of the best unbiased prediction. For this a general charcterization for the best prediction is derived. Then we present and analyze some mathematical characterizations of the partial least squares approach given in the literature as well as some new ones and develop analytical solutions to improve the precision of the prediction. Subsequently we provide a simulation study, in which the performance of the newly proposed modified partial least squares method is compared with that of the standard least squares method as well as with the competing methods of principal component and canonical regression analysis under the model assumption of a multivariate normal distribution with an autoregressive covariance structure in the example of the construction of a prediction interval. As a result the simulation study demonstrates the advantages of the newly suggested modified approach compared to the methods existing so far in the literature. We conclude with a discussion of the results and an outlook to possible future work.

Einleitung

In vielen Anwendungsbereiche der Statistik wie der Medizin oder Ökonomie werden in zunehmendem Maße große und insbesondere hochdimensionale Datensätze erhoben, um damit interessierende Fragestellungen untersuchen zu können. Verfahren zur Dimensionsreduktion werden in der Statistik vielfältig verwendet:

1. Um die Daten auf geeignete Weise in einen Raum mit niedrigerer Dimension zu transformieren, in welchem die weitere Analyse und statistische Aufgaben (Punktschätzung, Intervallschätzung, Test, und Vorhersage) vorgenommen werden.

2. Um bessere Anpassung an die Daten und einfache Modelle zu erreichen.

3. Zum Einsparen von Speicherplatz

4. Zur Visualisierung und damit zum besseren Erkunden der Daten.

Die Regressionsanalyse stellt ein bekanntes Verfahren zur Dimesionsreduktion dar. Andere Methoden, die mit dem Begriff der Dimensionsreduktion verbunden werden, sind die Hauptkomponentenanalyse (engl. principal component analysis, PCA) und die kanonische Korrelatioin (engl. canonical correlation, CC). Ein aktuelleres Verfahren zur Dimensionsreduktion und Vorhersage in hochdimensionalen Datensätzen, wie sie z.B. in den Neurowissenschaften bei MRT-Daten zur Analyse von Hirnaktivitäten oder bei der Bildanalyse vorliegen, ist die Methode der partiellen kleinsten Quadrate (PLS = partial least squares). Diese Methode wurde ursprünglich von H. Wold (1966) eingeführt und wird in verschiedenen Anwendungsgebieten zur Regressions- oder Vorhersage für die Analyse hochdimensionaler Daten eingesetzt. Die statistischen Eigenschaften der Methode der partiellen kleinsten Quadrate sind jedoch bisher wenig erforscht. Dies liegt zu einem nicht unwesentlichen Teil an der Problematik, dass diese Methode der partiellen kleinsten Quadrate bisher nur unzureichend mit wahrscheinlichkeitstheoretischen Mitteln beschrieben werden kann. Ein mathematisches Modell des PLS-Ansatzes wurde von Helland (1988, 1990) eingeführt. Einige Erweiterungen des Resultates von Helland in mehrdimensionalen Modelle und Anwendungen für Klassifikation wurden von Liu und Rayens (2007) betrachtet. Die Shrinkage-Eigenschaft wurde von Krämer (2007) untersucht. Eine Erweiterung der PLS-Methode wurde von Li, Udén und v.Rosen (2013) modifiziert. In der vorliegenden Arbeit befassen wir uns mit einem Ansatz zur Entwicklung theoretischer Grundlagen der Methode der partiellen kleinsten Quadrate und wenden diese zur Vorhersage und zur Dimensionsreduktion an. Hierzu führen wir zuerst eine etwas allgemeinere Theorie der Vorhersage ein. Dazu geben wir zuerst einen Abriss der für die Arbeit benötigten Grundlagen der mathematischen Statistik. Dies

beinhaltet insbesondere die Eigenschaften des besten erwartungstreuen Schätzers, d.h. desjenigen mit gleichmäßig kleinster Varianz. Im zweiten Kapitel stellen wir die in dieser Arbeit behandelten Assoziationsstrukturen vor. Darauf aufbauend entwickeln wir im dritten Kapitel die Theorie der Vorhersage und leiten die wesentlichen Eigenschaften der besten erwartungstreuen Vorhersage her. Das vierte Kapitel behandelt die Vorhersage im linearen Modell. Im fünften Kapitel werden mathematische Charakterisierungen der PLS-Methode vorgestellt und durch neue ergänzt sowie ein neues analytisches Verfahren zur Verbesserung der Vorhersage entwickelt und auf seine Tauglichkeit überprüft. Anschließend wird mit Hilfe einer Simulationsstudie unter der Verteilungsannahme einer multivariaten Normalverteilung mit autoregressiver Kovarianzstruktur am Beispiel des Vorhersageintervalls die Qualität des neu entwickelten modifizierten Verfahrens mit dem Standardansatz der partiellen kleinsten Quadrate sowie mit Methoden der Hauptkomponentenanalyse und der kanonischen Korrelationsanalyse verglichen. Als Ergebnis demonstriert die Simulationsstudie die Vorteile des neu vorgeschlagenen modifizierten Ansatzes gegenüber den bekannten Methoden. Zusammenfassend kann gesagt werden, dass diese Arbeit das Ziel verfolgt, und diese Interpretation bietet mögliche Erweiterungen der vorgestellten Konzepte für verschiedene Modelle zu entwickeln.

1. Allgemeine Grundlagen der mathematischen Statistik

In diesem Kapitel stellen wir die wichtigsten mathematischen Begriffe vor, auf denen die Entwicklungen in den folgenden Kapiteln beruhen. Dazu befassen wir uns mit Theorie und Anwendung der statistischen Modelle. Schließlich geben wir einen kurzen Überblick über die wichtigsten Resultate zur Parameterschätzung im linearen Normalverteilungsmodell und im Modell mit multivariaten, u.i.v. normalverteilten ZV'en und verweisen für Beweise und weitere Ergebnisse auf die Literatur, z.B. Shao (2003), Alvin (2008), Christensen (2011), Christensen und Lin (2013) und Hocking (2005). Zuletzt beleuchten wir kurz das AR(1)-Modell.

1.1. Statistisches Experiment

Grundlage der mathematischen Behandlung einer statistischen Problemstellung bildet das statistische Experiment welches das den Beobachtungen zugrundegelegte Modelle beschreibt.

Definition 1.1 ***(statistisches Experiment)***
Ein statistisches Experiment ist ein Tripel $E = \left(M, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\right)$ *bestehend aus*

- *einer nichtleeren Menge* M
- *einer* σ*-Algebra* $\mathcal{B}$ *über* M*, die die beobachtbaren Ereignisse enthält .*
- *einer Familie* $(P_\theta)_{\theta\in\Theta}$ *von Wahscheinlichkeitsmaßen auf* $(M, \mathcal{B})$*, die mit Element des Parameterraums* Θ *parametrisiert ist. Mit jedem statistischen Experiment kann die Beobachtung einer Zufallsvariablen*

  $$X : (\Omega, \mathcal{C}) \to (M, \mathcal{B})$$

  und eine Familie $(Q_\theta)_{\theta\in\Theta}$ *von W-Maßen auf* $(\Omega, \mathcal{C})$ *verbunden werden, so dass*

  $$Q_\theta^X = P_\theta \qquad \text{für alle } \theta \in \Theta.$$

Die Formulierung von statistischen Experimenten wird anhand der folgenden Beispiele illustriert.

Beispiele für statistische Experimente

Das statistische Experiment beschreibt stets das Ergebnis eines Zufallsexperiments, etwa die Werte einer erhaltenen Stichprobe oder gesammelte Messergebnisse eines Experiments. Somit ist die Verteilung der Zufallsvariable das Schlüsselelement. Das statistische Experiment ist dann eine geeignete Familie von solchen Verteilungen. Anhand von drei Beispielen wird im Folgenden die Formulierung von statistischen Experimenten illustriert.

(1) Vereinfachtes Normalverteilungsmodell
Wir betrachten das Modell mit $n \geq 2$ u.i.v.$N(\beta, \sigma^2)$-verteilten Zufallsvariablen $X_1, X_2, \ldots, X_n$, wobei $\beta \in \mathbb{R}$ und $\sigma^2 \in (0, \infty)$ die Parameter sind. Mit $\mathbf{X}$ bezeichen wir die $\mathbb{R}^n$-wertige Zufallsvariable $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$. Damit hat man das statistische Experiment

$$E = \Big(\mathbb{R}^n, \mathcal{B}^n, \big(\otimes_{i=1}^n N(\beta, \sigma^2)\big)_{(\beta,\sigma^2)\in\mathbb{R}\times(0,\infty)}\Big) \tag{1.1}$$

(2) Lineares Normalverteilungsmodell
Im Folgenden sei ein lineares Normalverteilungsmodell zu Grund gelegt:

$$\mathbf{X} \sim (\mathbf{B}\boldsymbol{\beta}, \sigma^2\mathbf{V}), \qquad \mathbf{V} \text{ bekannt}$$

Die Zufallsvariable $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ des Modells ist $\mathbb{R}^n$-wertig, wobei $\mathbf{B}$ eine gegebene reelle $n \times p$ Matrix vom Rang p, und $n > p$ vorausgesetzt ist, mit $\mathbf{V} \in PD(n)$ (Die Menge aller positiv definiten reellen $(n \times n)$-Matrizen wird mit PD(n) abgekürzt). Der Parameter ist $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T \in \mathbb{R}^p \times (0, \infty)$

$$E = \Big(\mathbb{R}^n, \mathcal{B}^n, N_n\big(\mathbf{B}\boldsymbol{\beta}, \sigma^2\big)\big)_{(\boldsymbol{\beta},\sigma^2)\in\mathbb{R}^p\times(0,\infty)}\Big) \tag{1.2}$$

(3) Modell mit multivariaten, u.i.v.normalverteilten ZV'en
Seien $X_1, X_2, \ldots, X_n$ stochastisch unabhängig und identisch $N_p(\boldsymbol{\beta}, \mathbf{V})$-verteilte $\mathbb{R}^p$-wertige Zufallsvariablen, wobei $(\boldsymbol{\beta}, \mathbf{V}) \in \mathbb{R}^p \times PD(p)$ die Parameter sind. Kurz:

$$E = \Big((\mathbb{R}^p)^n, (\mathcal{B}^p)^n, (\otimes_{i=1}^n N(\boldsymbol{\beta}, \mathbf{V}))_{(\boldsymbol{\beta},\mathbf{V})\in\mathbb{R}^p\times PD(p)}\Big) \tag{1.3}$$

Um die Aufgabenstellungen weiter zu erlätern und nachfolgende Begriffsbildungen vorzunehmen, werden wir im weiteren Verlauf unserer Betrachtungen neben dem statistischen Experiment einen weiteren Begriff benötigen, und zwar den der statistischen Entscheidungsprobleme

Definition 1.2 ***(statistisches Entscheidungsproblem)***
Ein statistisches Entscheidungsproblem ist ein Tripel (E, A, L) *bestehend aus*

- *einen statistischen Experiment* $E = \big(M, \mathcal{B}, (P_\theta)_{\theta\in\Theta}\big)$

- *einem Antwortraum* A
- *einer Verlustfunktion* $L : A \times \Theta \longrightarrow [0, \infty)$

Falls die Parametermenge eine Teilmenge des $\mathbb{R}^d$ ist, spricht man von einem parametrischen, ansonsten von einem nichtparametrischen Modell. Die Zustandsräume

$$\mathbb{R}^p \times (0, \infty), \qquad \mathbb{R}^p \times PD(p)$$

implizieren zum Beispiel parametrische Modelle. In dieser Arbiet beschränken wir uns im Wesentlichen auf parametrische Modelle.

1.2. Schätzproblem

Wir gehen wieder von einem allgemeinen statistischen Experiment aus,

$$E = \left(M, \mathcal{B}, \left(P_\theta^X\right)_{\theta \in \Theta}\right)$$

desweiteren sei ein Aspekt des Parameters gegeben, der einen interssierenden Teil-Parameter $g(\theta)$ beschreibt, $g : \Theta \to A$, wir nennen den Aspekt $g(\theta), \theta \in \Theta$ ebenfalls einen Parameter. Unter einem Schätzer für $g(\theta)$ verstehen wir eine Funktion

$$T : (M, \mathcal{B}) \longrightarrow (A, \mathcal{A}),$$

die also für jeden möglichen Wert $x \in M$ der Zufallsvariablen X die Schätzung $T(x)$ für $g(\theta)$ liefert, und T messbar ist. Sei

$$D = \{T | T : M \longrightarrow A \; messbar, E(T^2(X)) < \infty)\} \tag{1.4}$$

die Menge aller Schätzer. Um den stochastischen Kontext zu betonen, unterscheiden wir:

- $T(X)$ Schätzer für $g(\theta)$ Zufallsvariable auf $(M, \mathcal{A})$ (und die im Wessentlichen dasselbe wie T ist).
- $T(x)$ Schätzung für $g(\theta)$ Realisation.

Mit der Bezeichnung T statt $\widehat{g}$ eines Schätzers für $g(\theta)$ haben wir einen kleinen Notationswechsel vorgenommen, der natürlich inhaltlich Bedeutung hat. In dieser Arbeit betrachten wir (und vergleichen wir) mehrere/viele Schätzer für denselben Parameter $g(\theta)$, und Bezeichnungen wie T, T_1, T^* etc, für verschiedene Schätzer für $g(\theta)$. Das Ziel ist es. die Qualität eines Schätzers $T(X)$ für den Parameter $g(\theta)$ zu messen. Als Messung des Fehlers stellen wir zunächst die Verlustfunktion im ein- mehrdimensionale Fall vor und behandeln danach Risikofunktion. Abschließend wird sein, die Qualität des Schätzers anhand seiner Streuung zu messen.

Einige Verlustfunktionen:

(i) Der eindimensionale Fall ($A = \mathbb{R}$):
$L(z, \theta) = |z - g(\theta)|$ (absolute Abweichung),
$L(z, \theta) = (z - g(\theta))^2$ (quadratische Abweichung),
$L(z, \theta) = \frac{(z-g(\theta))^2}{g(\theta)^2}$ (relative quadratische Abweichung),
$L(z, \theta) = (w(\theta))^2(z - g(\theta))^2$ gewichtete quadratische Abweichung , mit einer gegebenen Gewichtsfunktion $w : \Theta \longrightarrow [0, \infty)$.

(ii) Der mehrdimensionale Fall ($A = \mathbb{R}^p$):
$L(\mathbf{z}, \boldsymbol{\theta}) = (\mathbf{z} - \mathbf{g}(\boldsymbol{\theta}))^T(\mathbf{z} - \mathbf{g}(\boldsymbol{\theta}))$ (quadratische Abweichung),
$L(\mathbf{z}, \boldsymbol{\theta}) = \frac{(\mathbf{z}-\mathbf{g}(\boldsymbol{\theta}))^T(\mathbf{z}-\mathbf{g}(\boldsymbol{\theta}))}{\mathbf{g}(\boldsymbol{\theta})^T\mathbf{g}(\boldsymbol{\theta})}$ (relative quadratische Abweichung),
$L(\mathbf{z}, \boldsymbol{\theta}) = (\mathbf{z} - \mathbf{g}(\boldsymbol{\theta}))^T\mathbf{W}(\boldsymbol{\theta})(\mathbf{z} - \mathbf{g}(\boldsymbol{\theta}))$ gewichtete quadratische Abweichung, mit einer gegebenen Gewichstransformation $\mathbf{W} : \Theta \longrightarrow PD(p)$.

1.2.1. Optimalitätskriterien

Zum Vergleich von Schätzer interessieren wir uns oft für die Frage: wie soll man das Verhalten eines Schätzers $T(X)$ für $g(\theta)$ messen? Daher gehen wir von einer gegebenen Verlustfunktion L aus

$$L : A \times \Theta \longrightarrow [0, \infty)$$

Der Vergleich verschiedener Schätzer für $g(\theta)$ orientiert sich an deren Risikofunktion.

Definition 1.3 ***(Risikofunktion)***
Sei (E,A,L) ein statistisches Entscheidungsproblem. Dann heißt

$$R(T(X), g(\theta)) = E_\theta(L(T(X), g(\theta))) \tag{1.5}$$

die Erwartung von einer Verlustfunktion des Schätzers $T(X)$ die Risikofunktion des Schätzers $T(X)$.

Das Ziel wird sein, ein Kriterium zu finden, welches bereits vor der Datenerhebung zur Beurteilung eines Schätzers genutzt werden kann. Dafür kommt unter anderem das im Folgenden vorgestellte Maß des mittleren quadratischen Fehlers in Frage.
Der mittlere quadratische Fehler, kurz MSE (engl. mean squared error), ist ein Gütemaß für den Punktschätzer. Er setzt sich zusammen aus dem Bias und der Varianz des Punktschätzers. Hauptsächlich betrachten wir in diesem Abschnitt den quadratischen Verlust.

Definition 1.4 ***(mittlerer quadratischer Fehler)***
Der mittlere quadratische Fehler eines Punktschätzers $T \in D$ für einen Parameter $g(\theta)$, ist definiert als

$$MSE_\theta(T(X), g(\theta)) = \mathbb{R}(T(X), g(\theta)) = E_\theta[(T(X) - g(\theta))^2] \tag{1.6}$$

Weiterhin heißt

$$b_\theta(T(X), g(\theta)) = E_\theta(T(X)) - g(\theta))$$

die Bias von T bzgl g. Gilt $b_\theta(T(X), g(\theta)) = 0, \ \forall\ \theta \in \Theta$, wenn T unverzerrt ist.

Daraus als Korollar zur Definition 1.4 ergibt sich die folgende wichtige Zerlegung des mittleren quadratischen Fehlers in Varianz des Schätzers und Quadrat des Biases:

$$\begin{aligned} MSE_\theta(T(X), g(\theta)) &= Var_\theta(T(\theta)) + [E_\theta(T(X)) - g(\theta)]^2 \\ &= Var_\theta(T(X)) + b^2_\theta(T(X), g(\theta)). \end{aligned}$$

Interessant in der statistischen Theorie ist der mehrdimensionale Fall, daher betrachten wir die Verallgemeinerungen der mittleren quadratischen Abweichung. Die Zweckmäßigkeit von verzerrten Schätzer im mehrparametrigen Fall lässt sich aus dem MSE-Kriterium herleiten. Es ist definiert durch:

Definition 1.5 ***(R^p-wertiger Parameter und MSE)*** *Schlittgen (2009, S.98,99)*
Im Fall $\mathbf{g}(\boldsymbol{\theta}) \in \mathbb{R}^p$ *nennt man*

(a) *den Ausdruck*

$$MSEM_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}}[(\mathbf{T}(\mathbf{X}) - \mathbf{g}(\boldsymbol{\theta}))(\mathbf{T}(\mathbf{X}) - \mathbf{g}(\boldsymbol{\theta}))^T]$$

die Matrix des mittleren quadratischen Fehlers.

(b)

$$MSE_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}}(\mathbf{T}(\mathbf{X}) - \mathbf{g}(\boldsymbol{\theta}))^T(\mathbf{T}(\mathbf{X}) - \mathbf{g}(\boldsymbol{\theta})) \tag{1.7}$$

den mittleren quadratischen Fehler

Beachte:

- Die Matrix des mittleren quadratischen Fehlers hat die einzelnen MSE's auf der Hauptdiaonalen.
- Er lässt sich durch Ergänzen und Ausmultiplizieren einfach umformen zu

$$\begin{aligned} MSE_\theta &= E_\theta[(\mathbf{T}(\mathbf{X}) - \mathbf{g}(\theta))^T((\mathbf{T}(\mathbf{X}) - \mathbf{g}(\theta))] \\ &= Spur(E_\theta[(\mathbf{T}(\mathbf{X}) - \mathbf{g}(\theta))((\mathbf{T}(\mathbf{X}) - \mathbf{g}(\theta))^T]) \\ &= \sum_{i=1}^{p} MSE_\theta\big(T_i(\mathbf{X}), g_i(\boldsymbol{\theta})\big) \quad (T_i, g_i \text{ der i-ten Komponente}) \\ &= \sum_{i=1}^{p} Var(T_i(\mathbf{X})) + (E[(\mathbf{T}(\mathbf{X})] - \mathbf{g}(\boldsymbol{\theta}))^T (E[(\mathbf{T}(\mathbf{X})] - \mathbf{g}(\boldsymbol{\theta})) \end{aligned}$$

1.2.2. Eigenschaften des UMVUE-Schätzers

UMVUE steht für engl. Uniformly Minimum Variance Unbiased Estimator. Für eine vorgegebne Funktion g des Parameters θ betrachten wir die Menge aller erwartungstreuen Schätzer für $g(\theta)$ mit endlicher Varianz

$$U_g = \{T | T : M \longrightarrow A \; messbar, \; E_\theta(T(X)) = g(\theta), \; E_\theta(T^2(X)) < \infty\}$$

und die Menge aller Schätzer für $g(\theta)$ mit endlicher Varianz und Erwartunswert gleich null

$$U_0 = \{T|T : M \longrightarrow A \; messbar, \;\; E_\theta(T(X)) = 0, \; E_\theta(T^2(X)) < \infty\}.$$

Definition 1.6 ***(UMVUE-Schätzer)***
Eine Statistik $T^* \in U_g$ heißt UMVUE, wenn gilt

$$Var(T^*) = \inf\{Var(T) : \; T \in U_g \; \}.$$

(a) Der Satz von Rao-Blackwell sagt aus, dass durch Bildung der bedingten Erwartung bezüglich einer suffizienten Statistik eines Schätzers mit gleichmäßig nicht größem Risiko entsteht.

Theorem 1.1 ***(Rao-Blackwell-Verbesserung)*** *Witting (1985, S. 349)*
Seien $T(X)$ eine suffiziente Statistik und $S(X)$ ein Schätzer mit $S \in U_g$ (S im Wessentlichen ist dasselbe wie $S(X)$). Definiere die Rao- Blackwell-Verbesserung $S^*(X) = E[S(X)|T(X)]$. Dann gilt:

1. *$S^* \in U_g$*
2. *$Var_\theta(S^*(X)) \leq Var_\theta(S(X)), \forall\theta \in \Theta$, d.h. S^* ist besser als S.*

(b) Wenn die Statistik $T(X)$ suffizient sogar vollständig ist, so für $S \in U_g$ liefert die Rao-Blackwell-Verbesserung $S^*(X) = E[S(X)|T(X)]$ bereits ein UMVUE-Schätzer, wie der nächste Satz von Lehmann-Scheffé aussagt:

Theorem 1.2 ***(Lehmann-Scheffé-Verbesserung)*** *Witting (1985, S. 354)*
Sei $T(X)$ vollständige suffiziente Statistik und $S \in U_g$. Dann ist der Schätzer

$$S^*(X) = E[S(X)|T(X)]$$

gleichmäßig optimal in der Menge aller erwartungstreuen Schätzer für $g(\theta)$, d.h. $S^*(X)$ ist ein erwartungstreuer Schätzer für $g(\theta)$ mit gleichmäßig minimaler Varianz (UMVUE).

Der Satz von Lehmann-Scheffé liefern ein Verfahren, einen gleichmäßig besten erwartungstreuen Schätzer zu finden. Mittels der Fisher-Information $I(\theta)$ ergibt sich nun eine untere Schranke für die Varianz eines Schätzers $T^* \in U_g$.

(c) Theorem 1.3 ***(Fisher-Information und UMVUE)*** *Shao (2003 S. 171,172)*
Falls $\mathcal{P} = \{f_\theta, \theta \in \Theta\}$ eine exponentielle Familie mit Darstellung

$$f_\theta(x) = \exp\left(c(\theta)T(x) + d(\theta) + l(x)\right)$$

ist, und $T^*(X)$ ein erwartungstreuer Schätzer für $g(\theta)$, so dass

$$Var_\theta(T^*(X)) = \frac{(\frac{\partial}{\partial\theta}g(\theta))^2}{I(\theta)}, \forall\theta \in \Theta. \tag{1.8}$$

Dann

1. *$T^*(x)$ UMVUE-Schätzung für $g(\theta)$.*
2. *$T^*(x)$ lineare Funktion in $T(x)$*

Die Optimalität der erwartungstreuen Schätzer lässt sich mit der Kovarianzmethode charakterisieren

(d) Theorem 1.4 ***(Kovarianzmethode)*** *Shao (2003, Kap. 3.1.2)*
Sei $T^* \in U_g$ erwartungstreuer Schätzer, dann gilt:

$$Cov_\theta(T^*(X), T_0(X)) = 0, \forall \theta \in \Theta \text{ und } \forall\ T_0 \in U_0 \ \Leftrightarrow\ T^*(X) \text{ UMVUE für } g(\theta) \tag{1.9}$$

Aus Theorem 1.4 ergibt sich insbesondere die folgenden Folgerung.

Folgerung 1.1 *Sei $T^* \in U_g$ ein erwartungstreuer Schätzer, dann gilt*

$$E_\theta(T^*(X)T_0(X)) = 0,\ \forall T_0 \in U_0 \quad \text{und} \quad \forall \theta \in \Theta \quad \Leftrightarrow \quad T^*(X) \text{ UMVUE}\,. \tag{1.10}$$

Beachte:

- Wenn T^* UMVUE-Schätzer ist, gilt

  $$Var_\theta(T^*) = Cov_\theta(T^*, T),\ \forall T \in U_g.$$

  und die Varianz von $T - T^*$ ist gleich die Differenz der Varianzen

  $$\begin{aligned} Var_\theta(T - T^*) &= Var_\theta(T^*) + Var_\theta(T) - 2\, Cov_\theta(T, T^*) \\ &= Var_\theta(T) - Cov_\theta(T^*, T). \\ &= Var_\theta(T) - Var_\theta(T^*). \end{aligned}$$

- Hieraus ergibt sich insbesondere

  $$\frac{Var_\theta(T - T^*)}{Var_\theta(T)} = 1 - Corr^2_\theta(T, T^*).$$

  Mit der Schreibweise

  $$Corr^2_\theta(T, T^*) = 1 - \frac{Var_\theta(T - T^*)}{Var_\theta(T)}\ \text{, in dem}\ \ 0 \le Corr^2_\theta(T, T^*) \le 1.$$

  ergibt sich:

  $$Corr^2_\theta(T, T^*) = 1 \Longleftrightarrow Var_\theta(T - T^*) = 0 \quad \text{d.h. } T = T^* \quad P_\theta \text{ fast sicher}$$

- Wenn T_1, T_2 zwei UMVUE-Schätzer für den Parameter $g(\theta)$ sind, dann ist $T_1 = T_2$ P_θ fast sicher, da

  $$\begin{aligned} E_\theta[(T_1 - T_2)^2] &= E_\theta[T_1(T_1 - T_2)] + E_\theta[T_2(T_2 - T_1)] \\ &= 0 \quad \text{(Folgerung(1.1))} \end{aligned}$$

Beispiel 1.1 ***Vereinfachtes Normalverteilungsmodell:*** *Betrachte das Modell*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\beta, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \text{ mit } \mathbf{V} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ bekannt und } |\rho| < 1.$$

Parameter des Modells ist: $\beta \in \mathbb{R}$.

- *Der gewichtete kleinste Quadrate Schätzer(WLS), bzw. ML-Schätzer ist:*

$$\begin{aligned} \widehat{\beta}_{WLS} &= \arg\min_{\beta}(\mathbf{X} - \mathbf{1}_2\beta)^T\mathbf{V}^{-1}(\mathbf{X} - \mathbf{1}_2\beta) \\ &= (\mathbf{1}_2^T\mathbf{V}^{-1}\mathbf{1}_2)^{-1}\mathbf{1}_2^T\mathbf{V}^{-1}\mathbf{X} \\ &= \left(\mathbf{1}_2^T\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\mathbf{1}_2\right)^{-1}\mathbf{1}_2^T\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\mathbf{X} \\ &= \frac{1}{2(1-\rho)}(1-\rho)(X_1 + X_2) \\ &= \bar{X} \end{aligned}$$

- *Der gewöhnliche kleinste Quadrate Schätzer (OLS) ist:*

$$\begin{aligned} \widehat{\beta}_{OLS} &= \arg\min_{\beta}(\mathbf{X} - \mathbf{1}_2\beta)^T(\mathbf{X} - \mathbf{1}_2\beta) \\ &= (\mathbf{1}_2^T\mathbf{1}_2)^{-1}\mathbf{1}_2^T\mathbf{X} \\ &= \frac{(X_1 + X_2)}{2} \\ &= \bar{X} \end{aligned}$$

Interpretation im Fall $A = \mathbb{R}^p$
Im Fall $A = \mathbb{R}^p$ steht MCUE für Minimum Covarince unbiased estimator. Betrachtet man nur unverzerrte Schätzer, so kann man die Karianz des Schätzers als Maß für die Qualität des Schätzers heranziehen, da unter Unverzerrtheit die Karianz des Schätzers gleich der mittleren quadratischen Fehler ist.

Definition 1.7 ***(Kovariaz-Matrix des Schätzers)*** *Rao, Shalabh, Toutenburg und Heumann (2008, Kap.3.4)*
Sei $\mathbf{T} = (T_1, \ldots, T_p)^T$ *ein unverzerrter Schätzer für den* $\mathbb{R}^p$*-Parameter*

$$\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta}))^T.$$

Dann heißt die $p \times p$*-Matrix* $Cov_\theta(\mathbf{T}) = E_{\boldsymbol{\theta}}\left[(\mathbf{T} - \mathbf{g}(\boldsymbol{\theta}))(\mathbf{T} - \mathbf{g}(\boldsymbol{\theta}))^T\right]$ *Kovariaz-Matrix von* $\mathbf{T}$.

Ein Schätzer ist in diesem Sinn besser als alle anderen unverzerrten Schätzer, falls seine Kovarianz minimal ist, was zu folgender Optimalitätseigenschaft führt.

Definition 1.8 ***(MCUE-Schätzer)*** *Rao, Shalabh, Toutenburg und Heumann (2008, Kap.3.4)*
Ein Schätzer $\mathbf{T} \in U_g$ *von* $\mathbb{R}^p$*-Parameter* $\mathbf{g}(\boldsymbol{\theta})$ *heißt MCUE, wenn*

$$Cov_{\boldsymbol{\theta}}(\mathbf{T}) \leq_{psd} Cov_{\boldsymbol{\theta}}(\mathbf{S}), \quad \forall\, \boldsymbol{\theta} \in \Theta \quad und\, \forall\, \mathbf{S} \in U_g. \tag{1.11}$$

gilt, wobei $\leq_{psd}$ *die Anordnung im Sinne der positiven Definitheit ist.*

Lemma 1.1 *Rao, Shalabh, Toutenburg und Heumann (2008, Kap.3.4)*
Die folgenden Bedingungen sind äquivalent:

1. $\mathbf{T}^* = (T_1^*, \ldots, T_p^*)^T$ *ein MCUE-Schätzer für den* $\mathbb{R}^p$*-Parameter*

$$\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta}))^T$$

2. $\mathbf{c}^T\mathbf{T}^*$ *ist UMVUE-Schätzer für den Parameter* $\mathbf{c}^T\mathbf{g}(\boldsymbol{\theta}), \forall \mathbf{c} \in \mathbb{R}^p$.

3. $Cov_{\boldsymbol{\theta}}(\mathbf{T}_0, \mathbf{T}^*) = \mathbf{0}_{q\times q}, \forall\ \mathbf{T}_0 \in U_0$

Bemerkung 1.1 *Wenn* $\mathbf{T}^*$ *MCUE-Schätzer ist, dann gilt:*

$$Cov_{\boldsymbol{\theta}}(\mathbf{T}^*) = Cov_{\boldsymbol{\theta}}(\mathbf{T}^*, \mathbf{T}), \quad \forall \mathbf{T} \in U_g$$

$$\begin{aligned} Cov_{\boldsymbol{\theta}}(\mathbf{T} - \mathbf{T}^*) &= Cov_{\boldsymbol{\theta}}(\mathbf{T}^*) + Cov_{\boldsymbol{\theta}}(\mathbf{T}) - 2\ Cov_{\boldsymbol{\theta}}(\mathbf{T}, \mathbf{T}^*) \\ &= Cov_{\boldsymbol{\theta}}(\mathbf{T}) - Cov_{\boldsymbol{\theta}}(\mathbf{T}^*, \mathbf{T}). \\ &= Cov_{\boldsymbol{\theta}}(\mathbf{T}) - Cov_{\boldsymbol{\theta}}(\mathbf{T}^*) \geq 0. \end{aligned}$$

Bemerkung 1.2 ***(BLUE-Schätzer)***
Ein linearer UMVUE-Schätzer (MCUE-Schätzer) heißt BLUE-Schätzer (best linear unbiased estimator)

1.3. Verteilungsmodelle

In diesem Abschnitt stellen wir $\mathbb{R}^{p+q}$-wertige verteilte Zufallsvariablen (ZV) vor.

(a) Seien $\mathbf{X} : (M, \mathcal{A}) \longrightarrow (\mathbb{R}^p, \mathcal{B}^p), \mathbf{Y} : (M, \mathcal{A}) \longrightarrow (\mathbb{R}^q, \mathcal{B}^q)$,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \text{ eine } \mathbb{R}^{p+q} - \text{wertige ZV, mit}$$

$$\boldsymbol{\mu} = E\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{pmatrix} \in \mathbb{R}^{p+q} \text{ bekannt}$$

$$\mathbf{V} = Cov\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{\mathbf{xx}} & \mathbf{V}_{\mathbf{xy}} \\ \mathbf{V}_{\mathbf{yx}} & \mathbf{V}_{\mathbf{yy}} \end{pmatrix} \in PD(p+q) \text{ bekannt .}$$

(b) $\mathbf{X}$ und $\mathbf{Y}$ gemeinsam normalverteilt:

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}_\mathbf{x}, \mathbf{V}_{\mathbf{xx}}) \text{ und } \mathbf{Y} \sim N_q(\boldsymbol{\mu}_\mathbf{y}, \mathbf{V}_{\mathbf{yy}})$$
$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N_{p+q}(\boldsymbol{\mu}, \mathbf{V}), \text{ mit } \boldsymbol{\mu} \in \mathbb{R}^{p+q} \text{ und } \mathbf{V} \in PD(p+q)$$

Wir partitionieren $\boldsymbol{\mu}$ und $\mathbf{V}$ entsprechend der Teil-Dimensionen p und q, wie im Teil (a).
Der bedingte Erwartungswert von $\mathbf{Y}$ unter $\mathbf{X}$ ist gegeben durch:

$$E(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\mu}_\mathbf{y} + \mathbf{V}_{\mathbf{yx}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}(\mathbf{X} - \boldsymbol{\mu}_\mathbf{x}). \tag{1.12}$$

Die bedingte Verteilung von $\mathbf{Y}$ unter $\mathbf{X}$ ist gegeben durch:

$$\mathbb{P}^{\mathbf{Y}|\mathbf{X}=\mathbf{x}} = N(\boldsymbol{\mu}_\mathbf{y} + \mathbf{V}_{\mathbf{yx}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}), \mathbf{V}_{\mathbf{yy}} - \mathbf{V}_{\mathbf{yx}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{V}_{\mathbf{yx}}^{\mathbf{T}}), \ \mathbf{x} \in \mathbb{R}^p \tag{1.13}$$

1.4. Lineares Normalverteilungsmodell

Sei ein lineares Normalverteilungsmodell zu Grunde gelegt

$$\begin{aligned} &\mathbf{X} \ \sim N_n(\mathbf{B}\boldsymbol{\beta}, \sigma^2 \ \mathbf{V}),\ (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^p \times (0, \infty) \\ &\mathbf{B} \text{ eine feste } n \times p - \text{Matrix, } Rang(\mathbf{B}) = p \\ &\mathbf{V} \in PD(n) \text{ bekannt, mit } p \ < \ n. \end{aligned}$$

Theorem 1.5 *Als ML-Schätzung* $(\widehat{\boldsymbol{\beta}}_{ML}(\mathbf{x}), \widehat{\sigma^2}_{ML}(\mathbf{x}))$ *für* $(\boldsymbol{\beta}, \sigma^2)$ *ergibt sich (wobei wir* $n > p$ *voraussetzen):*

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{ML}(\mathbf{x}) &= (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}\mathbf{x}, \\ \widehat{\sigma^2}_{ML}(\mathbf{x}) &= \frac{1}{n}(\mathbf{x} - \mathbf{B}(\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}\mathbf{x})^T(\mathbf{x} - \mathbf{B}(\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}\mathbf{x}). \end{aligned}$$

- Als Gauß-Markov-Schätzung $\widehat{\boldsymbol{\beta}}_{GM}(\mathbf{x})$ für $\boldsymbol{\beta}$ ergibt sich (wobei $\mathbf{V}$ bekannt ist):

$$\widehat{\boldsymbol{\beta}}_{GM}(\mathbf{x}) = (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}\mathbf{x} \tag{1.14}$$
$$\text{BLUE-Schätzung für } \beta. \tag{1.15}$$

- Als gewöhnliche kleinste-Quadrate-Schätzung $\widehat{\boldsymbol{\beta}}_{KQ}(\mathbf{x})$ für $\boldsymbol{\beta}$ ergibt sich:

$$\widehat{\boldsymbol{\beta}}_{KQ}(\mathbf{x}) = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{x}. \tag{1.16}$$

- Es gilt auch :

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{GM} \\ \widehat{\boldsymbol{\beta}}_{KQ} \end{pmatrix} \sim N_{2p}\left(\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{pmatrix} (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1} & (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1} \\ (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1} & (\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{V}\mathbf{B})(\mathbf{B}^T\mathbf{B})^{-1} \end{pmatrix}\right)$$

  Hieraus ergibt sich auch:

$$\begin{aligned} Cov(\widehat{\boldsymbol{\beta}}_{GM}, \widehat{\boldsymbol{\beta}}_{GM} - \widehat{\boldsymbol{\beta}}_{KQ}) &= Cov(\widehat{\boldsymbol{\beta}}_{GM}) - Cov(\widehat{\boldsymbol{\beta}}_{GM}, \widehat{\boldsymbol{\beta}}_{KQ}) \\ &= (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1} - (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1} \\ &= \mathbf{0} \end{aligned}$$

  Damit ist die Behauptung von Theorem(1.3) erfüllt.

Theorem 1.6 *1) Als BLUE-Schätzung $\widehat{\boldsymbol{\beta}}(\mathbf{x})$ für $\boldsymbol{\beta}$ ergibt sich:*

$$\begin{aligned} \widehat{\boldsymbol{\beta}}(\mathbf{x}) &= (\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}\mathbf{x} \\ &= \textit{MCUE-Schätzung für } \boldsymbol{\beta}. \end{aligned}$$

2) Als UMVUE-Schätzung $\widehat{\sigma^2}(\mathbf{x})$ für σ^2 ergibt sich:

$$\widehat{\sigma^2}(\mathbf{x}) = \frac{1}{n-p}(\mathbf{x} - \mathbf{B}(\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}\mathbf{x})^T(\mathbf{x} - \mathbf{B}(\mathbf{B}^T\mathbf{V}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{V}^{-1}\mathbf{x})$$

1.5. Modell mit multivariaten, u.i.v. normalverteilten ZV'en

Seien $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ stochastisch unabhängig und identisch $N_p(\boldsymbol{\beta}, \mathbf{V})$-verteilte $\mathbb{R}^p$-wertige Zufallsvariablen, wobei $(\boldsymbol{\beta}, \mathbf{V}) \in \mathbb{R}^p \times PD(p)$ die Parameter sind. Wir fassen die Zufallsvariablen zu einer $\mathbb{R}^{n\times p}$-wertigen Zufallsvariablen zusammen,

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T.$$

Die Lebesgue-Dichte der Verteilung $P^{\mathbf{X}}_{\boldsymbol{\beta},\mathbf{V}}$ ist gegeben durch:

$$\begin{aligned} f_{\boldsymbol{\beta},\mathbf{V}}(\mathbf{x}) &= \prod_{i=1}^{n}\left[(2\pi)^{-p/2}\det(\mathbf{V})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{x}_i - \boldsymbol{\beta})\right)\right] \\ &= (2\pi)^{-np/2}\det(\mathbf{V})^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{x}_i - \boldsymbol{\beta})\right) \end{aligned}$$

für alle $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$, wobei wir $n > p$ voraussetzen.
Die Log-Likelhood Funktion ist

$$l_{\mathbf{x}}(\boldsymbol{\beta}, \mathbf{V}) = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\det(\mathbf{V})\right) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{x}_i - \boldsymbol{\beta}).$$

Wir betrachten die Statistik

$$\begin{aligned}\overline{\mathbf{x}} &= \frac{1}{n}\mathbf{x}^T\mathbf{1}_n \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \\ \mathbf{S}(\mathbf{x}) &= \sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^T \\ &= \mathbf{x}^T\mathbf{x}-\overline{\mathbf{x}}\,\overline{\mathbf{x}}^T.\end{aligned}$$

Elementare Eigenschaften von S und $\overline{\mathbf{x}}$ Schlittgen (2009, S. 101)

1. Es lässt sich zeigen, dass $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\beta}, \frac{1}{n}\mathbf{V})$.
2. $\mathbf{S}(\mathbf{x}) \in PSD(p),\ \forall \mathbf{x} \in \mathbb{R}^{n\times p}$.
3. $\mathbf{S}(\mathbf{X}) = \mathbf{X}^T(\mathbf{I}_n - \frac{\mathbf{1}}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X} \sim W_p(n-1,\mathbf{V})$
4. $\frac{1}{n-1}\mathbf{S}(\mathbf{X})$ ist ein erwartungstreuer Schätzer für $\mathbf{V}$ ($n \geq 2$ vorausgesetzt).
5. Zusätzlich lässt sich zeigen, dass $\mathbf{S}, \bar{\mathbf{X}}$ unabhängig sind.

Theorem 1.7 *Als ML-Schätung $(\widehat{\boldsymbol{\beta}}_{ML}(\mathbf{x}), \widehat{\mathbf{V}}_{ML}(\mathbf{x}))$ für $(\boldsymbol{\beta}, \mathbf{V})$ ergibt sich (wobei wir $n > p$ voraussetzen):*

$$\widehat{\boldsymbol{\beta}}_{ML}(\mathbf{x}_1,\cdots,\mathbf{x}_n) = \overline{\mathbf{x}}, \tag{1.17}$$

$$\widehat{\mathbf{V}}_{ML}(\mathbf{x}_1,\cdots,\mathbf{x}_n) = \frac{1}{n}S(\mathbf{x}) \tag{1.18}$$

Theorem 1.8 *1) Als BLUE-Schätzung für $\boldsymbol{\beta}$ ergibt sich: $\widehat{\boldsymbol{\beta}}(\mathbf{x}_1,\cdots,\mathbf{x}_n) = \overline{\mathbf{x}}$.*

2) Als MCUE-Schätzung(engl. minimum coriance unbiased estimator MCUE) für $\mathbf{V}$ ergibt sich :

$$\widehat{\mathbf{V}}(\mathbf{x}_1,\cdots,\mathbf{x}_n) = \frac{1}{n-1}S(\mathbf{x}). \tag{1.19}$$

(i) Der mittlere quadratische Fehler für $\widehat{\boldsymbol{\beta}}$ ist gegeben durch:

$$\begin{aligned}MSE(\bar{\mathbf{X}},\boldsymbol{\beta}) &= R(L(\bar{\mathbf{X}},\boldsymbol{\beta})) \\ &= E_\theta[(\bar{\mathbf{X}}-\boldsymbol{\beta})^T(\bar{\mathbf{X}}-\boldsymbol{\beta})] \\ &= Spur(Cov(\bar{\mathbf{X}})) \\ &= \frac{1}{n}Spur(\mathbf{V})\end{aligned}$$

(ii) Der gewichtete quadratische Fehler für $\widehat{\boldsymbol{\beta}}$, mit der gegeben Gewichtsfunktion

$$\mathbf{W} : \mathbb{R}^p \times PD(p) \to PD(p), \mathbf{W}(\boldsymbol{\beta}, \mathbf{V}) = \mathbf{V}^{-1} \tag{1.20}$$

ist gegeben durch:

$$\begin{aligned} MSE(\bar{\mathbf{X}}, \boldsymbol{\beta}) &= R(L(\bar{\mathbf{X}}, \boldsymbol{\beta})) \\ &= E_{\boldsymbol{\theta}}[(\bar{\mathbf{X}} - \boldsymbol{\beta})^T \mathbf{V}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\beta})] \\ &= Spur(\mathbf{V}^{-1} Cov(\bar{\mathbf{X}})) \\ &= \frac{p}{n}. \end{aligned}$$

Modell mit multivariaten gemeinsam, u.i.v. normalverteilten ZV'en

Wir betrachten das Modell mit $n \geq (p+q)$ u.i.v.$\mathbb{R}^{p+q}$-wertigen, normal-verteilten Zufallsvariablen, wobei der ($(p+q)$-dim.) Erwartungswert und die (positiv definite $(p+q) \times (p+q)$) Kovarianzmatrix die Parameter sind:

$$\begin{aligned} &\mathbf{X}_i \sim N_p(\boldsymbol{\beta}_{\mathbf{x}}, \mathbf{V}_{\mathbf{xx}}),\ \mathbf{Y}_i \sim N_q(\boldsymbol{\beta}_{\mathbf{y}}, \mathbf{V}_{\mathbf{yy}}) \\ &\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Y}_i \end{pmatrix} \sim N_{p+q}(\boldsymbol{\beta}, \mathbf{V}),\ (1 \leq i \leq n)\ u.i.v., \\ &\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{\mathbf{x}} \\ \boldsymbol{\beta}_{\mathbf{y}} \end{pmatrix} \in \mathbb{R}^{p+q} \text{ und } \mathbf{V} = \begin{pmatrix} \mathbf{V}_{\mathbf{xx}} & \mathbf{V}_{\mathbf{xy}} \\ \mathbf{V}_{\mathbf{yx}} & \mathbf{V}_{\mathbf{yy}} \end{pmatrix} \in PD(p+q). \end{aligned}$$

Wir haben also effektiv den Parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\beta}_{\mathbf{y}}, \mathbf{V}_{\mathbf{xx}}, \mathbf{V}_{\mathbf{yy}}, \mathbf{V}_{\mathbf{xy}})$ mit fünf Komponenten und den Parameterbereich:

$$\begin{aligned} \Theta = \{&\boldsymbol{\theta} = (\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\beta}_{\mathbf{y}}, \mathbf{V}_{\mathbf{xx}}, \mathbf{V}_{\mathbf{yy}}, \mathbf{V}_{\mathbf{xy}}) : \boldsymbol{\beta}_{\mathbf{x}} \in \mathbb{R}^p, \boldsymbol{\beta}_{\mathbf{y}} \in \mathbb{R}^q, \mathbf{V}_{\mathbf{xx}} \in PD(p), \\ &\mathbf{V}_{\mathbf{yy}} \in PD(q), \mathbf{V}_{\mathbf{xy}} \in \mathbb{R}^{p \times q} \text{ mit } \mathbf{V}_{\mathbf{yy}} - \mathbf{V}_{\mathbf{yx}} \mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}} \mathbf{V}_{\mathbf{yx}}^T \in PD(q)\}. \end{aligned}$$

Wir betrachten die Statistiken:

$$\begin{aligned} &\overline{\mathbf{x}} = \frac{1}{n} \mathbf{x}^T \mathbf{1}_n, \qquad \overline{\mathbf{y}} = \frac{1}{n} \mathbf{y}^T \mathbf{1}_n \qquad \text{mit } \mathbf{x} \in \mathbb{R}^{n \times p}, \mathbf{y} \in \mathbb{R}^{n \times q} \\ &\mathbf{S}_{\mathbf{xx}} = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T, \qquad \mathbf{S}_{\mathbf{yy}} = \sum_{i=1}^n (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^T \\ &\mathbf{S}_{\mathbf{xy}} = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})^T, \qquad \mathbf{S}_{\mathbf{yx}} = \sum_{i=1}^n (\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{x}_i - \overline{\mathbf{x}})^T. \end{aligned}$$

Mit Hilfe von Theorem 1.2 erhält man:

Theorem 1.9 *1) Als BLUE-Schätzung $(\widehat{\boldsymbol{\beta}}_{\mathbf{x}}, \widehat{\boldsymbol{\beta}}_{\mathbf{y}})$ für $(\boldsymbol{\beta}_{\mathbf{x}}, \boldsymbol{\beta}_{\mathbf{y}})$ ergibt sich :*

$$\widehat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{x}) = \overline{\mathbf{x}} \qquad \widehat{\boldsymbol{\beta}}_{\mathbf{y}}(\mathbf{y}) = \overline{\mathbf{y}}.$$

2) Als MCUE-Schätzung $(\widehat{\mathbf{V}}_{\mathbf{xx}}, \widehat{\mathbf{V}}_{\mathbf{yy}}, \widehat{\mathbf{V}}_{\mathbf{xy}})$ für $(\mathbf{V}_{\mathbf{xx}}, \mathbf{V}_{\mathbf{yy}}, \mathbf{V}_{\mathbf{xy}})$ ergibt sich:

$$\begin{aligned} &\widehat{\mathbf{V}}_{\mathbf{xx}}(\mathbf{x}) = \frac{1}{n-1} \mathbf{S}_{\mathbf{xx}},\ \widehat{\mathbf{V}}_{\mathbf{yy}}(\mathbf{y}) = \frac{1}{n-1} \mathbf{S}_{\mathbf{yy}} \\ &\widehat{\mathbf{V}}_{\mathbf{xy}}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \mathbf{S}_{\mathbf{xy}}, \quad \textit{mit } \mathbf{x} \in \mathbb{R}^{n \times p}, \mathbf{y} \in \mathbb{R}^{n \times q}. \end{aligned}$$

1.6. Autoregressives Modell

In der Praxis spielen autoregressive Modelle eine wichtige Rolle, bei denen sich der aktuelle Messwert als Funktion des vorherigen Messwertes und eines Fehlerterms ergibt. Im Folgenden betrachten wir das autoregressive Modell von der Ordnung 1, kurz AR(1). Für Beweise und weitere Ergebnisse auf die Literatur, z.B. Kreiß und Neuhaus (2006, Kap.2, Kap.3)

Definition 1.9 ***(AR(1)-Modell)*** *Die Komponenten der $\mathbb{R}^p$-wertigen Zufallsvariable* $\mathbf{Y}$ *sind autoregressiv AR(1), falls:*

(i) $Y_1 = e_1 \sim N_1(0, \frac{1}{1-\rho^2}\sigma^2) \qquad |\rho| < 1.$

(ii) $Y_i = \rho\, Y_{i-1} + e_i$, *mit* $\quad e_i \sim N_1(0, \sigma^2)$, *für* $\quad i = 2, \ldots, p$ *und* $e_1, \ldots, e_p$ *unabhängig sind.*

Interpretation: Der neue Wert Y_i hängt direkt vom unmittelbaren Vorgänger Y_{i-1} ab,

$$\begin{aligned} Y_i &= e_i + \rho Y_{i-1} \\ &= e_i + \rho(\rho Y_{i-2} + e_{i-1}) \\ &= e_i + \rho e_{i-1} + \rho^2 Y_{i-2} \end{aligned}$$

Eigenschaften des AR(1)-Modells

1. **Erwartungswert:** $E(Y_i) = 0.$
2. **Varianz:** $Var(Y_i) = \frac{1}{1-\rho^2}\sigma^2.$
3. **Autokovarianz:** $Cov(Y_i, Y_{i-1}) = \rho Var(Y_{i-1}) = \rho\frac{\sigma^2}{1-\rho^2}$
   und für $s \leq i-1$ gilt auch:

$$Cov(Y_i, Y_{i-s}) = \rho^s \frac{\sigma^2}{1-\rho^2}.$$

Im Rahmen dieser Arbeit betrachten wir die standardisierte Version der Kovarianzmatrix mit $\sigma^2 = 1 - \rho^2$, z.B. für $p = 4$ erhalten wir:

$$Cov(\mathbf{Y}) = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}.$$

2. Assoziationsstrukturen

In diesem Abschnitt werden Regressionsmodelle und Korrelationskoeffizienten für eindimensionale Zufallsvariable Y und einen Zufallsvektor $\mathbf{X}$ behandelt. Danach kann die Kanonishe Korrelation als Verallgemeinerung der Korrelationskoeffizienten aufgefasst werden. Daran schließt sich der wichtige Spezialfall unter Annahme der Normalverteilung an.

2.1. Theorie der Regressionsmodelle

In dem folgenden Abschnitt wird kurz in grundlegende Prinzipien der Regressionsmodelle eingeführt. Speziell der multiple Korrelationskoeffizient wird dann definiert. Wir verweisen für Beweise und weitere Ergebnisse auf die Literatur, z.B. Puntanen, Styan und Isotalo (2011, Kap.9) und Rao (1973, Kap.4g) zur Regressionsmodelle und Fujikoshi, Ulyanov und Shimizu (2010, Kap.4.2) und Alvin (2008, Kap.10.3) zum multiplen Korrelationskoeffizient.
Modell: In diesem Abschnitt betrachten wir das Modell 1.3.a mit $q = 1$. Kurz:

$$\begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \text{ eine } \mathbb{R}^{p+1} \text{ wertige ZV, mit}$$
$$\boldsymbol{\mu} \in \mathbb{R}^{p+1}, \mathbf{V} \in PD(p+1)$$
$$\boldsymbol{\mu}, \mathbf{V} \text{ bekannt}$$

Notation:

- Bezeichne

$$SP = \{g | g : \mathbb{R}^p \to \mathbb{R},\ g \text{ messbar und } E(g^2(\mathbf{X})) < \infty\}$$

  die Menge aller Funktionen g, so dass $g(\mathbf{X})$ eine quadratintegrierbare reelle Zufallsvariable auf $\mathbb{R}^p$ ist.

- $$LP = \{g | g : \mathbb{R}^p \to \mathbb{R}, g(\mathbf{x}) = b + \mathbf{a}^T\mathbf{x}, \mathbf{a} \in \mathbb{R}^p, b \in \mathbb{R}\} \subseteq SP$$

  die Menge aller linearen reellwertigen Transformationen ist.

Interessieren wir uns für den Zusammenhang zwischen der Zufallsvariable Y und Zufallsvektor $\mathbf{X}$. So gibt sich es verschiedene Ansätze, diesen zu modellieren. Einer der bekanntesten ist das Regressionsmodell.

Definition 2.1 *Eine Funktion $m \in SP$ heißt Regressionsfunktion von Y unter $\mathbf{X}$, wenn gilt*

$$E[(Y - m(\mathbf{X}))^2] = \min_{g \in SP} E[(Y - g(\mathbf{X}))^2].$$

Definition 2.2 *Eine lineare Funktion $f \in LP$ heißt lineare Regressionsfunktion von Y unter $\mathbf{X}$, wenn gilt*

$$E[(Y - f(\mathbf{X}))^2] = \min_{l \in LP} E[(Y - l(\mathbf{X}))^2].$$

Bestimmung der Regressionsfunktion

Aus der Definition des bedingten Erwartungswertes folgt

Satz 2.1 *Rao (1973, Kap.4g.1) Sei $\widehat{Y} = m(\mathbf{X}) = E(Y|\mathbf{X})$, dann ist $E(Y - g(\mathbf{X}))^2$ minimal, wenn $g(\mathbf{X}) = m(\mathbf{X})$ ist.*

Nun fragen wir nach der besten linearen Vorhersage von Y unter $\mathbf{X}$ in der linearen Klasse LP

Satz 2.2 *Rao (1973, Kap.4g.1) Sei $\widehat{Y} = f(\mathbf{X}) = \mu_y + \mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{x}})$, dann ist $E(Y - l(\mathbf{X}))^2$ minimal, wenn $l(\mathbf{X}) = f(\mathbf{X})$ ist*

und es gilt:

Satz 2.3 *In der Klasse der linearen Regressionsmodelle gilt, dass*

$$Var(Y - \mathbf{V}_{y\mathbf{x}}V_{\mathbf{xx}}^{-1}\mathbf{X}) \leq Var(Y - \mathbf{a}^T\mathbf{X}), \quad \forall \mathbf{a} \in \mathbb{R}^p \tag{2.1}$$

ist.

Bemerkung 2.1 $\widehat{\boldsymbol{\beta}} = \mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{V}_{\mathbf{x}y}$ *heißt Regressionskoeffizient von Y unter $\mathbf{X}$.*

Die Abweichungen zwischen den tatsächlichen Werte Y und den geschätzten Werten $\widehat{Y}$:

$$\epsilon = Y - \widehat{Y} \qquad \text{bezeichnet man als Residuen.}$$

Statistische Eigenschaften der Residuen:

Wir beenden diesen Teil mit der Untersuchung der statistischen Eigenschaften der Residuen $\epsilon = Y - E(Y|\mathbf{X})$

1. Erwartungswert: $E(\epsilon) = E(Y - E(Y|\mathbf{X})) = 0$, d.h. die Residuen sind im Mittel Null.

2. Varianz: Es gilt

$$\begin{aligned} Var(\epsilon) &= Var(Y - E(Y|\mathbf{X})) \\ &= Var(Y) + Var(E(Y|\mathbf{X})) - 2\ Cov(Y, E(Y|\mathbf{X})) \\ &= Var(Y) + Var(E(Y|\mathbf{X})) - 2\ Cov(E(Y|\mathbf{X}), E(Y|\mathbf{X})) \\ &= Var(Y) - Var(E(Y|\mathbf{X})). \end{aligned}$$

In der Klasse der linearen Regression erhalten wir:

$$\begin{aligned} Var(\epsilon) &= Var(Y - f(\mathbf{X})) && (2.2) \\ &= Var\big(Y - \mu_y - \mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{x}})\big) && (2.3) \\ &= V_{yy} - \mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{V}_{\mathbf{x}y} && (2.4) \end{aligned}$$

Hieraus erhalten wir:

Satz 2.4 *Die folgenden Bedingungen sind äquivalent*

1. $Var(Y - \mathbf{b}^T\mathbf{X}) \geq Var(Y - \mathbf{a}^T\mathbf{X}),\ \forall \mathbf{b} \in \mathbb{R}^p$
2. $Cov(Y - \mathbf{a}^T\mathbf{X}, \mathbf{c}^T\mathbf{X}) = 0, \forall \mathbf{c} \in \mathbb{R}^p$.
3. $\mathbf{a} = \mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{X}_{\mathbf{x}y}$.

Im Kapitel 3 werden wir die Eigenschaften der Residuen der Vorhersage in verschiedenen Klassen untersuchen.

Multipler Korrelationskoeffizient

Eine wichtige Maßzahl zur Beurteilung der Güte einer Regressionsfunktion ist der multiple Korrelationskoeffizient

Definition 2.3 ***(Multipler Korrelationskoeffizient)***
Der multiple Korrelationskoeffizient $\rho_{Y|\mathbf{X}}$ ist bestimmt als die Korrelation zwischen Y und Regressionskoeffizient von Y unter $\mathbf{X}$ und gegeben durch:

$$\rho_{Y|\mathbf{X}} = \frac{Cov(\boldsymbol{\beta}^T\mathbf{X}, Y)}{\sqrt{Var(\boldsymbol{\beta}^T\mathbf{X})}\sqrt{Var(Y)}}$$

Eigenschaften des multiplen Korrelationskoeffizients

1. Es gilt:
$$Var(\mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{X}) = \mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{V}_{\mathbf{x}y} = Cov(\mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{X}, Y) \quad (2.5)$$
Aus (2.3) und (2.4) ergibt sich die Zerlegung
$$Var(Y) = Var(\mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{X}) + Var(\epsilon)$$

2. Der quadrierte multiple Korrelationskoeffizient stellt sich damit in der Form

$$\rho^2_{Y|\mathbf{X}} = \frac{Var(\mathbf{V}_{y\mathbf{x}}\mathbf{V}^{-1}_{\mathbf{xx}}\mathbf{X})}{Var(Y)} = 1 - \frac{Var(\epsilon)}{Var(Y)}$$

dar

3. Der multiple Korrelationskoeffizient $\rho_{Y|\mathbf{X}}$ ist die maximale Korrelation zwischen Y und einer Linearkombination $\mathbf{a}^T\mathbf{X}$

$$\rho_{Y|\mathbf{X}} = \max_{\mathbf{a}\in\mathbb{R}^p}(\mathbf{a}^T\mathbf{X}, Y) \tag{2.6}$$

4. Aus (2.3) und (2.4) erhalten wir

$$\begin{aligned} Var(\epsilon) &= V_{yy} - \mathbf{V}_{y\mathbf{x}}\mathbf{V}^{-\mathbf{1}}_{\mathbf{xx}}\mathbf{V}_{\mathbf{x}y} && (2.7)\\ &= Var(Y)(1-\rho^2_{Y|\mathbf{X}}), && (2.8)\end{aligned}$$

d.h. je größer $\rho^2_{Y|\mathbf{X}}$, desto kleiner ist $Var(\epsilon)$, das beduetet in diesem Sinn, dass $\rho^2_{Y|\mathbf{X}}$ ein Maß für die Güte der Vorhersage von Y unter $\mathbf{X}$ durch die Regressionsfunktion ist.

Insgesamt erhalten wir somit das folgende Ergebnis, dass der multiple Korrelationskoeffizient die Stärke des Zusammenhangs einer Zufallsvariable Y mit einem Zufallsvektor $\mathbf{X}$ erfasst. In Kapitel 3 werden wir auf die Begriffe Vorhersage und multipler Korrelationskoeffizient zurückkommen.

2.2. Kanonische Korrelation

Im Folgenden geben wir einen kurzen Überblick über die wichtigsten Resultate zum Begriff der kanonischen Korrelation und verweisen für Beweise und weitere Ergebnisse auf die Literatur, z.B. Schlittgen (2009, Kap.12) und Fujikoshi, Ulyanov und Shimizu (2010, Kap.11). Der Ansatz der kanonischen Korrelation bietet eine Möglichkeit Zusammenhänge zwischen zwei Gruppen von Variablen herzuleiten. Hierbei bestimmen wir ein Paar Linearkombinationen der beiden Variablen, so dass diese die größte mögliche Korrelation aufweisen. Dies ist die Aufgabe der sogenannten kanonischen Korrelationsanalyse, die von Hotelling in den 30 er Jahren entwickelt wurde.

Die Ausgangsfragestellung der kanonischen Korrelation

Im letzten Abschnitt wurde der Korrelationskoeffizienten für eine eindimensionale Variable Y und einen $\mathbb{R}^p$-Zufallsvektor $\mathbf{X}$ dargestellt. Die kanonische Korrelation kann in folgendem Sinn als Verallgemeinerung aufgefasst werden: Nun ist auch $\mathbf{Y}$ ein $\mathbb{R}^q$ Zufallsvektor ($q > 1$). Die Idee der kanonischen Korrealtionsanalyse ist die Analyse der Struktur des gemeinsamen Zusammenhangs von zwei Merkmalsgruppen in niedrigerer

Dimension mit Hilfe von Projektionen.
Wir betrachten wieder in diesem Abschnitt das Modell 1.3.a mit $q > 1$, d.h.

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \text{ eine } \mathbb{R}^{p+q}\text{-wertige ZV, mit}$$
$$\boldsymbol{\mu} \in \mathbb{R}^{p+q}, \mathbf{V} \in PD(p+q)$$
$$\boldsymbol{\mu}, \mathbf{V} \text{ bekannt.}$$

Gesucht sind Linearkombinationen $\mathbf{a}^T\mathbf{X}$ und $\mathbf{b}^T\mathbf{Y}$ mit $\mathbf{a} \in \mathbb{R}^p \setminus \{0\}, \mathbf{b} \in \mathbb{R}^q \setminus \{0\}$ derart, dass die Korrelation zwischen diesen Linearkombinationen maximal wird. Dazu betrachtet man

$$Corr(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}) = \frac{Cov(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y})}{\sqrt{Var(\mathbf{a}^T\mathbf{X})}\sqrt{Var(\mathbf{b}^T\mathbf{Y})}} = \frac{\mathbf{a}^T\mathbf{V_{xy}}\mathbf{b}}{\sqrt{\mathbf{a}^T\mathbf{V_{xx}}\mathbf{a}}\sqrt{\mathbf{b}^T\mathbf{V_{yy}}\mathbf{b}}}$$

Vereinbarung:

- die (größte) kanonische Korrelation zwischen $X_1, \ldots, X_p$ und $Y_1, \ldots, Y_q$ ist die maximale Korrelation zwischen allen Linearkombinationen $\mathbf{a}^T\mathbf{X}$ und $\mathbf{b}^T\mathbf{Y}$,

$$Corr_1(\mathbf{X}, \mathbf{Y}) = \max_{\mathbf{a}\in\mathbb{R}^p\setminus\{0\}, \mathbf{b}\in\mathbb{R}^q\setminus\{0\}} Corr(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}).$$

- die Linearkombinationen $\mathbf{U} = \mathbf{a}^T\mathbf{X}$ und $\mathbf{V} = \mathbf{b}^T\mathbf{Y}$, für welche das Maximum angenommen wird, heißen kanonische Variablen.

Das Ziel ist, die Gewichtungsvektoren $\mathbf{a}, \mathbf{b}$ zu bestimmen unter den Normierungsbedingungen $\mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} = 1, \mathbf{b}^T\mathbf{V_{yy}}\mathbf{b} = 1$.

Bestimmung der Gewichtungsvektoren

Kleines Hilfsresultat Seien $\mathbf{A}$ eine $q \times p$-Matrix, $\mathbf{B}$ eine $p \times q-$Matrix. Die Matrizen $\mathbf{AB}$ und $\mathbf{BA}$ haben die gleichen von Null verschiedenen Eigenwerte.
Es lässt sich zeigen: die Gewichtungsvektoren $\mathbf{a}_{cc}, \mathbf{b}_{cc}$ (cc= canonical correlation), sind die Lösung von

$$(\mathbf{V_{xx}^{-1}V_{xy}V_{yy}^{-1}V_{yx}} - r^2\mathbf{I}_p)\mathbf{a} = \mathbf{0}_{p\times p} \quad \text{und} \quad (\mathbf{V_{yy}^{-1}V_{yx}V_{xx}^{-1}V_{xy}} - r^2\mathbf{I}_q)\mathbf{b} = \mathbf{0}_{q\times q},$$

wobei r der größte Eigenwert von

$$\mathbf{V_{xx}^{-1}V_{xy}V_{yy}^{-1}V_{yx}} \text{ bzw. } \mathbf{V_{yy}^{-1}V_{yx}V_{xx}^{-1}V_{xy}} \text{ ist.}$$

Beachte $\mathbf{A} = \mathbf{V_{yy}^{-1}V_{yx}}$, $\mathbf{B} = \mathbf{V_{xx}^{-1}V_{xy}}$. Dann

$$\mathbf{V_{xx}^{-1}V_{xy}V_{yy}^{-1}V_{yx}} \quad , \mathbf{V_{yy}^{-1}V_{yx}V_{xx}^{-1}V_{xy}}$$

besitzen denselben Eigenwert.
Die Gewichtungsvektoren sind die zugehörigen Eigenvektoren, die zu dem größten Eigenwert r gehören.

$$\mathbf{a}_{cc} = e.v_{max}(\mathbf{V_{xx}^{-1}V_{xy}V_{yy}^{-1}V_{yx}}) \quad \text{und} \qquad \mathbf{b}_{cc} = e.v_{max}(\mathbf{V_{yy}^{-1}V_{yx}V_{xx}^{-1}V_{xy}})$$

und es gilt auch:

1) $\mathbf{a}_{cc} = \mathbf{V_{xx}^{-1}V_{xy}b}_{cc}$ und $\mathbf{b}_{cc} = \mathbf{V_{yy}^{-1}V_{yx}a}_{cc}$.

2) Das Minimum $\min\limits_{\mathbf{a,b}} E[(\mathbf{a}^T\mathbf{X} - \mathbf{b}^T\mathbf{Y})^2]$ wird angenommen für

$$\mathbf{a}_{cc} = \mathbf{V_{xx}^{-1}V_{xy}b}_{cc} \quad \text{und} \qquad \mathbf{b}_{cc} = \mathbf{V_{yy}^{-1}V_{yx}a}_{cc}.$$

Andere Zugänge zur Bestimmung der Gewichtungsvektoren $\mathbf{a}, \mathbf{b}$:
Das Optimierungsproblem

$$\max_{\mathbf{a,b}} Corr^2(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y})$$

kann in zwei Schritten erfolgen:

1: wenn $\mathbf{b}$ fest ist, erhalten wir

$$\begin{aligned}
\max_{\mathbf{a}} Corr^2(\mathbf{a}^T\mathbf{X}, \mathbf{b}^T\mathbf{Y}) &= \max_{\mathbf{a}} \frac{(\mathbf{a}^T\mathbf{V_{xy}b})^2}{(\mathbf{a^T V_{xx} a})(\mathbf{b^T V_{yy} b})} \\
&= \frac{1}{\mathbf{b^T V_{yy} b}} \max_{\mathbf{a}} \frac{(\mathbf{a}^T\mathbf{V_{xy}b})^2}{\mathbf{a^T V_{xx} a}} \\
&= \frac{\mathbf{b}^T\mathbf{V_{yx}V_{xx}^{-1}V_{xy}b}}{\mathbf{b^T V_{yy} b}} \quad (\text{ nach Satz A.4}).
\end{aligned}$$

2: danach

$$\max_{\mathbf{b}} \frac{\mathbf{b}^T\mathbf{V_{yx}V_{xx}^{-1}V_{yx}^T b}}{\mathbf{b^T V_{yy} b}} = \lambda_{max}(\mathbf{V_{yy}^{-1}V_{yx}V_{xx}^{-1}V_{yx}^T}) \quad (\text{ nach Satz A.3}).$$

Eigenschaften der kanonischen Korrelation

Die kanonische Korrelation hat folgende Eigenschaften:

1. Ihr Wertebereich ist bestimmt durch

$$0 \leq Corr_1(\mathbf{X}, \mathbf{Y}) \leq 1$$

2. Im Fall $q = 1$ und $p > 1$ ergibt sich der multiple Korrelationskoeffizient zwischen Y und $\mathbf{X} = (X_1, \ldots, X_p)$

3. Für $\mathbf{a} \in \mathbb{R}^p \setminus \{0\}$ und $\mathbf{b} \in \mathbb{R}^q \setminus \{0\}$ gilt :

$$E[(\mathbf{a}^T\mathbf{X} - \mathbf{b}^T\mathbf{Y})^2] \geq 2(1 - Corr_1^2(\mathbf{Y}, \mathbf{X}))$$
$$\text{unter den Nebenbedingungen:} \quad \mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} = 1, \mathbf{b}^T\mathbf{V_{yy}}\mathbf{b} = 1.$$

4. Wenn $\mathbf{Y}^* = \mathbf{F}\mathbf{Y} + \mathbf{f}$ und $\mathbf{X}^* = \mathbf{G}\mathbf{X} + \mathbf{g}$ sind, wobei $\mathbf{G}: \ p \times p$ und $\mathbf{F}: \ q \times q$ reguläre Matrizen und $\mathbf{f} : p \times 1$ und $\mathbf{g} : q \times 1$ feste Vektoren sind, dann ist die kanonische Korrelation zwischen $\mathbf{Y}^*$ und $\mathbf{X}^*$ die gleiche wie zwischen $\mathbf{Y}$ und $\mathbf{X}$, das heißt die kanonische Korrelation ist invariant unter allen linearen Transformationen, die vollen Rang haben.

   Bemerkung 2.2 *Wenn $q = 1$ ist, dann ist der multiple Korrelationskoeffizient zwischen Y und $\mathbf{X}^*$ der gleiche wie zwischen Y und $\mathbf{X}$, das heißt der multiple Korrelationskoeffizient ist invariant unter allen linearen Transformationen von $\mathbf{X}$ die vollen Rang haben.*

2.3. Normalverteilungsmodell

Betrachte des Modells 1.3.b. Kurz:

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \boldsymbol{\mu}_\mathbf{x} \\ \boldsymbol{\mu}_\mathbf{y} \end{pmatrix}, \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{yx}} & \mathbf{V_{yy}} \end{pmatrix}\right),$$
$$\begin{pmatrix} \boldsymbol{\mu}_\mathbf{x} \\ \boldsymbol{\mu}_\mathbf{y} \end{pmatrix} \in \mathbb{R}^{p+q}, \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{yx}} & \mathbf{V_{yy}} \end{pmatrix} \in PD(p+q).$$

Aus den Verteilungsresultaten für (mehrdimensionale) normalverteilte Zufallsvariablen und aus der Modelldefinition des Regressionsmodells ergibt sich

- Die Regressionsfunktion von $\mathbf{Y}$, die auf $\mathbf{X}$ basiert, ist

$$m(\mathbf{X}) = E(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\mu}_\mathbf{y} + \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}(\mathbf{X} - \boldsymbol{\mu}_\mathbf{x})$$

- Die Regressionsfunktion schreibt sich als lineare Funktion um

$$\begin{aligned} E(\mathbf{Y}|\mathbf{X}) &= \boldsymbol{\mu}_\mathbf{y} + \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}(\mathbf{X} - \boldsymbol{\mu}_\mathbf{x}) \\ &= \boldsymbol{\mu}_\mathbf{y} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\boldsymbol{\mu}_\mathbf{x} + \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{X} \\ &= \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1^T\mathbf{X}, \end{aligned}$$

  wobei

$$\begin{aligned} \boldsymbol{\alpha}_0 &= \boldsymbol{\mu}_\mathbf{y} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\boldsymbol{\mu}_\mathbf{x} \\ \boldsymbol{\alpha}_1 &= \mathbf{V_{xx}^{-1}}\mathbf{V_{xy}} \end{aligned}$$

  sind.

- Die Kovarianzmatrix von $(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))$ ist minimal und gegeben durch:

$$\begin{aligned}\boldsymbol{\Sigma} &= Cov(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X})) \\ &= E\left((E(\mathbf{Y}|\mathbf{X}) - \mathbf{Y})(E(\mathbf{Y}|\mathbf{X}) - \mathbf{Y})^T\right) \\ &= \mathbf{V_{yy}} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{V_{xy}}\end{aligned}$$

- Der mittlere quadratische Fehler der Regressionsfunktion von $\mathbf{Y}$, die auf $\mathbf{X}$ basiert, ist gegeben durch den Ausdruck:

$$\begin{aligned}MSE(\mathbf{Y}, E(\mathbf{Y}|\mathbf{X})) &= E\left((\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))^T(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X}))\right) \\ &= Spur\left(Cov\left(\mathbf{Y} - E(\mathbf{Y}|\mathbf{X})\right)\right) \\ &= Spur\left(\mathbf{V_{yy}} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{V_{xy}}\right)\end{aligned}$$

- Verteilung der Residuen $\mathbf{Y} - m(\mathbf{X})$ mit Normalverteilungsannahme:

$$\boldsymbol{\epsilon} \sim N_q(\mathbf{0}, \mathbf{V_{yy}} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{V_{xy}}) \tag{2.9}$$

- Verteilung der Residuenquadratsumme:

$$\boldsymbol{\epsilon}^T\mathbf{D}^{-1}\boldsymbol{\epsilon} \sim \chi^2_q, \text{ mit } \mathbf{D} = \mathbf{V_{yy}} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{V_{xy}} \in PD(q). \tag{2.10}$$

Schätzung von $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1, \boldsymbol{\Sigma}$

Wir betrachten das Modell 1.5 (Modell mit multivariaten gemeinsam, u.i.v. normalverteilten ZV'en): Kurz

$$\begin{pmatrix}\mathbf{X}_i \\ \mathbf{Y}_i\end{pmatrix} \sim N_{p+q}(\boldsymbol{\beta}, \mathbf{V}), \quad (1 \le i \le n) \quad u.i.v.,$$
$$\boldsymbol{\beta} = \begin{pmatrix}\boldsymbol{\beta_x} \\ \boldsymbol{\beta_y}\end{pmatrix} \in \mathbb{R}^{p+q} \text{ und } \mathbf{V} = \begin{pmatrix}\mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{yx}} & \mathbf{V_{yy}}\end{pmatrix} \in PD(p+q).$$

Hieraus und aus Theorem 1.10 ergibt sich als MCUE-Schätzung(engl. minimum covariance unbiased MCUE) für $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1$ und $\boldsymbol{\Sigma}$ nach Rao, Shalabh, Toutenburg und Heumann (2008, Theorem 3.16 S.74-75)

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_0(\mathbf{x}, \mathbf{y}) &= \overline{\mathbf{y}} - \mathbf{S_{yx}}\mathbf{S_{xx}^{-1}}\overline{\mathbf{x}} \\ \widehat{\boldsymbol{\alpha}}_1(\mathbf{x}, \mathbf{y}) &= \mathbf{S_{xx}^{-1}}\mathbf{S_{xy}} \\ \widehat{\boldsymbol{\Sigma}}(\mathbf{x}, \mathbf{y}) &= \frac{1}{n-1}(\mathbf{S_{yy}} - \mathbf{S_{yx}}\mathbf{S_{xx}^{-1}}\mathbf{S_{xy}}) \\ \widehat{MSE}(\mathbf{Y}, E(\mathbf{Y}|\mathbf{X})) &= Spur[\frac{1}{n-1}(\mathbf{S_{yy}} - \mathbf{S_{yx}}\mathbf{S_{xx}^{-1}}\mathbf{S_{xy}})],\end{aligned}$$

mit $\mathbf{x} \in \mathbb{R}^{n\times p}, \mathbf{y} \in \mathbb{R}^{n\times q}$.

3. Einige Aspekte der Theorie der Vorhersage

Eine der wichtigsten Anwendungen in der Statistik ist die Prognose. Das folgende Kapitel beschäftigt sich mit der Optimalität von Vorhersagen. Hierfür wird der klassische Zugang der Effizienz, welche am mittleren quadratischen Fehler von dem zur schätzenden Vorhersage gemessen wird, betrachtet. Es stellt sich heraus, das zusätzlich zu einem Abstandskriterium eine zweite Bedingung, die Unverzerrtheit, gefordert werden muss, um hinreichend allgemeine Aussagen treffen zu können. Anschließend werden als Erweiterung die Eigenschaften des mittleren quadratischen Fehlers behandelt, welche auch für die Darstellung der Methode der partiellen kleinsten Quadrate anwendbar sind.

3.1. Allgemeine Resultate

Das Ziel dieses Abschnitts wird sein, ein Kriterium zu finden, welches bereits zur Beurteilung einer Vorhersage genutzt werden kann, um die Qualität einer Vorhersage zu messen.
Modell: Seien $(\Omega, \mathcal{C}, \mathbb{P})$ ein W-Raum und

$$X : (\Omega, \mathcal{C}) \to (M_1, \mathcal{A}_1)$$
$$Y : (\Omega, \mathcal{C}) \to (M_0, \mathcal{A}_0)$$

zwei Zufallsvariablen gegeben, wobei

- $(M_i, \mathcal{A}_i)$, $i = 0, 1$ Messräume sind.
- Y nicht beobachtbar ist.
- Die Verteilung $P^{(Y,X)}$ wird als gemeinsame Verteilung der Zufallsvariablen (Y, X) bezeichnet, und ist ein W-Verteilung auf der Produkt Sigma-Algebra $\mathcal{A}_0 * \mathcal{A}_1$.

In der Statistik ist die Frage von Interesse: wie lässt sich Y durch X approximativ beschreiben? oder anders gesagt, im Fall, dass Y nicht beobachtbar ist, welche Prognosen lassen sich über Y machen, wenn X beobachtbar ist?

Definition 3.1 *(**Vorhersage**) Rao, Shalabh, Toutenburg und Heumann (2008, Kap.6) Unter einer Vorhersage $\widehat{Y} = g(X)$ für Y verstehen wir eine Funktion $g : M_1 \to M_0$, die*

für jeden möglichen Wert $x \in M_1$ der Zufallsvariable X einen Vorhersagewert $g(x)$ von Y liefert, dabei soll g messbar sein:

$$g : (M_1, \mathcal{A}_1) \to (M_0, \mathcal{A}_0).$$

Bezeichne

$$SP = \{g | g : M_1 \longrightarrow M_0 \text{ messbar}, E\big(g^2(X)\big) < \infty)\} \tag{3.1}$$

die Menge aller quadratintegrierbaren Vorhersagen von Y unter X.
Mit der Bezeichnung $\widehat{Y}$ statt g für eine Vorhersage für Y haben wir einen kleinen Notationswechsel vorgenommen, der natürlich inhaltlich Bedeutung hat. In diesem Kapitel betrachten wir verschiedene Vorhersagen für die selbe unbeobachtbare Zufallsvariable Y, und bezeichnen diese mit $g, g_1, \widehat{Y}, Y^*$ etc. für verschiedene Vorhersagen für Y. Das Konzept zur Beurteilung bzw. zum Vergleich verschiedener Vorhersagen für Y geht von einer gegebenen quadratischen Verlustfunktion L aus (steht für engl. Loss):

$$L : M_0 \times M_0 \to [0, \infty) \text{ messbar .}$$

Bemerkung 3.1 *Im Rahmen dieser Arbeit betrachten wir die Verlustfunktionen*

$$L : M_0 \times M_0 \to [0, \infty),$$

die die folgenden Bedingungen erfüllen:

1. $L \geq 0$
2. *die Funktion L ist konvex.*
3. $L(y_1, y_1) = 0$.

Speziell wird als Verlustfunktion verwendet $L(y_1, y_2) = \phi(y_1 - y_2)$, mit einer konvexen und Nullpunkt-symmetrischen Funktion ϕ.

Einige Verlustfunktionen:

(i) Im Fall $M_1 = \mathbb{R}^p, M_0 = \mathbb{R}$:

$$L(y_1, y_2) = (y_1 - y_2)^2 \quad \text{(quadratische Abweichung)}$$

$$L(y_1, y_2) = |y_1 - y_2|, \quad \text{(absolute Abweichung)}$$

(ii) Im Fall $M_1 = \mathbb{R}^p, M_0 = \mathbb{R}^q$:

$$L(\mathbf{y}_1, \mathbf{y}_2) = (\mathbf{y}_1 - \mathbf{y}_2)^T (\mathbf{y}_1 - \mathbf{y}_2) \quad \text{(quadratische Abweichung)}$$

$$L_{\mathbf{W}}(\mathbf{y}_1, \mathbf{y}_2) = (\mathbf{y}_1 - \mathbf{y}_2)^T \mathbf{W} (\mathbf{y}_1, \mathbf{y}_2), \mathbf{W} \in PD(q) \text{(gewichtete quadratische Abweichung).}$$

Für das theoretische Verständnis der Vorhersage bzw. der partiellen kleinsten Quadrate spielt der mittlere quadratische Fehler der Vorhersage eine wichtige Rolle. Dazu führen wir folgende Definition ein.

Definition 3.2 ***(Risikofunktion einer Vorhersage)*** *Rao, Shalabh, Toutenburg und Heumann (2008, Kap.6.4):*
Für eine Vorhersage $\widehat{Y} = g(X) : (M_1, \mathcal{A}_1) \to (M_0, \mathcal{A}_0)$ *heißt*

$$R(Y, \widehat{Y}(X) = E(L(Y, \widehat{Y}(X))) \tag{3.2}$$

Risikofunktion der Vorhersage $\widehat{Y}$*.*

Definition 3.3 ***(mittlerer quadratischer Fehler der Vorhersage MSEP)***

(i) *Für die quadratische Verlustfunktion* $L(Y, \widehat{Y}) = (Y - \widehat{Y})^2$ *heißt die Risikofunktion*

$$R(Y, \widehat{Y}) = E((Y - \widehat{Y})^2), \tag{3.3}$$

der mittlere quadratische Fehler der Vorhersage: $MSEP(Y, \widehat{Y})$*, (engl. Mean Squared Error of Prediction)*

(ii) *Für die gewichtete quadratische Verlustfunktion*

$$L_{\mathbf{W}}(\mathbf{Y}, \widehat{\mathbf{Y}}) = (\mathbf{Y} - \widehat{\mathbf{Y}})^T \mathbf{W} (\mathbf{Y} - \widehat{\mathbf{Y}}), \ \mathbf{W} \in PD(p) \qquad \textit{ist:}$$

$$R_{\mathbf{W}}(\mathbf{Y}, \widehat{\mathbf{Y}}) = E((\mathbf{Y} - \widehat{\mathbf{Y}})^T \mathbf{W} (\mathbf{Y} - \widehat{\mathbf{Y}})) \quad \textit{(Weighted Mean Squared Error).} \tag{3.4}$$

der gewichtete mittlere quadratische Fehler der Vorhersage (engl. Weighted Mean Squared Error)

Der mittlere quadratische Fehler der Vorhersage (MSEP) kann als Vergleichskriterium für verschiedene Vorhersagen herangezogen werden. Vorhersagen mit kleinem MSEP sind dabei vorzuziehen.

Folgerung 3.1 *Für den mittleren quadratischen Fehler der Vorhersage gilt:*

$$\begin{aligned}
MSEP(Y, \widehat{Y}) &= E[(Y - \widehat{Y})^2] \\
&= E[(Y - \widehat{Y} - E(Y - \widehat{Y}) + E(Y - \widehat{Y}))^2] \\
&= E[(Y - \widehat{Y} - E(Y - \widehat{Y}))^2] + (E(Y - \widehat{Y}))^2 \\
&\quad +2\ E[E(Y - \widehat{Y})(Y - \widehat{Y})] - 2\ E(Y - \widehat{Y})E(Y - \widehat{Y}) \\
&= E(Y - \widehat{Y} - E(Y - \widehat{Y}))^2 + (E(Y - \widehat{Y}))^2 \\
&= Var(Y - \widehat{Y}) + [E(Y) - E(\widehat{Y})]^2.
\end{aligned}$$

Daraus erhalten wir folgende wichtige Zerlegung des mittleren quadratischen Fehlers der Vorhersage in die Varianz der Vorhersage und das Quadrat der Verzerrung:

$$\begin{aligned} R(Y,\widehat{Y}) &= MSEP(Y,\widehat{Y}) \\ &= Var(Y-\widehat{Y}) + [E(Y) - E(\widehat{Y})]^2. \end{aligned}$$

Im mehrdimensionalen Fall formulieren wir den mittleren quadratischen Fehler der Vorhersage wie folgt:

$$\begin{aligned} R(\mathbf{Y},\widehat{\mathbf{Y}}) &= MSEP(\mathbf{Y},\widehat{\mathbf{Y}}) \\ &= E((\mathbf{Y}-\widehat{\mathbf{Y}})^T(\mathbf{Y}-\widehat{\mathbf{Y}})) \\ &= Spur\ [Cov(\mathbf{Y}-\widehat{\mathbf{Y}})] + [E(\mathbf{Y}-\widehat{\mathbf{Y}})]^T[E(\mathbf{Y}-\widehat{\mathbf{Y}})]. \end{aligned}$$

Definition 3.4 ***(optimale Vorhersage)*** *Rao, Shalabh, Toutenburg und Heumann (2008, Kap.6.4)*
Die Vorhersage $g^* \in G$ heißt optimale Vorhersage in der Klasse $G \subseteq SP$, wenn für jede Vorhersage $g \in G$ gilt:

$$\begin{aligned} &R(Y,g^*(X)) \leq R(Y,g(X)) \ \textit{bzw.} \\ &R_{\mathbf{W}}(\mathbf{Y},g^*(\mathbf{X})) \leq R_{\mathbf{W}}(\mathbf{Y},g(\mathbf{X})), \quad \mathbf{W} \in PD(p). \end{aligned}$$

Definition 3.5 ***(erwartungstreue Vorhersage)***
die Vorhersage $g(X)$ heißt (engl. unbiased) erwartungstreue (unverzerrte) Vorhersage, wenn $E(g(X)) = E(Y)$ ist.

3.2. Optimalitätskonzepte

Das Ziel dieses Abschnittes besteht darin, die beste Vorhersage bzw. beste lineare Vorhersage zu bestimmen. Zu Beginn werden die folgenden Lemmas vorgestellt, welche die nachfolgenden Begriffe motivieren.

Lemma 3.1 ***(Minimaleigenschaft des Mittelwerts)***
Sei X quadratintegrierbare reelle Zufallsvariable auf M_1 mit $Var(X) < \infty$ und $E(X) = \mu$. Dann gilt

$$E[(X-a^*)^2] = \min_{a\in\mathbb{R}} E[(X-a)^2] \quad \Leftrightarrow \quad a^* = \mu,$$

die Varianz ist ein Maß dafür, wie weit die Werte von X im Schnitt auseinander fallen

Lemma 3.2 *Sei X quadratintegrierbare reelle Zufallsvariable auf M_1 mit $Var(X) < \infty$ und $E(X) = \mu$. Dann gilt*

$$\begin{bmatrix} \textit{für } a \in \mathbb{R} \textit{ und} \\ \\ E[(X-a)^2] = E(X^2) - a^2 \end{bmatrix} \Leftrightarrow \begin{bmatrix} a = 0, \\ \textit{oder} \\ a = \mu \end{bmatrix}.$$

Beweis:
" $\Leftarrow$ " wenn $a = \mu$, folgt $E(X-\mu)^2 = E(X^2) - \mu^2$.
" $\Rightarrow$ " Sei $E[(X-a)^2] = E(X^2) - a^2$,

$$E[(\mu - a)^2] \;\; = \;\; E(X^2) + a^2 - 2a\mu,$$

Hieraus und aus $E[(X-a)^2] = E(X^2) - a^2$ ergibt sich:

$$a(a-\mu) = 0 \Longrightarrow a = \mu \quad \text{ oder } a = 0$$

und die Varianz ist gerade die mittlere quadratische Abweichung vom Mittelwert. Um Optimalitätsaussagen machen zu können, braucht man die folgenden Begriffe:

Definition 3.6 ***(Beste Vorhersage)*** *Rao (1973, Kap. 4g.1)*
Die Vorhersage $\widehat{Y} \in G$ heißt beste Vorhersage von Y unter $\mathbf{X}$ in der Klasse $G \subseteq SP$ im Sinne des mittleren quadratischen Fehlers, wenn gilt:

$$E(Y - \widehat{Y}(X))^2 = \min_{g \in G} E(Y - g(\mathbf{X}))^2.$$

Einige Klassen von G:

1. $SP = \{g | g : \mathbb{R}^p \to \mathbb{R}, g \text{ messbar } E(g^2(\mathbf{X})) < \infty\}$, die Menge aller quadratintegrierbaren reellen Zufallsvariablen

2. $UP = \{g | g : \mathbb{R}^p \to \mathbb{R}, E(Y) - E(g(\mathbf{X})) = 0, E(g^2(\mathbf{X})) < \infty\} \subseteq SP$, die Teilmenge der unverzerrten quadratintegrierbaren reellen Zufallsvariablen

3. $UP_0 = \{g | g : \mathbb{R}^p \to \mathbb{R}, E(g(\mathbf{X})) = 0, E(g^2(\mathbf{X})) < \infty\} \subseteq SP$, die Menge der quadratintegrierbaren reellen Zufallsvariablen mit Erwartungswert null.

4. $G_0 = \{g | g : \mathbb{R}^p \to \mathbb{R}, g(\mathbf{x}) = b, b \in \mathbb{R}\} \subseteq SP$, die Menge aller konstanten Abbildungen

5. $LP = \{g | g : \mathbb{R}^p \to \mathbb{R}, g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b, \mathbf{a} \in \mathbb{R}^p, b \in \mathbb{R}\} \subseteq SP$, die Menge aller linearen reellwertigen Transformationen.

6. $LUP = \{g | g : \mathbb{R}^p \to \mathbb{R}, g(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + b, E(Y) - E(g(\mathbf{X})) = 0, \ \mathbf{a} \in \mathbb{R}^p, b \in \mathbb{R}\} \subseteq LP$ die Menge aller unverzerrten linearen reellwertigen Transformationen

7. $LP_2 = \{g | g : \mathbb{R}^p \to \mathbb{R}, g(\mathbf{x}) = \mu_y + \mathbf{a}^T(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}), \ \mathbf{a} \in \mathbb{R}^p, \text{ mit } \mathbf{a}^T\mathbf{V}_{\mathbf{xx}}\mathbf{a} = 1\} \subseteq LP$ die Menge aller unverzerrten, über $Cov(\mathbf{X})$ standardisierten, linearen, reellwertigen Transformationen.

8. $LP_1 = \{g | g : \mathbb{R}^p \to \mathbb{R}, g(\mathbf{x}) = \mu_y + \mathbf{a}^T(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}), \ \mathbf{a} \in \mathbb{R}^p \text{ mit } \mathbf{a}^T\mathbf{a} = 1\} \subseteq LP$, die Menge aller unverzerrten, standardisierten, linearen, reellwertigen Transformationen.

Beachte: Folgende Aspekte sollte man beachten:

1) $LUP = LP \cap UP$

2) $G \cap UP$ die Teilmenge von G der unverzerrten quadratintegrierbaren reellen Zufallsvariablen

3) $G \cap UP_0$ die Teilmenge der quadratintegrierbaren reellen Zufallsvariablen mit Erwartungswert null.

Bezeichnungen: Eine gewisse Abschwächung der Behauptungen von Definition (3.6) werden wir im Folgenden erklären:

(1) Eine Vorhersage $\widehat{Y} \in SP$ mit $E[(Y - \widehat{Y})^2] = \inf\{E[(Y - g(\mathbf{X}))^2] : g \in SP\}$, heißt beste Vorhersage.

(2) Eine Vorhersage $\widehat{Y} \in UP$ mit $Var(Y - \widehat{Y}) = \inf\{Var(Y - g(\mathbf{X})) : g \in UP\}$, heißt beste erwartungstreue (unverzerrte) Vorhersage.

(3) Eine Vorhersage $\widehat{Y} \in G \cap UP$ mit $Var(Y - \widehat{Y}) = \inf\{Var(Y - g(\mathbf{X})) : g \in G \cap UP\}$, heißt beste erwartungstreue (unverzerrte) Vorhersage in der Klasse G.

(4) Eine Vorhersage $\widehat{Y} \in LUP$ mit $Var(Y - \widehat{Y}) = \inf\{Var(Y - g(\mathbf{X})) : g \in LUP\}$, heißt beste lineare erwartungstreue (unverzerrte) Vorhersage.

(5) Seien G_1, G_2 zwei Klassen von G, wenn $G_1 \subset G_2$ ist, gilt:

$$\min_{h \in G_2} E(Y - h(\mathbf{X}))^2 \leq \min_{g \in G_1} E(Y - g(\mathbf{X}))^2,$$
$$\max_{h \in G_2} Corr^2(Y, h(\mathbf{X}))^2 \geq \max_{g \in G_1} Corr^2(Y, g(\mathbf{X}))^2$$

3.3. Die beste erwartungstreue Vorhersage

Betrachtet man nur erwartungstreue (unverzerrte) Vorhersagen, so kann man die Varianz der Vorhersage als Maß für die Qualität der Vorhersage heranziehen, da unter Erwartungstreue die Varianz der Vorhersage gleich dem mittleren quadratischen Fehler ist. Eine Vorhersage ist in diesem Sinn besser als alle anderen unverzerrten Vorhersagen, falls seine Varianz minimal auf G ist. Für eine erwartungstreue (unverzerrte) Vorhersage $\widehat{Y} \in G \cap UP$ von Y gilt die folgende Optimalitätseigenschaft:

$$E[(Y - \widehat{Y}(X))^2]) = \min_{g \in G} E[(Y - g(\mathbf{X}))^2], \text{ d.h.} \tag{3.5}$$
$$Var(Y - \widehat{Y}(X)) = \min_{g \in G \cap UP} Var(Y - g(\mathbf{X})). \tag{3.6}$$

BUP steht für beste erwartungstreue Vorhersage (engl. Best Unbiased Prediction).

Die Eigenschaften der besten erwartungstreuen Vorhersage

Die Optimalität von erwartungstreuen Vorhersagen lässt sich mit der Kovarianzmethode charakterisieren.

Lemma 3.3 ***(Kovarianzmethode)***
Sei $\widehat{Y}$ eine erwartungstreue Vorhersage für Y in der Klasse G ($\widehat{Y} \in G \cap UP$), dann gilt:

$$\widehat{Y} \text{ ist BUP in } G \iff Cov(Y - \widehat{Y}(\mathbf{X}), g(\mathbf{X})) = 0, \forall g \in G \cap UP_0.$$

Auf Grund der Modellannahme gilt folgende erweiterte Version der Kovarianzmethode

Theorem 3.1 ***(Kovarianzmethode- Version 2)***
Sei $\widehat{Y}$ eine erwartungstreue Vorhersage für Y in der Klasse G ($\widehat{Y} \in G \cap UP$), dann gilt:

$$\widehat{Y} \text{ ist BUP in } G \iff Cov(Y - \widehat{Y}(\mathbf{X}), g(\mathbf{X})) = 0, \forall g \in G.$$

Beweis: Notwendig: " $\Rightarrow$ "
Seien $\widehat{Y}$ eine BUP für Y in der Klasse G und $g \in G$, dann gilt $\forall \lambda \in \mathbb{R}$

$$\begin{aligned} MSEP\big(Y, \widehat{Y} - \lambda g(\mathbf{X})\big) &= Var\big(Y - \widehat{Y} + \lambda g(\mathbf{X})\big) + \lambda^2 E^2(g(\mathbf{X})) \\ &= Var(Y - \widehat{Y}) + \lambda^2\ Var(g(\mathbf{X})) \\ &\quad +2\lambda Cov(Y - \widehat{Y}, g(\mathbf{X})) + \lambda^2\ E^2(g(\mathbf{X})) \\ &\quad \geq\ Var(Y - \widehat{Y}) \quad (\text{ da } \widehat{Y} \text{ BUP ist}) \end{aligned}$$

$$\begin{aligned} &\Rightarrow\ \lambda^2 Var(g(\mathbf{X})) + 2\ \lambda\ Cov(Y - \widehat{Y}, g(\mathbf{X})) + \lambda^2 E(g^2(\mathbf{X})) \geq 0 \\ &\Rightarrow\ \lambda^2(Var(g(\mathbf{X})) + E^2(g(\mathbf{X}))) + 2\ \ \lambda Cov(Y - \widehat{Y}, g(\mathbf{X})) \geq 0 \\ &\Rightarrow\ \lambda^2 E(g^2(\mathbf{X})) + 2\ \lambda Cov(Y - \widehat{Y}, g(\mathbf{X})) \geq 0 \\ &\Rightarrow\ Cov(Y - \widehat{Y}, g(\mathbf{X})) = 0, \text{ da} \end{aligned}$$

wenn $Cov(Y - \widehat{Y}, g(\mathbf{X})) < 0$ ist, $\exists \lambda \in [0, -2\frac{Cov(Y-\widehat{Y}, g(\mathbf{X}))}{E(g^2(\mathbf{X}))}]$, so dass

$$\lambda^2 E(g^2(\mathbf{X})) + 2\ \lambda Cov(Y - \widehat{Y}, g(\mathbf{X})) < 0.$$

ist, und wenn $Cov(Y - \widehat{Y}, g(\mathbf{X})) > 0$ ist, $\exists \lambda \in [-2\frac{Cov(Y-\widehat{Y}, g(\mathbf{X}))}{E(g^2(\mathbf{X}))}, 0]$, so dass

$$\lambda^2 E(g^2(\mathbf{X})) + 2\ \lambda Cov(Y - \widehat{Y}, g(\mathbf{X})) < 0.$$

ist.
Hieraus folgt, dass $\widehat{Y}$ nicht BUP ist, wenn $Cov(Y - \widehat{Y}, g(\mathbf{X})) \neq 0$ ist, im Widerspruch zur Annahme, dass $\widehat{Y}$ BUP für Y in der Klasse G.

Hinreichend: " $\Leftarrow$ "
Sei nun umgekehrt $Cov\big(Y - \widehat{Y}(\mathbf{X}), g(\mathbf{X})\big) = 0, \forall g \in G$, dann für $\widetilde{Y} \in G \cap UP$ gilt:

$$Cov(Y - \widehat{Y}, \widehat{Y}) = 0 \qquad Cov(Y - \widehat{Y}, \widetilde{Y}) = 0$$

Hieraus ergibt sich, dass

$$Cov(Y - \widehat{Y}, \widehat{Y} - \widetilde{Y}) = 0$$

und

$$\begin{aligned} Var(Y - \widetilde{Y}) &= Var(Y - \widehat{Y} + \widehat{Y} - \widetilde{Y}) \\ &= Var(Y - \widehat{Y}) + Var(\widehat{Y} - \widetilde{Y}) \\ &\quad +2\ Cov(Y - \widehat{Y}, \widehat{Y} - \widetilde{Y}) \\ &= Var(Y - \widehat{Y}) + Var(\widehat{Y} - \widetilde{Y}). \end{aligned}$$

Dieses impliziert, dass

$$Var(Y - \widehat{Y}) \leq Var(Y - \widetilde{Y}).$$

Also ist $\widehat{Y}$ beste lineare erwartungstreue Vorhersage (BUP) und damit folgt die Behauptung.

Beispiel 3.1

(a) *Für $\widehat{Y} \in UP$ gilt das Äquivalent:*

$$Cov(Y - \widehat{Y}, g(\mathbf{X})) = 0, \forall g \in SP \quad \Longleftrightarrow \quad \widehat{Y} = m(\mathbf{X}),$$

wobei $m(\mathbf{X}) = E(Y|X)$ ist.
" $\Leftarrow$ " *Wenn $\widehat{Y} = m(\mathbf{X})$ ist, erhalten wir aus den Eigenschaften der bedingten Erwartung,*

$$\begin{aligned} Cov(Y - m(\mathbf{X}), g(\mathbf{X})) &= Cov(Y, g(\mathbf{X})) - Cov(m(\mathbf{X}), g(\mathbf{X})), \forall g \in SP \\ &= Cov(m(\mathbf{X}), g(\mathbf{X})) - Cov(m(\mathbf{X}), g(\mathbf{X})) \\ &= 0, \end{aligned}$$

da

$$\begin{aligned} Cov(Y, g(\mathbf{X})) &= E\big[\big(Y - E(Y)\big)\big(g(\mathbf{X}) - E(g(\mathbf{X}))\big)\big] \\ &= E\Big[\big(g(\mathbf{X}) - E(g(\mathbf{X}))\big)E\big[\big(Y - E(Y)\big)\big]|\mathbf{X}\Big] \\ &= Cov(m(\mathbf{X}), g(\mathbf{X})) \end{aligned}$$

" $\Rightarrow$ " *Sei nun umgekehrt $\widehat{Y} \in UP$ mit $Cov(Y - \widehat{Y}, g(\mathbf{X})) = 0, \forall g \in SP$, dann ist $\widehat{Y} = m(\mathbf{X})$. Also genügt zu zeigen, dass*

$$E(m(\mathbf{X}) - \widehat{Y}) = 0 \qquad \textit{und} \qquad Var(m(\mathbf{X}) - \widehat{Y})) = 0$$

ist.

$$\begin{aligned} Var(m(\mathbf{X}) - \widehat{Y}) &= Cov(m(\mathbf{X}) - \widehat{Y}, m(\mathbf{X}) - \widehat{Y}) \\ &= Cov(Y - \widehat{Y} - (Y - m(\mathbf{X})), m(\mathbf{X}) - \widehat{Y}) \\ &= Cov(Y - \widehat{Y}, m(\mathbf{X}) - \widehat{Y}) - Cov(Y - m(\mathbf{X}), m(\mathbf{X}) - \widehat{Y}) \\ &= 0 - 0 \end{aligned}$$

Dieses impliziert, dass $\widehat{Y} = m(\mathbf{X})$ *und das Theorem(3.1) in der SP erfült ist.*

(b) *Auf analoge Weise gilt für* $\widehat{Y} \in LP \cap UP$*, dass:*

$$Cov(Y - \widehat{Y}, g(\mathbf{X})) = 0,\ \forall g \in LP \setminus G_0 \Longleftrightarrow \widehat{Y} = \mu_y + \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}(\mathbf{X} - \boldsymbol{\mu}_\mathbf{x}).$$

Zum Beispiel:

1) $Cov(Y - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{X}, \mathbf{X}) = \mathbf{0}$

2) $Cov(Y - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{X}, \mathbf{V_{xx}^{-1}}\mathbf{X}) = \mathbf{0}$

3) $Cov(Y - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{X}, X_1) = 0, \qquad X_1 = \mathbf{e_1}^T\mathbf{X}.$

(c) *Im Fall, dass* Y *und* $\mathbf{X}$ *unabhängig sind, gilt:*

$$Cov(Y - \widehat{Y}, g(\mathbf{X})) = 0,\ \forall g \in LP \setminus G_0 \Longleftrightarrow \widehat{Y} = \mu_y.$$

Beispiel 3.2 ***AR(1)-Modell***
Wir betrachten das Modell 1.6 (Normalverteilung mit autoregressiver Struktur AR(1))

$$\begin{pmatrix} X_1 \\ X_2 \\ Y \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \mu_y \end{pmatrix}, \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \right) \quad mit \ \ |\rho| < 1.$$

Seien

$$\begin{aligned} G_1 &= \{g | g : \mathbb{R}^2 \to \mathbb{R}, g(x_1, x_2) = a_1 x_1 : \quad a_1 \in \mathbb{R}\} \\ G_2 &= \{g | g : \mathbb{R}^2 \to \mathbb{R}, g(x_1, x_2) = c_1 x_1 + c_2 x_2 : \quad c_1, c_2 \in \mathbb{R}\} \end{aligned}$$

Die beste lineare erwartungstreue Vorhersage für Y *unter* $\mathbf{X}$ *in den Klassen* G_1, G_2 *und der zugehörige mittlere quadratische Fehler sind:*

G	$BLUP$	$MSEP$
G_1	$\rho^2 X_1$	$1 - \rho^4$
G_2	ρX_2	$1 - \rho^2$

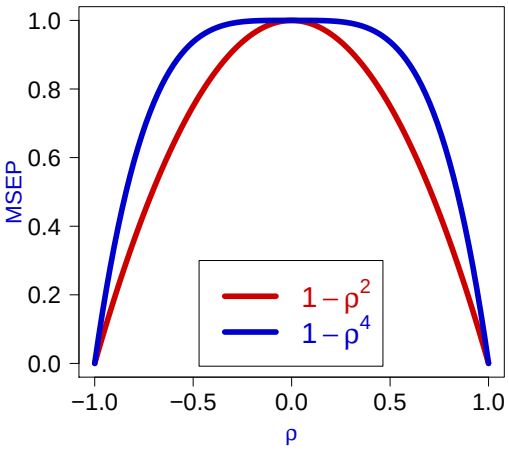

Abbildung 3.1.: Vergleich der MSEP

Die Abbildung (3.1) zeigt den mittleren quadratischen Fehler der Vorhersagen MSEP in den Klassen G_1 und G_2 in Abhängigkeit von ρ. Es gilt auch:

$$\begin{aligned} Cov(Y - \rho^2 X_1, aX_1) &= Cov(Y, aX_1) - Cov(\rho^2 X_1, aX_1), \forall a \in \mathbb{R} \\ &= \rho^2 a - \rho^2 a \\ &= 0 \end{aligned}$$

$$\begin{aligned} Cov(Y - \rho X_2, c_1 X_1 + c_2 X_2) &= Cov(Y, c_1 X_1 + c_2 X_2) - Cov(\rho X_2, c_1 X_1 + c_2 X_2), \forall c_1, c_2 \in \mathbb{R} \\ &= \rho^2 c_1 + \rho c_2 - \rho^2 c_1 - \rho c_2 \\ &= 0. \end{aligned}$$

Weiterhin kann man bemerken:

$$\begin{aligned} Cov(Y - \rho^2 X_1, X_2) &= Cov(Y, X_2) - Cov(\rho^2 X_1, X_2) \\ &= \rho^2 - \rho^3 \end{aligned}$$

Beachte:

- Der Vorhersagefehler ist somit mit der für die Vorhersage zur Verfügung stehenden Information unkorreliert. In diesem Sinn kann die Vorhersage nicht mehr verbessert werden. Der Prognosefehler $Y - \widehat{Y}$ steht somit orthogonal zu allen in die Prognosefunktion eingehenden Variablen $X_1, \ldots, X_p$.
- Geometrisch ausgedrückt besagt diese Eigenschaft, dass man die beste lineare Vorhersage eines Punktes, hier y, in einem Teilraum, hier der durch $x_1, \ldots, x_p$ aufgespannten lineare Teilraum, als orthogonale Projektion des Punktes auf diesen Teilraum erhält.

- da $LP \subseteq SP$ ist, ist die lineare Vorhersage nicht immer optimal. Als Beispiel dafür ist:

  Beispiel 3.3 *Kreiß und Neuhaus (2006, Aufgabe(4.4))*
  Seien X, Z u.i.$N(0,1)$ -verteilt. Dann gilt für $Y = X^2 + Z$, dass

$$E(Y|X) = E\big((X^2+Z)|X\big) = E\big(X^2|X\big) + E(Z) = X^2.$$

$$\widehat{Y} = g(X) = BLUP(Y,X) = E(Y) = 1, \, da$$
$$E(X) = 0, \;\; E(X^3) = 0 \;\; und \quad Cov(X^2, X) = 0.$$

  Außerdem gilt für die Vorhersagefehler:

$$E\big(Y - E(Y|X)\big)^2 = Var\big(Y - E(Y|X)\big) = 1.$$

$$E\big(Y - \widehat{Y}\big)^2 = E\big(Y - E(Y)\big) = Var(Y) = 3.$$

  Auf ähnliche Weise gilt für $Y = X^2 + X + Z$, dass

$$E(Y|X) = E\big(X^2 + X + Z)|X\big) = X^2 + X$$

$$\widehat{Y}_1 = BLUP(Y,X) = E(Y) + \frac{Cov(Y,X)}{Var(X)}X = X + 1$$
$$E(Y - E(Y|X))^2 = 1 \qquad E(Y - \widehat{Y}_1)^2 = 3$$

  sind.

Außerdem sind die folgenden Eigenschaften der Vorhersage nützlich, die sich unmittelbar aus dem Theorem (3.1) ergeben.

Folgerung 3.2 *Wenn $\widehat{Y}$ eine erwartungstreue Vorhersage für Y in der Klasse G ist, dann gilt:*

$$\widehat{Y} \text{ ist } BUP \text{ in } G \qquad \Longleftrightarrow \qquad E[(Y - \widehat{Y})g(\mathbf{X})] = 0, \;\; \forall g \in G$$

Folgerung 3.3 *Seien $\widetilde{Y}, \widehat{Y}$ zwei beste erwartungstreue Vorhersagen für Y in der Klasse G, d.h. $(\widetilde{Y}, \widehat{Y} \in G \cap UP)$. Dann gilt: $\widetilde{Y} = \widehat{Y}$ fast sicher.*

Beweis:
Kleines Hilfsresultat: $\qquad (a-b)^2 = a(a-b) - b(a-b)$

$$\begin{aligned} E[(\widetilde{Y} - \widehat{Y})^2] &= E\big[\big(Y - \widehat{Y} - (Y - \widetilde{Y})\big)^2\big] \\ &= E\big[(Y - \widehat{Y})(\widetilde{Y} - \widehat{Y})\big] - E\big[(Y - \widetilde{Y})(\widetilde{Y} - \widehat{Y})\big] \\ &= 0 \qquad \text{(Folgerung 3.2)}, \end{aligned}$$

hieraus ergibt sich, dass fast sicher $\widetilde{Y} = \widehat{Y}$ ist.

Folgerung 3.4 *Sei $\widehat{Y}$ BUP für Y in der Klasse $G \subseteq SP$, dann gilt:*

$$\begin{aligned} Var(Y-\widehat{Y}) &= Cov(Y-\widehat{Y}, Y-\widehat{Y}) \\ &= Cov(Y-\widehat{Y}, Y) - Cov(Y-\widehat{Y}, \widehat{Y}) \\ &= Cov(Y-\widehat{Y}, Y) \end{aligned}$$

.

Theorem 3.2 *Sei $\widehat{Y}$ BUP für Y in der Klasse $G \subseteq SP$, dann gilt:*

$$Corr^2(Y, \widehat{Y}) = \frac{Var(\widehat{Y})}{Var(Y)}.$$

Beweis:
Da $\widehat{Y}$ BUP ist, dann folgt $Cov(Y-\widehat{Y}, \widehat{Y}) = 0$, und

$$\begin{aligned} Corr^2(Y, \widehat{Y}) &= \frac{Cov^2(\widehat{Y}, Y)}{Var(\widehat{Y})Var(Y)} \\ &= \frac{Var^2(\widehat{Y})}{Var(\widehat{Y})Var(Y)} \\ &= \frac{Var(\widehat{Y})}{Var(Y)} \end{aligned}$$

Beispiel 3.4 *Für $m(\mathbf{X}) = E(Y|\mathbf{X})$ gilt, dass*

$$\min_{g \in SP} Var(Y - g(\mathbf{X})) = Var(Y - m(\mathbf{X})).$$

und aus den Eigenschaften der bedingten Erwartung erhalten wir

$$\begin{aligned} Corr^2(Y, m(\mathbf{X})) &= \frac{Cov^2(m(\mathbf{X}), Y)}{Var(m(\mathbf{X}))Var(Y)} \\ &= \frac{Var^2(m(\mathbf{X}))}{Var(m(\mathbf{X}))Var(Y)} \\ &= \frac{Var(m(\mathbf{X}))}{Var(Y)} \end{aligned}$$

Theorem 3.3 *Sei $\widehat{Y}$ BUP in der Klasse $G \subseteq SP$, dann gilt*

$$Corr^2(Y, \widehat{Y}) = \max_{g \in G} Corr^2(Y, g(\mathbf{X})).$$

Beweis: Es gilt:

$$\begin{aligned} Corr^2(Y, g(\mathbf{X})) &= \frac{Cov^2(Y, g(\mathbf{X}))}{Var(Y)Var(g(\mathbf{X}))} \\ &= \frac{Cov^2(\widehat{Y}, g(\mathbf{X}))}{Var(\widehat{Y}))Var(g(\mathbf{X}))} \ \frac{Var(\widehat{Y})}{Var(Y)} \qquad \text{(Theorem 3.1)} \\ &= Corr^2(\widehat{Y}, g(\mathbf{X})) \ \frac{Cov^2(Y, \widehat{Y})}{Var(\widehat{Y})Var(Y)} \\ &= Corr^2(\widehat{Y}, g(\mathbf{X})) \ Corr^2(Y, \widehat{Y}). \end{aligned}$$

Da $Corr^2(\widehat{Y}, g(\mathbf{X})) \in [0, 1]$ ist, folgt

$$Corr^2(Y, \widehat{Y}) \geq Corr^2(Y, g(\mathbf{X})).$$

3.4. Eigenschaften des mittleren quadratischen Fehlers

In diesem Abschnitt leiten wir verschiedene Formen für den mittleren quadratischen Fehler der Vorherssage her, welche für die nachfolgenden Begriffsbildungen und Aufgabenstellungen nützlich sind.

Folgerung 3.5 *Sei $\widehat{Y}$ BUP in der Klasse $G \subseteq SP$, dann gilt*

$$MSEP(Y, \widehat{Y}) = Var(Y) - Var(\widehat{Y})$$

Beweis:

$$\begin{aligned} MSEP(Y, \widehat{Y}) &= Var(Y - \widehat{Y}) \\ &= Var(Y) + Var(\widehat{Y}) - 2Cov(Y, \widehat{Y}) \\ &= Var(Y) + Var(\widehat{Y}) - 2Var(\widehat{Y}) \\ &= Var(Y) - Var(\widehat{Y}) \end{aligned}$$

Folgerung 3.6 *Sei $\widehat{Y}$ BUP für Y in der Klasse $G \subseteq SP$, dann*

$$MSEP(Y, \widehat{Y}) = Var(Y) - Cov(Y, \widehat{Y}).$$

Folgerung 3.7 *Sei $\widehat{Y}$ BUP für Y in der Klasse $G \subseteq SP$, dann*

$$MSEP(Y, \widehat{Y}) = Var(Y)(1 - Corr^2(Y, \widehat{Y}))$$

Beweis:

$$\begin{aligned} MSEP(Y,\widehat{Y}) &= Var(Y-\widehat{Y}) \\ &= Var(Y) - Corr^2(Y,\widehat{Y})Var(Y) \quad \text{(Theorem 3.2)} \\ &= Var(Y)(1-Corr^2(Y,\widehat{Y})). \end{aligned}$$

Für die beste erwartungstreue Vorhersagen $\widehat{Y}$ in der Klasse $G \subseteq SP$ lässt sich anschaulich die Korrelation $Corr^2(Y,\widehat{Y})$ interpretieren: $Corr^2(Y,\widehat{Y})$ ist ein Maß dafür, um wie viel besser Y durch eine Zufallsvariable $\widehat{Y}$ angenähert werden kann (im Sinne des mittleren quadratischen Fehlers) als durch eine Vorhersage $g_0 \in G_0$. Es gilt nämlich

$$\min_{g\in G} E\big[\big(Y-g(X)\big)^2\big] = \big(1-Corr^2(Y,\widehat{Y})\big) \min_{g_0\in G_0} E\big[\big(Y-g_0(X)\big)^2\big]$$

Zum Beweis: $E\big[\big(Y-g_0(X)\big)^2\big] = Var(Y) + \big(E(Y)-g_0(X)\big)^2$, hat $Var(Y)$ als minimalen Wert, während

$$\begin{aligned} E\big[\big(Y-g(X)\big)^2\big] &= E\big[\big(Y-\widehat{Y}+\widehat{Y}-g(X)\big)^2\big] \\ &= Var(Y)(1-Corr^2(Y,\widehat{Y})) + \big(g(X)-\widehat{Y}\big)^2 \qquad \text{(Folgerung 3.7)} \end{aligned}$$

als Minimum $Var(Y)(1-Corr^2(Y,\widehat{Y}))$ für $\widehat{Y} = g(X)$ annimmt.

Bemerkung 3.2 *Aus Theorem 3.2 und Folgerung 3.7 ergibt sich, dass*

$$Corr^2(Y,\widehat{Y}) = 1 - \frac{Var(Y-\widehat{Y})}{Var(Y)} \tag{3.7}$$

Interpretation:

(i) $Corr^2(Y,\widehat{Y})$ ist ein Maß der Genauigkeit der Vorhersage, das man beim Vergleich verschiedener Situationen oder verschiedener Klassen von Vorhersagen verwendet.

(ii) Je näher der Ausdruck (3.7) bei 1 liegt, desto kleiner ist der Vorhersagefehler

$$Var(Y-\widehat{Y})$$

d.h. desto besser ist die Vorhersage.

(iii) Im Extremfall gilt:

$Corr^2(Y,\widehat{Y}) = 1 \Longleftrightarrow Var(Y-\widehat{Y}) = 0$ d.h. $Y = \widehat{Y}$ f.s. und die Vorhersage ist perfekt
$Corr^2(Y,\widehat{Y}) = 0 \Longleftrightarrow Var(Y) = Var(Y-\widehat{Y})$ d.h. $\widehat{Y} \in G_0$ und die Vorhersage ist gleich dem Mittelwert $\widehat{Y} = \mu_y$ f.s.

Beachte:
Wenn $\widehat{Y}$ BUP für Y auf G ist, ergibt sich hieraus und aus den Darstellungsformeln für den mittleren quadratischen Fehler, dass

a)
$$Var(Y) - Var(\widehat{Y}) = \min_{g \in G} MSEP(Y, g(\mathbf{X})).$$

b)
$$Var(Y) - Cov(Y, \widehat{Y}) = \min_{g \in G} MSEP(Y, g(\mathbf{X})).$$

c)
$$Var(Y)(1 - Corr^2(Y, \widehat{Y})) = \min_{g \in G} MSEP(Y, g(\mathbf{X})).$$

Mit Hilfe der obigen Formeln lässt sich die optimale Lösung der Minimierung des mittleren quadrierten Vorhersagefehlers wie folgt schreiben:

a)
$$Var(Y) - Var(\widehat{Y}) = \min_{g \in G \cap UP} Var(Y - g(\mathbf{X}))$$

b)
$$Var(Y) - Cov(Y, \widehat{Y}) = \min_{g \in G \cap UP} Var(Y - g(\mathbf{X}))$$

c)
$$Var(Y)(1 - Corr^2(Y, \widehat{Y})) = \min_{g \in G \cap UP} Var(Y - g(\mathbf{X})).$$

Eine nützliche Erweiterung obiger Begriffe ist die folgende Äquivalenz

Lemma 3.4 *Es gilt:*

$$\left[\begin{array}{c} \widehat{Y}(\mathbf{X}) \text{ erwartungstreu in der Klasse } G \subseteq SP \text{ mit} \\ \\ MSEP(Y, \widehat{Y}(\mathbf{X})) = Var(Y) - Var(\widehat{Y}(\mathbf{X})) \end{array} \right] \Leftrightarrow \left[\begin{array}{c} \widehat{Y}(\mathbf{X}) = \mu_y \\ oder \\ \widehat{Y}(\mathbf{X}) \ BUP \end{array} \right].$$

3.5. Modell der Dimensionsreduktion

Das Ziel einer Dimensionsreduktion ist, eine Abbildung zu finden, die von einem höherdimensionalen Raum in einen niederdimensionalen Raum abbildet. So werden häufig Projektionen der Daten auf einen niedrigdimensionalen Unterraum betrachtet, wobei dann zu entscheiden ist, welche der möglichen Projektionen, z.B. alle zweidimensionalen Koordinaten, auszuwählen sind. Aus mathematischer Sicht hat eine Dimensionsreduktion häufig das Ziel, eine niedrigere Dimensionalität zu erreichen, indem eine Abbildung $f : \mathbb{R}^p \to \mathbb{R}^d, d < p$, benutzt wird. Dabei werden wir in dieser Arbeit nur endlich

dimensionale Vektorräume betrachtet. Bei statistischen Fragestellungen kann unter Dimensionsreduktion sowohl eine Reduzierung der Anzahl von Variablen als auch eine Verringerung der Anzahl von Beobachtungen verstanden werden. Hier bezieht sich der Begriff Dimensionsreduktion auf die Anzahl von Variablen. Natürlich sollte eine Reduktion so durchgeführt werden, dass keine wesentlichen Informationen verloren gehen. Um zu einer mathematischen Präzisierung dieser Forderung zu gelangen, stellen wir zunächst folgende Überlegung an:

Modelldefinition

Wir betrachten in diesem Abschnitt weiter das Modell 1.3.a mit $q = 1,\ p > 1$, d.h.

$$\begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \text{ eine } \mathbb{R}^{p+1} \text{ wertige ZV, wobei}$$
$$\boldsymbol{\mu} \in \mathbb{R}^{p+1}, \mathbf{V} \in PD(p+1) \text{und bekannt sind.}$$

Ein solches Modell für Dimensionsreduktion wird von Li (1991) und Cook (1994) verwendet und im Folgenden vorgestellt:
Zwischen einer Zielgrößer Y und einem Vektor $\mathbf{X}$ gelte der Zusammenhang:

$$\begin{aligned} Y &= d(Z_1, \cdots, Z_k, \epsilon), \quad \text{mit} \quad k < p, \\ &= d(\mathbf{Z}, \epsilon), \end{aligned}$$

so dass $Z_1 = \mathbf{a}_1^T\mathbf{X}, \cdots, Z_k = \mathbf{a}_k^T\mathbf{X}$ sind. Die Matrixformulierung ist:

$$\mathbf{Z} = \mathbf{U}^T\mathbf{X} \quad \text{mit} \quad \mathbf{U} = [\mathbf{a}_1, \ldots, \mathbf{a}_k],$$

wobei die folgenden Voraussetzungen vereinbart werden:

(i) ϵ sei eine $\mathbb{R}$-wertige ZV, so dass $\mathbf{X}, \epsilon$ unabhängig seien.

(ii) Die Kovarianz V_ϵ der Residuen ϵ sei unbekannt.

(iii) Die Vektoren $\mathbf{a}_1, \cdots, \mathbf{a}_k \in \mathbb{R}^p$ heißen dimensionsreduzierende Richtung (Abkürzung DR-Richtung), mit $\mathbf{U} = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$ eine $p \times k$-Matrix.

(iv) Die unbekannte Abbildung $d : \mathbb{R}^k \to \mathbb{R}$ werde als Linkfunktion der Reduktion bezeichnet.

Interpretation:

1. Der Spaltenraum von $\mathbf{U}$ wird mit $\mathcal{U} = span(\mathbf{U})$ bezeichnet. Dieser Unterraum heißt dimensionsreduzierender Unterraum (DRU). Weiterhin sei angenommen, dass $\mathcal{U}$ ein eindeutig bestimmer DRU minimaler Dimension k sei mit $1 \leq k \leq p$ d.h. man hofft, dass k klein ist.

2. Die Vektoren $\mathbf{a}_1, \ldots, \mathbf{a}_k \in \mathbb{R}^p$ bzw. der von ihnen aufgespannte Raum $\mathcal{U} = span(\mathbf{U})$ sowie die minimale Anzahl k sind unbekannt.

3. Die Abbildung $d : \mathbb{R}^k \to \mathbb{R}$ ist unbekannt. Das primäre Ziel besteht also zunächst darin, eine Dimensionsreduktion zu erreichen, die die relevanten Informationen in $\mathbf{X}$ repräsentiert. Aus mathematischer Sicht besteht das Ziel darin, einen Unterraum zu suchen, der durch $\mathbf{U} = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$ aufgespannt wird, so dass $k << p$. Da $\mathbf{U} = [\mathbf{a}_1, \ldots, \mathbf{a}_k]$ nicht eindeutig bestimmt ist, wird vorausgesetzt, dass $\mathbf{a}_1, \ldots, \mathbf{a}_k$ unabhängig sind und eine Orthonormalbasis von $\mathcal{U} = span(\mathbf{U})$ bilden, d.h. es gilt, dass $\mathbf{U}^T\mathbf{U} = \mathbf{I}_k$ ist.

4. Die Effizienz dieser Reduktion ist gegeben durch:

$$Eff(d(Z)) = \frac{MSEP(Y, BLUP(Y, \mathbf{X}))}{MSEP(Y, BLUP(Y, d(Z)))}$$

Beispiel 3.5 ***(AR(1)-Modell)***
Wir betrachten das Modell 1.6 (Normalverteilung mit autoregressiver Struktur AR(1):)

$$\begin{pmatrix} X_1 \\ X_2 \\ Y \end{pmatrix} \sim N_3 \left(\begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \\ \mu_y \end{pmatrix}, \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \right) \quad mit \;\; |\rho| < 1.$$

Die Unterräume:

$$\mathcal{U}_1 = span((1,0)^T),\ \mathcal{U}_2 = span((0,1)^T),\ \mathcal{U}_3 = span((\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T)$$

sind dimensionsreduzierende Unterräume DRU.

- *Das reduzierte Modell unter Unterraum $\mathcal{U}_1$ ist:*

$$\begin{pmatrix} Z_1 \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_{x_1} \\ \mu_y \end{pmatrix}, \begin{pmatrix} 1 & \rho^2 \\ \rho^2 & 1 \end{pmatrix} \right), \; mit$$

$$Z_1 = X_1 \quad und \;\; d(Z_1) = BLUP(Y, Z_1).$$

$$MSEP(Y, BLUP(Y, Z_1)) = 1 - \rho^4 \quad und \quad Eff(d(Z_1)) = \frac{1}{1+\rho^2}.$$

- *Das reduzierte Modell unter Unterraum $\mathcal{U}_2$ ist:*

$$\begin{pmatrix} Z_2 \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_{x_2} \\ \mu_y \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \; mit$$

$$Z_2 = X_2 \quad und \;\; d(Z_2) = BLUP(Y, Z_2).$$

$$MSEP(Y, BLUP(Y, Z_2)) = 1 - \rho^2 \quad und \quad Eff(d(Z_1)) = 1$$

$$Eff(d(Z_1)) = 1 \qquad (Eigenschaften\ des\ AR(1)\text{-}Modells)$$

- *Das reduzierte Modell unter Unterraum $\mathcal{U}_3$ ist:*

$$\begin{pmatrix} Z_3 \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \frac{1}{\sqrt{2}}(\mu_{x_1} + \mu_{x_2}) \\ \mu_y \end{pmatrix}, \begin{pmatrix} 1+\rho & \frac{1}{\sqrt{2}}(\rho^2+\rho) \\ \frac{1}{\sqrt{2}}(\rho^2+\rho) & 1 \end{pmatrix} \right), mit$$

$$Z_3 = \frac{1}{\sqrt{2}}(X_1 + X_2) \quad und \quad d(Z_3) = BLUP(Y, Z_3)$$

$$MSEP(Y, BLUP(Y, Z_3)) = 1 - \frac{\rho^2(1+\rho)}{2}, \qquad Eff(d(Z_3)) = \frac{1-\rho^2}{1-\frac{\rho^2(1+\rho)}{2}}$$

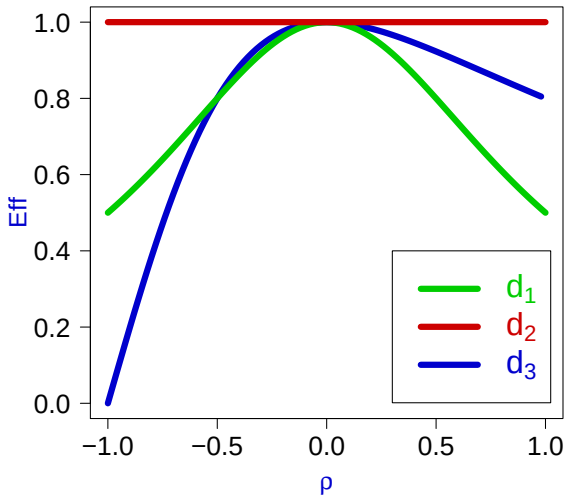

Abbildung 3.2.: Effizienz der Reduktioen nach d_1, d_2, d_3

Die Abbildung (3.2) zeigt die Effizienzfunktionen der Dimensionsreduktionen d_1, d_2 und d_3 in Abhängigkeit von ρ. Man sieht, falls ρ an der Stelle nahe bei 0 liegt, ist die Effizienz von d_1, d_3 nahe 1 und es gibt keinen Informationsverlust bei $d_2, \forall\ |\rho| < 1$.

4. Vorhersage in Linearen Modellen

Eines der wichtigsten Ziele der Regressionsrechnung ist es, den gefundenen Zusammenhang auszunutzen, um zukünftige Werte zu prognostizieren, wenn man eine neue Beobachtung für den Regressor erhält. Dabei geht man davon aus, dass der durch die Regressionsgerade beschriebene Zusammenhang auch in der Zukunft zumindest für einen gewissen Zeitraum Gültigkeit besitzt.
Wir beginnen mit dem folgenden Beispiel, um einige Ideen einzubringen und die nachfolgende Begriffsbildungen zu motivieren.

Beispiel 4.1 *(Stochastisch unabhängige identische reelle Zufallsvariablen)*

$$X_1, \ldots, X_n, X_{n+1} \quad u.i.v.\ N(\beta, \sigma^2), \tag{4.1}$$

wobei, der Parameter $\theta = (\beta, \sigma^2)^T \in \mathbb{R} \times (0, \infty)$ *unbekannt ist. Die Zufallsvariablen* $X_1, \ldots, X_n$ *seien beobachtbar und* X_{n+1} *nicht beobachtbar. Das Ziel ist, die Vorhersage von* X_{n+1} *zu bestimmen.*
Maßtheoretische Modellformulierung:

$$\left(\mathbb{R}^{n+1}, \mathcal{B}^{n+1}, (P_{\beta,\sigma^2})_{(\beta,\sigma^2)\in\mathbb{R}\times(0,\infty)}\right), \text{ wobei } P_{\beta,\sigma^2} = [N(\beta, \sigma^2)]^n \ \otimes \ N(\beta, \sigma^2).$$

- *Die beste lineare Vorhersage von* X_{n+1} *unter* $X_1, \ldots, X_n$ *ist*

$$\begin{aligned} \widehat{X}_{n+1} &= \widehat{\beta} \\ &= \frac{1}{n}\mathbf{1}_n^T\mathbf{X} \\ &= \bar{X} \end{aligned}$$

- *Damit gilt insbesondere:*

$$\begin{aligned} Var(X_{n+1} - \widehat{X}_{n+1}) &= Var(X_{n+1}) + Var(\widehat{X}_{n+1}) - 2\ Cov(X_{n+1}, \widehat{X}_{n+1}) \\ &= \sigma^2 + \frac{\sigma^2}{n}. \end{aligned}$$

Nach diesem einführenden Beispiel kehren wir zur Modelldarstellung zurück.

4.1. Modelldefinition

Definition 4.1 ***(Lineares Modell)***
Seien $p, n \in \mathbb{N}$ *mit* $p < n$. *Ein lineares Modell für* $\mathbb{R}^n$*-wertige beobachtete Zufallsvariable* $\mathbf{X}$ *mit unbekanntem* $(p+1)$*-dimensionalen Parameter* $\theta = (\beta^T, \sigma^2)^T \in \mathbb{R}^p \times (0, \infty)$ *besteht aus:*

- *einer reellen bekannten $n \times p-$Designmatrix $\mathbf{F}$ von vollem Rang p, und*
- *einem Zufallsvektor $\mathbf{e_x} = (e_1, \ldots, e_n)^T$ von n standardisierten Zufallsvariablen, den Fehlern.*

Der $n-$dimensionale Beobachtungsvektor $\mathbf{X}$ ergibt sich durch die Matrixformulierung:

$$\mathbf{X} = \mathbf{F}\beta \ + \mathbf{e_x}, \qquad \text{mit} \quad Cov(\mathbf{e_x}) = \sigma^2 \mathbf{V_{xx}}, \ \mathbf{V_{xx}} \text{ bekannt} \tag{4.2}$$

Das zugehörige statistische Experiment unter Normalverteilungsannahme ist:

$$\left(\mathbb{R}^n, \mathcal{B}^n, N_n(F\beta, \sigma^2 V)_{\theta \in \Theta}\right) : \Theta = \mathbb{R}^p \times (0, \infty), \quad \theta = (\beta^T, \sigma^2)^T.$$

Basierend auf der Definition des linearen Modells ergibt sich die Vorhersage im linearen Modell durch die Matrixformulierung: (siehe Puntanen, Styan und Isotalo (2011, Kap.10.9))

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{F} \\ \mathbf{F}_0 \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{e_x} \\ \mathbf{e_y} \end{pmatrix}, \tag{4.3}$$

mit

- $\mathbf{F}_0$ einer reellen bekannten $q \times p-$ Designmatrix von vollem Rang $q : q < p$
- $\mathbf{Y}$ nicht beobachtbare ($q-$dimensionale) Zufallsvariable.

$$Y : (\Omega, \mathcal{C}) \to (\mathbb{R}^q, \mathcal{B}^q),$$

- $\mathbf{e_x}, \mathbf{e_y}$ gemeinsam verteilt, mit Kovarianzmatrix:

$$\Sigma = Cov \begin{pmatrix} \mathbf{e_x} \\ \mathbf{e_y} \end{pmatrix} = \sigma^2 \mathbf{V}, \quad \mathbf{V} = \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{xy}^T} & \mathbf{V_{yy}} \end{pmatrix} \in PD(n+q) \ \text{ bekannt.}$$

Maßtheoretische Modellformulierung zum Beispiel unter Normalverteilungsannahme:

$$\left(\mathbb{R}^{n+q}, \mathcal{B}^{n+q}, \left(N_{n+q}\left(\begin{bmatrix} \mathbf{F}\boldsymbol{\beta} \\ \mathbf{F}_0\boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{xy}^T} & \mathbf{V_{yy}} \end{bmatrix}\right)\right)_{(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^p \times (0,\infty)}\right). \tag{4.4}$$

Definition 4.2 ***(lineare Vorhersage)*** *Puntanen, Styan und Isotalo (2011, Kap.10.9) Eine spezielle Form der linearen Vorhersage für $\mathbf{Y}$ in diesem Modell ist:*

$$\mathbf{Y}(\widehat{\boldsymbol{\beta}}) = \mathbf{F}_0\widehat{\boldsymbol{\beta}} + \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}(\mathbf{X} - \mathbf{F}\widehat{\boldsymbol{\beta}}), \tag{4.5}$$

wobei $\widehat{\boldsymbol{\beta}}$ ein Schätzer für $\boldsymbol{\beta}$ ist, (engl. linear prediction (LP)).

Als Folgerung der Definition 4.2 ergibt sich:

Folgerung 4.1 ***(BLUP-Vorhersage)*** *Puntanen, Styan und Isotalo (2011, Kap.10.9) und Isotalo, Puntanen (2009)*

1) *Eine lineare Vorhersage* $\mathbf{Y}(\widehat{\boldsymbol{\beta}})$ *in der Form (4.5) heißt erwartungstreue Vorhersage, wenn* $\widehat{\boldsymbol{\beta}}$ *erwartungstreuer Schätzer für* $\boldsymbol{\beta}$ *ist, (engl. linear unbiased prediction (LUP)).*

2) *Wenn* $\widehat{\boldsymbol{\beta}}$ *ein BLUE-Schätzer ist, heißt*

$$\begin{aligned} \mathbf{Y}(\widehat{\boldsymbol{\beta}}) &= \mathbf{F}_0\widehat{\boldsymbol{\beta}} + \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}(\mathbf{X} - \mathbf{F}\widehat{\boldsymbol{\beta}}) && (4.6)\\ &= BLUE(\mathbf{F}_0\boldsymbol{\beta}) + \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\big(\mathbf{X} - BLUE(\mathbf{F}\boldsymbol{\beta})\big) && (4.7) \end{aligned}$$

die beste lineare erwartungstreue Vorhersage (engl. best linear unbiase prediction (BLUP)).

Einige Eigenschaften

Lemma 4.1 *Schmidt(1988)*
Seien $\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$, $\mathbf{G} = \begin{pmatrix} \mathbf{F}\boldsymbol{\beta} \\ \mathbf{F}_0\boldsymbol{\beta} \end{pmatrix}$ *und* $\mathbf{V} = \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{xy}^T} & \mathbf{V_{yy}} \end{pmatrix}$, *dann gilt:*

$$E\big((\mathbf{Z} - \mathbf{G}\widehat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}(\mathbf{Z} - \mathbf{G}\widehat{\boldsymbol{\beta}})\big) = E\big((\mathbf{X} - \mathbf{F}\widehat{\boldsymbol{\beta}})^T\mathbf{V_{xx}^{-1}}(\mathbf{X} - \mathbf{F}\widehat{\boldsymbol{\beta}})\big) +$$

$$E\big((\mathbf{Y} - \mathbf{F}_0\widehat{\boldsymbol{\beta}})^T(\mathbf{V_{yy}} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{V_{xy}})^{-1}(\mathbf{Y} - \mathbf{F}_0\widehat{\boldsymbol{\beta}})\big),$$

wenn $\widehat{\boldsymbol{\beta}}$ *ein linearer erwartungstreuer Schätzer ist.*

Lemma 4.2 ***(Kovarianz der Residuen)*** *Sengupta und Jammalamadaka (2003, Kap.7.13.1)*
Wenn $\widehat{\boldsymbol{\beta}}$ *ein linearer erwartungstreuer Schätzer ist, dann hat der mittlere quadratische Fehler der Vorhersage die folgende Form*

$$\begin{aligned} Cov(\mathbf{Y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}})) &= \sigma^2(\mathbf{V_{yy}} - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{V_{xy}}) \\ &\quad + (\mathbf{F}_0 - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{F})\; Cov(\widehat{\boldsymbol{\beta}})\; (\mathbf{F}_0^T - \mathbf{F}^T\mathbf{V_{xx}^{-1}}\mathbf{V_{xy}}) \end{aligned}$$

Im Hinblick auf die Definition 4.2 gilt die folgende spezielle Version der Kovarianzmethode (siehe Lemma 3.3).

Lemma 4.3 ***(Kovarianzmethode)***
Seien $\widetilde{\boldsymbol{\beta}}, \boldsymbol{\beta}^*$ *und* $\boldsymbol{\beta}^{**}$ *lineare erwartungstreue Schätzer, dann gilt:*

$$Cov(\mathbf{Y} - \mathbf{Y}(\widetilde{\boldsymbol{\beta}}), \mathbf{Y}(\boldsymbol{\beta}^*) - \mathbf{Y}(\boldsymbol{\beta}^{**})) = \mathbf{0}_{q\times q} \quad \Leftrightarrow \quad \mathbf{Y}(\widetilde{\boldsymbol{\beta}}) \text{ ist BLUP},$$

Beweis: (Für eindimensionale Aspekte siehe Beweis von Theorem 3.1)
Notwendig: Seien $\mathbf{Y}(\widetilde{\boldsymbol{\beta}})$ BLUP, dann ist $\mathbf{Y}(\widetilde{\boldsymbol{\beta}})-\lambda(\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))$ LUP für $\mathbf{Y}$, $\forall\lambda\in\mathbb{R}$ mit

$$E\big(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}})+\lambda(\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))\big)=\mathbf{0}_{q\times 1}.$$

Hieraus ergibt sich, dass

$$\begin{aligned} Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}})+\lambda(\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))) &= Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}))+\lambda^2 Cov(\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**})) \\ &\quad +2\,\lambda Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**})) \\ &\quad \geq Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}})) \end{aligned}$$

$$\Rightarrow \lambda^2 Cov(\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))+2\lambda Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))\geq\mathbf{0}_{q\times q}$$

Hieraus ergibt sich

$$Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))=\mathbf{0}_{q\times q},\quad \text{da}$$

wenn $Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))<\mathbf{0}_{q\times q}$, $\exists\lambda\in\mathbb{R}$, so dass die Matrix

$$\lambda^2 Cov(\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))+\lambda Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))<\mathbf{0}_{q\times q}.$$

ist. Wenn $Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))>\mathbf{0}_{q\times q}$, $\exists\lambda\in\mathbb{R}$, mit

$$\lambda^2 Cov(\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))+\lambda Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))<\mathbf{0}_{q\times q}.$$

ist, d.h. $\mathbf{Y}(\widetilde{\boldsymbol{\beta}})$ nicht BUP, wenn $Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\boldsymbol{\beta}^*)-\mathbf{Y}(\boldsymbol{\beta}^{**}))\neq\mathbf{0}_{q\times q}$
Hinreichend: $\mathbf{Y}(\boldsymbol{\beta}^*)$ ist eine lineare erwartungstreue Vorhersage für $\mathbf{Y}$, dann ist $E(\mathbf{Y}(\widetilde{\boldsymbol{\beta}})-\mathbf{Y}(\boldsymbol{\beta}^*))=\mathbf{0}_{q\times 1}$,

$$Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\widetilde{\boldsymbol{\beta}})-\mathbf{Y}(\boldsymbol{\beta}^*))=\mathbf{0}_{q\times q}$$

und

$$\begin{aligned} Cov(\mathbf{Y}-\mathbf{Y}(\boldsymbol{\beta}^*)) &= Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}})+\mathbf{Y}(\widetilde{\boldsymbol{\beta}})-\mathbf{Y}(\boldsymbol{\beta}^*)) \\ &= Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}))+Cov(\mathbf{Y}(\widetilde{\boldsymbol{\beta}})-\mathbf{Y}(\boldsymbol{\beta}^*)) \\ &\quad +2Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}),\mathbf{Y}(\widetilde{\boldsymbol{\beta}})-\mathbf{Y}(\boldsymbol{\beta}^*)) \\ &= Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}))+Cov(\mathbf{Y}(\widetilde{\boldsymbol{\beta}})-\mathbf{Y}(\boldsymbol{\beta}^*)). \end{aligned}$$

Damit ist $Cov(\mathbf{Y}-\mathbf{Y}(\widetilde{\boldsymbol{\beta}}))\leq Cov(\mathbf{Y}-\mathbf{Y}(\boldsymbol{\beta}^*))$ bzw. $\mathbf{Y}(\widetilde{\boldsymbol{\beta}})$ BLUP.

Spezialfall:

Wir betrachten das Modell mit der Annahme, dass $\mathbf{V_{xy}} = \mathbf{0}_{n\times q}$ ist, dann $V = \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{0} \\ \mathbf{0} & \mathbf{V_{yy}} \end{pmatrix}$. Aus der Theorie der linearen Modelle erhalten wir:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{GM} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{X} - \mathbf{F}\boldsymbol{\beta})^T \mathbf{V}_{\mathbf{xx}}^{-1} (\mathbf{X} - \mathbf{F}\boldsymbol{\beta}) \\ &= (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{X}. \\ &\quad \text{BLUE für } \boldsymbol{\beta} \\ \widehat{\boldsymbol{\beta}}_{OLS} &= \arg\min_{\boldsymbol{\beta}} (\mathbf{X} - \mathbf{F}\boldsymbol{\beta})^T (\mathbf{X} - \mathbf{F}\boldsymbol{\beta}) \\ &= (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{X}. \end{aligned}$$

Statistische Eigenschaft von $\widehat{\boldsymbol{\beta}}_{GM}$ und $\widehat{\boldsymbol{\beta}}_{OLS}$:

- Erwartungswert:

$$\begin{aligned} E(\widehat{\boldsymbol{\beta}}_{GM}) &= E((\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{X}) \\ &= (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \\ E(\widehat{\boldsymbol{\beta}}_{OLS}) &= E((\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{X}) \\ &= (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{F}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

- Kovarianzmatrix:

$$\begin{aligned} Cov(\widehat{\boldsymbol{\beta}}_{GM}) &= Cov((\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{X}) \\ &= (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{V}_{\mathbf{xx}} \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F} (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \\ &= (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \\ Cov(\widehat{\boldsymbol{\beta}}_{OLS}) &= Cov((\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{X}) \\ &= (\mathbf{F}^T \mathbf{F})^{-1} (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}} \mathbf{F}) (\mathbf{F}^T \mathbf{F})^{-1}. \end{aligned}$$

Hieraus ergibt sich

$$\begin{aligned} \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}) &= \mathbf{F}_0 (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{X}. \\ \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{OLS}) &= \mathbf{F}_0 (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{X} \end{aligned}$$

Statistische Eigenschaft der besten linearen Vorhersage:

- Erwartungswert und Kovarianzmatrix:

$$\begin{aligned} E(\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM})) &= \mathbf{F}_0 (\mathbf{F}^T \mathbf{V}_{\mathbf{xx}}^{-1} \mathbf{F})^{-1} \mathbf{F}^T \mathbf{V}_{\mathbf{xx}} \mathbf{F}\boldsymbol{\beta} \\ &= \mathbf{F}_0 \boldsymbol{\beta} = E(\mathbf{Y}). \\ E(\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{OLS})) &= \mathbf{F}_0 (\mathbf{F}^T \mathbf{F})^{-1} \mathbf{F}^T \mathbf{F}\boldsymbol{\beta} = \mathbf{F}_0 \boldsymbol{\beta} = E(\mathbf{Y}). \end{aligned}$$

$$\begin{aligned} Cov(\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM})) &= \mathbf{F}_0(\mathbf{F}^T\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{F})^{-1}\mathbf{F}_0^T. \\ Cov(\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{OLS})) &= \mathbf{F}_0(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}^T\mathbf{V}_{\mathbf{xx}}\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}\mathbf{F}_0^T. \end{aligned}$$

und

$$Cov(\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}), \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{OLS}) = \mathbf{F}_0(\mathbf{F}^T\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{F})^{-1}\mathbf{F}_0^T.$$

- Hieraus ergibt sich:

$$\begin{aligned} Cov\big(\mathbf{Y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}), \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}) - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{OLS})\big) &= Cov\big(\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}), \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{OLS}) - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM})\big) \\ &= \mathbf{F}_0(\mathbf{F}^T\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{F})^{-1}\mathbf{F}_0^T - \mathbf{F}_0(\mathbf{F}^T\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{F})^{-1}\mathbf{F}_0^T \\ &= \mathbf{0}. \end{aligned}$$

Durch obigen Spezialfall ergibt sich die Bemerkung:

Bemerkung 4.1 *Im Fall, dass* $\mathbf{X}$ *und* $\mathbf{Y}$ *unabhängig sind, hat das Lemma 4.3 die folgende Form:*
Sei $\widetilde{\boldsymbol{\beta}}$ *linearer erwartungstreuer Schätzer, dann gilt:*

$$\widetilde{\boldsymbol{\beta}} \text{ ist BLUP} \iff Cov(\mathbf{Y} - \mathbf{Y}(\widetilde{\boldsymbol{\beta}}), \mathbf{Y}(\boldsymbol{\beta}^*)) = \mathbf{0}, \quad \forall \boldsymbol{\beta}^* \in U_0.$$

Beispiel 4.2 ***(Autoregressive Kovarianz)***
Wir betrachten das Modell:

$$\begin{pmatrix} X_1 \\ X_2 \\ Y \end{pmatrix} \sim N_3\left(\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\beta, \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}\right), \qquad \text{mit } |\rho| \leq 1 \text{ bekannt.}$$

Parameter des Modells ist: $\beta \in \mathbb{R}$.
Modellannahme: $\mathbf{X}$ *beobachtbar,* Y *nicht beobachtbar. Aus der Theorie der linearen Modelle und mit Hilfe der obigen Definitionen und Theoreme sieht man, dass*

- *Der Gauss Markov-Schätzer, ML-Schätzer ist:*

$$\begin{aligned} \widehat{\beta}_{GM} &= (\mathbf{1}_2^T\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{1}_2)^{-1}\mathbf{1}_2^T\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{X} \\ &= (\mathbf{1}_2^T\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\mathbf{1}_2)^{-1}\mathbf{1}_2^T\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\mathbf{X} \\ &= \frac{1}{2(1-\rho)}(1-\rho)(X_1 + X_2) \\ &= \bar{\mathbf{X}} \\ &= \widehat{\beta}_{OLS}. \end{aligned}$$

- *Die beste lineare erwartungstreue Vorhersage ist:*

$$\begin{aligned} Y(\widehat{\beta}_{GM}) &= \widehat{\beta}_{GM} + \mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-1}(\mathbf{X} - \widehat{\beta}_{GM}) \\ &= \widehat{\beta}_{GM} + \frac{\rho}{1-\rho^2}\begin{pmatrix} \rho & 1 \end{pmatrix}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}(\mathbf{X} - \widehat{\beta}_{GM}) \\ &= \bar{\mathbf{X}} + \rho\begin{pmatrix} 0 & 1 \end{pmatrix}(\mathbf{X} - \bar{\mathbf{X}}) \\ &= \bar{\mathbf{X}} + \rho(\mathbf{X}_2 - \bar{\mathbf{X}}) \\ &= \frac{(1-\rho)X_1 + (1+\rho)X_2}{2}. \end{aligned}$$

- *Wegen Lemma(4.3) ist*

$$\begin{aligned} Cov(Y - Y(\widehat{\beta}_{GM}), X_2 - X_1) &= Cov(Y, X_2 - X_1) - Cov(Y(\widehat{\beta}_{GM}), X_2 - X_1) \\ &= \rho - \rho^2 - \frac{1}{2}\, Cov\big((1-\rho)X_1 + (1+\rho)X_2, X_2 - X_1\big) \\ &= \rho - \rho^2 - \frac{1}{2}(2\rho - 2\rho^2) = 0 \end{aligned}$$

4.2. Vorhersageintervalle

Die obige Vorgehensweise lässt sich auf die Vorhersage eines Werts anwenden. Die Idee besteht darin die gemeinsame Verteilung des Zufallsvektors X und der unbeobachtbaren Zufallsvariablen Y zu betrachten. Es ist allerdings unerlässlich, neben einem Vorhersagewert stets eine Angabe über seine Qualität oder seine Präzision zu machen. So kann man beispielsweise mit einigen wenigen Beobachtungen einen Vorhersagewert ausrechnen und diesen angeben, dieser hat aufgrund seiner großen Varianz eine geringe Aussagekraft. Erst durch eine ausreichend hohe Stichprobenzahl kann eine hinreichende Präzision garantiert werden. Naturlich hängt die Präzision immer mit dem gewählten Modell und der Aufgabenstellung zusammen, so dass allein die Größe der Stichproben allein kein zuverlässiges Qualitätsmerkmal darstellt. Ein zuverlässiges und allgemeines Merkmal für die Qualtität einer Vorhersage ist ein Vorhersageintervall. Für Beweise und weitere Ergebnisse verweisen wir auf die Literatur, z.B. Shao (2003, Kap.7.1.4), Rao, Shalabh, Toutenburg und Heumann (2008, Kap.6.7) und Christensen (2013).

Grundkonzept:

Definition 4.3 *(**Vorhersagebereich,** $(1-\alpha)$**-Vorhersagebereich**)*

(i) *Eine Abbildung $\mathcal{I} : \mathbb{R}^n \longrightarrow \mathcal{B}(\mathbb{R}^q)$ heißt Vorhersagebereich für $\mathbf{Y}$, wenn*

$$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{z} \in \mathcal{I}(\mathbf{x})\} \in \mathcal{B}^n, \forall \mathbf{z} \in \mathbb{R}^q.$$

(ii) *Die Wahrscheinlichkeit*

$$P(\mathbf{Y} \in \mathcal{I}(\mathbf{X})), \tag{4.8}$$

heißt die Überdeckungswahrscheinlichkeit des Vorhersagebereichs $\mathcal{I}$.

(iii) *Für ein kleines Toleranzniveau $\alpha \in (0,1)$ heißt $\mathcal{I}$ ein $(1-\alpha)$-Vorhersagebereich, wenn gilt:*

$$P(\mathbf{Y} \in \mathcal{I}(\mathbf{X})) \geq 1 - \alpha. \tag{4.9}$$

Beachte: Ziel ist die Konstruktion von $(1-\alpha)$-Vorhersagebereichen für ein vorgegebenes kleines α-Niveau (z.B. $\alpha = 0,05$), die möglichst kleine Bereiche liefert, welche die Überdeckungseigenschaft (4.9) erfüllen, das heißt die Vorhersagebereiche sollten möglichst klein sein und möglichst wenige falsche Werte überdecken.

Beispiele und Bemerkungen:

Bei einer Prognose macht man zwangsläufig einen Prognosefehler. Der Prognosefehler ist die Abweichung der Prognose von dem wahren Wert $\mathbf{y}$. Mit Hilfe der Eigenschaften der Parameterschätzer kann ein Intervall für den Prognosefehler hergeleitet werden. Dies ist zur Beurteilung der Prognose oft sehr hilfreich. Wir untersuchen diesen Begriff des Vorhersageintervalls in einigen Beispielen.

(i) Wir gehen wieder von folgendem Modell aus:

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N_{n+q}\left(\begin{bmatrix} \mathbf{F}\boldsymbol{\beta} \\ \mathbf{F}_0\boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{xy}^T} & \mathbf{V_{yy}} \end{bmatrix} \right) \text{ mit } (\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^p \times (0, \infty).$$

Die beste lineare erwartungstreue Vorhersage für $\mathbf{Y}$ ist gegeben durch:

$$\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}) = \mathbf{F}_0\widehat{\boldsymbol{\beta}}_{GM} + \mathbf{V_{xy}^T}\mathbf{V_{xx}^{-1}}(\mathbf{X} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{GM}).$$

Wegen $\widehat{\boldsymbol{\beta}}_{GM} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{F}^T\mathbf{V_{xx}^{-1}}\mathbf{F})^{-1})$ gilt für die Linearkombination

$$\mathbf{Y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}) \sim N_q(0, \sigma^2\mathbf{D}),$$

wobei

$$\mathbf{D} = Cov(\mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}) - \mathbf{Y})$$

Aus den Verteilungsresultaten der Normalverteilung erhalten wir, dass

$$\frac{(\mathbf{Y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}))^T\mathbf{D}^{-1}(\mathbf{Y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}))}{\sigma^2} \sim \chi_q^2$$

ist. Wenn wir σ^2 durch die Schätzung $\widehat{\sigma^2}$ ersetzen, ist der resultierende Ausdruck F-verteilt mit $(q, n-p)$ Freiheitsgraden und es gilt

$$P\left((\mathbf{Y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}))^T\mathbf{D^{-1}}(\mathbf{Y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM})) \leq qF_{q,n-p,1-\alpha}S^*\right) = 1 - \alpha,$$

mit

$$S^*(\mathbf{x}) = \widehat{\sigma^2} = \frac{1}{n-p}(\mathbf{x} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{GM})^T\mathbf{V_{xx}^{-1}}(\mathbf{x} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{GM}).$$

Hieraus ergibt sich der folgende Vorhersagebereich.

Satz 4.1 *Im Fall* $Rang(\mathbf{F}_0) = q$ *ist ein* $(1-\alpha)-$ *Vorhersageellipsoid für* $\mathbf{Y}$ *gegeben durch:*

$$\mathcal{I}(\mathbf{x}) = \left\{ \mathbf{y} \in \mathbb{R}^q : \frac{1}{q}\frac{(\mathbf{y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}))^T\mathbf{D}^{-1}(\mathbf{y} - \mathbf{Y}(\widehat{\boldsymbol{\beta}}_{GM}))}{S^*(\mathbf{x})} \leq F_{q,n-p,1-\alpha} \right\}. \quad (4.10)$$

(ii) Seien $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ stochastisch unabhängig und identisch $N_p(\boldsymbol{\beta}, \mathbf{V})$-verteilte $\mathbb{R}^p$-wertige Zufallsvariablen, $(\boldsymbol{\beta}, \mathbf{V}) \in \mathbb{R}^p \times PD(p)$ ist der Parameter. Kurz

$$\left((\mathbb{R}^n)^p, \mathcal{B}^{n\times p}, (\otimes_1^n N(\boldsymbol{\beta}, \mathbf{V}))_{(\boldsymbol{\beta},\mathbf{V})\in\mathbb{R}^p\times PD(p)}\right). \tag{4.11}$$

Sei $\mathbf{X}_{n+1}$ ein weiterer Zufallsvektor aus dieser Verteilung, nicht beobachtbar, unabhängig von den ersten n. Wir fassen die Zufallsvariablen zu einer $\mathbb{R}^{n\times p}$-wertigen Zufallsvariablen zusammen:

$$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T.$$

Das Modell schreibt sich dann so:

$$\begin{pmatrix} Vec(\mathbf{X}) \\ \mathbf{X}_{n+1} \end{pmatrix} \sim N_{np+p}\left(\begin{bmatrix} \mathbf{1}_n \otimes \boldsymbol{\beta} \\ \boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} I_n \otimes \mathbf{V} & 0 \\ 0 & \mathbf{V} \end{bmatrix}\right).$$

Die optimale Prognose für eine neue (zukünftige) Beobachtung $\mathbf{X}_{n+1}$ ist gegeben durch:

$$\begin{aligned} \widehat{\mathbf{X}}_{n+1} &= \widehat{\boldsymbol{\beta}} \\ &= \mathbf{X}^T \frac{1}{n} \mathbf{1}_n \\ &= \bar{\mathbf{X}}. \end{aligned}$$

Neben der Punktschätzung der Prognose ist man in der Regel auch an einem Vorhersageintervall interessiert. Ein Vorhersageintervall für $\mathbf{X}_{n+1}$ lässt sich folgendermaßen konstruieren. Wegen

$$\begin{aligned} \mathbf{D} = Cov(\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1}) &= Cov(\mathbf{X}_{n+1}) + Cov(\widehat{\mathbf{X}}_{n+1}) - Cov(\mathbf{X}_{n+1}, \widehat{\mathbf{X}}_{n+1}) \\ &\quad - Cov(\widehat{\mathbf{X}}_{n+1}\mathbf{X}_{n+1}) \\ &= \mathbf{V} + \frac{\mathbf{V}}{n}. \end{aligned}$$

gilt

$$\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1} \sim N_p(0, \mathbf{D}).$$

Kleines Hilfsresultat: Seber(1984, S. 30)
Wenn $\mathbf{Y} \sim N_p(0, \mathbf{V})$, $\mathbf{A} \sim W_p(m, \mathbf{V})$ und $\mathbf{Y}, \mathbf{A}$ unabhängig sind, dann gilt

$$\frac{m-p+1}{p} \mathbf{Y^T A^{-1} Y} \sim F_{p,m-p+1} \tag{4.12}$$

Wir betrachten die Statistik:

$$\begin{aligned} \mathbf{S}(\mathbf{x}) &= \sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \\ &= \mathbf{x}^T\mathbf{x} - \overline{\mathbf{x}}\ \overline{\mathbf{x}}^T. \end{aligned}$$

Durch geeignetes Standardisieren erhalten wir:

$$\mathbf{Q} = \sqrt{\frac{n}{n+1}}\mathbf{D}^{-1/2}(\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1}) \sim N_p(0, I_p)$$

$$\mathbf{B} = \mathbf{V}^{-1/2}\mathbf{S}\mathbf{V}^{-1/2} \sim W_p(n-1, I_p).$$

Im Hinblick, dass $\mathbf{B}$ und $\mathbf{Q}$ unabhängig sind, und unter Verwendung der Darstellung (4.12) erhalten wir

$$\frac{n-p}{p}\mathbf{Q}^T\mathbf{B}^{-1}\mathbf{Q} \sim F_{p,n-p}.$$

Wenn wir $\mathbf{Q}$ durch $\sqrt{\frac{n}{n+1}}\mathbf{D}^{-1/2}(\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1})$ ersetzen ist der resultierende Ausdruck

$$\frac{(n-p)n}{(n+1)p}(\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1})^T\mathbf{S}^{-1}(\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1}) \sim F_{p,n-p}$$

und es gilt

$$P\left(\frac{(n-p)n}{(n+1)p}(\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1})^T\mathbf{S}^{-1}(\mathbf{X}_{n+1} - \widehat{\mathbf{X}}_{n+1}) \leq F_{p,n-p,1-\alpha}\right) = 1-\alpha.$$

Somit erhält man

$$\mathcal{I}(\mathbf{x}) = \left\{\mathbf{x}_{n+1} \in \mathbb{R}^p : \frac{(n-p)n}{(n+1)p}(\mathbf{x}_{n+1} - \overline{\mathbf{x}})^T\mathbf{S}^{-1}(\mathbf{x})(\mathbf{x}_{n+1} - \overline{\mathbf{x}}) \leq F_{p,n-p,1-\alpha}(1-\alpha)\right\}, \tag{4.13}$$

als Vorhersageintervall für $\mathbf{X}_{n+1}$ zum Niveau $1-\alpha$.

4.3. Weitere simultane Vorhersageintervalle unter Normalverteilungsannahme

Dieser Abschnitt stellt zunächst Vorhersageintervalle im ein- und mehrdimensionalen Fall vor und behandelt danach das simultane Vorhersageintervall nach den Ansätzen von Bonferroni und Scheffé. Für Beweise und weitere Eigenschaften verweisen wir auf die Literatur, z.B. Hocking (2005, Kap. 18.2) und Christensen(2011).
Wir betrachten das Modell 1.3.b mit $\mathbb{R}^{p+q}$-wertigen, normal-verteilten Zufallsvariablen,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \boldsymbol{\mu}_\mathbf{x} \\ \boldsymbol{\mu}_\mathbf{y} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_{\mathbf{xx}} & \mathbf{V}_{\mathbf{xy}} \\ \mathbf{V}_{\mathbf{yx}} & \mathbf{V}_{\mathbf{yy}} \end{pmatrix}\right),$$

wobei

- $\mathbf{X} \sim N_p(\boldsymbol{\mu}_\mathbf{x}, \mathbf{V}_{\mathbf{xx}})$ (beobachtbar)
- $\mathbf{Y}$ eine $\mathbb{R}^q$-wertige Zufallsvariable (nicht beobachtbar)

Als BLUP-Vorhersage $\widehat{\mathbf{Y}}$ für $\mathbf{Y}$ ergibt sich

$$\begin{aligned}\widehat{\mathbf{Y}} &= \arg\min_{g\in SP} E[\mathbf{Y}-g(\mathbf{X})]^2 \\ &= E(\mathbf{Y}|\mathbf{X}) \\ &= \boldsymbol{\mu}_{\mathbf{y}} + \mathbf{V}_{\mathbf{yx}}\mathbf{V}_{\mathbf{xx}}^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{x}}) \text{ (nach 1.12)}\end{aligned}$$

Verteilung der Residuen $\mathbf{Y}-\widehat{\mathbf{Y}}$

Aus den Verteilungsresultaten für (mehrdimensionale) normalverteilte Zufallsvariablen erhalten wir, dass

1. $\mathbf{Y}-\widehat{\mathbf{Y}} \sim N_q(0,\mathbf{C})$ das heißt $\mathbf{C}^{-1/2}(\widehat{\mathbf{Y}}-\mathbf{Y}) \sim N_q(0,I_q)$, wobei

$$\mathbf{C} = Cov(\mathbf{Y}-\widehat{\mathbf{Y}}) = \mathbf{V}_{\mathbf{yy}} - \mathbf{V}_{\mathbf{yx}}\mathbf{V}_{\mathbf{xx}}^{-\mathbf{1}}\mathbf{V}_{\mathbf{xy}} \tag{4.14}$$

2. Hieraus ergibt sich, dass

$$(\mathbf{Y}-\widehat{\mathbf{Y}})^T\mathbf{C}^{-1}(\mathbf{Y}-\widehat{\mathbf{Y}}) \sim \chi^2_q$$

3. $Y_i-\widehat{Y}_i \sim N(0,C_{ii}) \Rightarrow \frac{Y_i-\widehat{Y}_i}{\sqrt{C_{ii}}} \sim N(0,1)$ mit $C_{ii}=\mathbf{e}_i^T\mathbf{C}\mathbf{e}_i$ und $\mathbf{e}_i$ ist $i-ter$ Einheitsvektor.

Prognoseintervall für einzelne Koeffizienten Y_i

- Ein $(1-\alpha)$ Vorhersageintervall für Y_i ist gegeben durch:

$$\mathcal{I}^{(i)}(\mathbf{x}) = [\widehat{y}_i(\mathbf{x}) - \sqrt{C_{ii}}z_{1-\frac{\alpha}{2}}, \widehat{y}_i(\mathbf{x}) + \sqrt{C_{ii}}z_{1-\frac{\alpha}{2}}],$$

  wobei $z_{1-\frac{\alpha}{2}} = \Phi^{-1}(1-\frac{\alpha}{2})$ das $(1-\frac{\alpha}{2})$ - Quantil von $N(0,1)$ ist.

- Die Länge des zufälligen Intervalls ist bekannt (hängt nicht von $\widehat{Y}_i$ ab) gleich

$$2\sqrt{C_{ii}}z_{1-\frac{\alpha}{2}}$$

Simultanes Prognoseintervall (Prognosebereich) für $\mathbf{Y}$:
Hocking(2005, Kap.18.1.1.2)

Oft interessieren nicht nur Vorhersageintervalle für eine einzelne Komponente Y_i, sondern auch Intervalle für mehrere Kombinationen $\mathbf{e}_i^T\mathbf{Y}$, $i=1,\ldots,r$, z.B. alle Komponenten $Y_i, i=1,\ldots,q$. Ein $(1-\alpha)$ Vorhersagebereich $\mathcal{I}$ für $\mathbf{Y}$ heißt ein $(1-\alpha)$ simultanes Vorhersageintervall für $Y_i,\ i=1,\ldots,q$, wenn $\mathcal{I}$ von der folgenden Form ist:

$$\mathcal{I}(\mathbf{x}) = \times_{i=1}^{q}\mathcal{I}^{(i)}(\mathbf{x}), \quad \forall \mathbf{x}\in\mathbb{R}^p,$$

mit Intervall $\mathcal{I}^{(i)}(\mathbf{x}) \subset \mathbb{R}$ für alle $i = 1, \ldots, q$ und $\mathbf{x} \in \mathbb{R}^p$, d.h. $\mathcal{I}(\mathbf{x})$ ein achsenparalleler Quader in $\mathbb{R}^q$ für jedes $\mathbf{x}$.
Ein simultanes $(1-\alpha)$ Vorhersageintervall für $\mathbf{Y}$ nach der Bonferroni-Methode ist gegeben durch:

$$\mathcal{I}_{Bonf}(\mathbf{x}) = \times_{i=1}^{q} \big[\mathbf{e}_i^T \widehat{\mathbf{y}}(\mathbf{x}) - \sqrt{C_{ii}} z_{1-\frac{\alpha}{2q}}, \mathbf{e}_i^T \widehat{\mathbf{y}}(\mathbf{x}) + \sqrt{C_{ii}} z_{1-\frac{\alpha}{2q}}\big], \tag{4.15}$$

wobei $z_{1-\frac{\alpha}{2q}}$ das $(1-\frac{\alpha}{2q})$-Quantil von $N(0,1)$ ist.
Ein anderes$(1-\alpha)$ simultanes Vorhersageintervall für die $\mathbb{R}^q$-wertige Zufallsvariable $\mathbf{Y}$ ist das Vorhersageellipsoid nach Scheffé.

Theorem 4.1 ***(Vorhersageellipsoid nach Scheffé)***
Ein $(1-\alpha)$ Vorhersageellipsoid für $\mathbf{Y}$ ist gegeben durch:

$$\mathcal{I}_{Sch}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^q : (\mathbf{y} - \widehat{\mathbf{y}}(\mathbf{x}))^T \mathbf{C}^{-1} (\mathbf{y} - \widehat{\mathbf{y}}(\mathbf{x})) \leq \chi^2_{q,1-\alpha}\}, \tag{4.16}$$

wobei $\mathcal{I}(\mathbf{x})$ ein Ellipsoid in $\mathbb{R}^q$ mit dem Mittelpunkt $\widehat{\mathbf{y}}(\mathbf{x})$ ist.

Ein ellipsenförmiger Bereich als Vorhersageintervall für die $\mathbb{R}^q$-wertige Vorhersage ist in den Anwendungen nicht so gut interpretierbar wie ein Achsen-paralleler Quader.

Herleitung von marginalen Vorhersageintervallen nach Scheffé:

Lemma 4.4 ***(andere Interpretation des Vorhersageellipsoids)***
Für gegebenes $\alpha \in (0,1)$ gilt:

$$P\Big(|Y_i - \widehat{Y}_i| \leq \sqrt{C_{ii}\chi^2_{q,1-\alpha}}\ \Big) = 1 - \alpha$$

Beweisidee:
Theorem 4.1 sagt, dass

$$P\Big(\big(\mathbf{Y} - \widehat{\mathbf{Y}}\big)^T \mathbf{C}^{-1} \big(\mathbf{Y} - \widehat{\mathbf{Y}}\big) \leq \chi^2_{q,1-\alpha}\Big) = 1 - \alpha. \tag{4.17}$$

Mit A.4 können wir schreiben:

$$\big(\mathbf{y} - \widehat{\mathbf{y}}\big)^T \mathbf{C}^{-1} \big(\mathbf{y} - \widehat{\mathbf{y}}\big) = \max_{\mathbf{a} \in R^q} \frac{\Big(\mathbf{a}^T \big(\mathbf{y} - \widehat{\mathbf{y}}\big)\Big)^2}{\mathbf{a}^T \mathbf{C} \mathbf{a}} \tag{4.18}$$

Folgerung 4.2 ***(marginales Vorhersageintervall)*** *Hocking (2005, Kap.18.1.1.3)*
Das marginale $(1-\alpha)$-Vorhersageintervall für $\mathbf{Y}$ nach Scheffé (Scheffé Rechteck Abk.: Schr) ist gegeben durch:

$$\mathcal{I}_{Schr}(\mathbf{x}) = \times_{i=1}^{q} \Big[\mathbf{e}_i^T \widehat{\mathbf{y}}(\mathbf{x}) - \sqrt{C_{ii}\chi^2_{q,1-\alpha}},\ \mathbf{e}_i^T \widehat{\mathbf{y}}(\mathbf{x}) + \sqrt{C_{ii}\chi^2_{q,1-\alpha}}\ \Big]. \tag{4.19}$$

wobei $\chi^2_{q,1-\alpha}$ das $(1-\alpha)$-Quantil der χ^2_q-Verteilung ist.

Beachte: Die Intervall-Längen der Bonferroni-Methode und der Scheffé-Methode sind unterschiedlich. Im vorliegenden Fall läuft es auf den Vergleich der Quantile

$$z_{1-\frac{\alpha}{2q}} \qquad \text{und} \qquad \sqrt{\chi^2_{q,1-\alpha}}$$

hinaus. Zum Beispiel:

	$\alpha = .01$			$\alpha = .05$	
q	$z_{1-\frac{\alpha}{2q}}$	$\sqrt{\chi^2_{q,1-\alpha}}$	q	$z_{1-\frac{\alpha}{2q}}$	$\sqrt{\chi^2_{q,1-\alpha}}$
2	2.807034	3.034854	2	2.241403	2.447747
3	2.935199	3.368214	3	2.39398	2.795483
4	3.023341	3.643721	4	2.497705	3.080216
5	3.090232	3.884105	5	2.575829	3.327236
10	3.290527	4.817598	10	2.807034	4.278672
30	3.587915	7.133876	30	3.14398	6.616115

In den obigen Beispielen gilt:

$$I_{Bonf}(\mathbf{x}) \subseteq I_{Schr}(\mathbf{x}), \ \forall \mathbf{x} \in \mathbb{R}^p$$

Daher liegt die Vermutung nahe, dass die Bonferroni-Methode besser ist als die Scheffé-Methode. Es bleibt Ansatzpunkt für weitere Untersuchungen. Eine Fragestellung für die Zukunft besteht darin, ob es einen Fall gibt, so dass $I_{Schr}(\mathbf{x}) \subseteq I_{Bonf}(\mathbf{x})$ ist.

5. Partielle kleinste Quadrate

In vielen Anwendungsfeldern der Statistik wie Medizin oder Ökonomie wird eine große Anzahl von Variablen erhoben, auch weil man nicht von vornherein weiß, welches die eigentlich wichtigen sind. Das macht die Entdeckung von Strukturen in den Daten schwierig, zumal die Variablen oft stark untereinander korrelieren, mithin gleichartige Information enthalten, ist es selten ratsam, einfach Variablen wegzulassen. Als Alternative können wir geeignete Linearkombinationen der Variablen suchen, die in geringerer Zahl möglichst alle relevanten Informationen des Datensatzes enthalten.
Für die Hauptkomponentenanalyse als datenanalytische Technik und Dimensionsreduktion kann man sagen, dass die Hauptkomponentenanalyse eine orthogonale Transformation im m-dimensionalen Raum der Originalvariablen in eine neue Variablenmenge ist. Als Dimensionsreduktion hat die Hauptkomponentenanalyse das Ziel, eine möglichst geringe Anzahl von Linearkombinationen der beobachteten Variablen zu finden, welche die Daten, möglichst gut repräsentieren. In manchen Anwendungen sind die resultierenden, als Hauptkomponenten bezeichneten Variablen die Größen, die von Interesse sind.
Die Hauptkomponentenanalyse besitzt jedoch einige Nachteile. Die Ergebnisse sind abhängig von der Skalierung und daher nicht eindeutig. Die Hauptkomponenten sind schwierig zu interpretieren. Die Interpretation ist zumeist subjektiv. Auch die Anzahl der auszuwählenden wesentlichen Hauptkomponenten ist nicht eindeutig bestimmt.
Eine weitere Möglichkeit zur Dimensionsreduktion bietet die sogenannte Methode der partiellen kleinsten Quadrate(engl. partial least squares, PLS), die ursprünglich von Wold (1966) eingeführt wurde. Die PLS-Methode hat gewisse Ähnlichkeiten mit der Hauptkomponentenanalyse PCA (engl. Principal component analysis). Bei der PCA-Methode werden ebenfalls zueinander orthogonale neue Variablen als Linearkombination aus den ursprünglichen Variablen bestimmt. Im Unterschied zur PLS werden bei der PCA (als Ansatz für Dimensionsreduktion und Vorhersage) nicht die Kovarianz zwischen Y und $\mathbf{a^T X}$, sondern nur die Varianz von $\mathbf{a^T X}$ maximiert (siehe Puntanen und Styan und Isotalo (2013, Kap.9.4)). Sie wird in verschiedenen Anwendungsgebieten zur Regression oder Diskriminanzanalyse und Vorhersage eingesetzt, insbesondere für die Analyse hochdimensionaler Daten. Besonders verbreitet ist die Anwendung der PLS-Methode in der Chemometrie zur spektrometrischen Kalibrierung. Prädiktoren sind dabei die Emissionsintensitäten zu den verschiedenen Frequenzen eines gewissen Spektrums, abhängige Variablen sind die Anteile von bestimmten Stoffen in einem Stoffgemisch (vgl.Grüning (2005, Kap.4.2.1)).
Im Folgenden werden wir zunächst die theoretischen Grundlagen der Methode der Partiellen kleinsten Quadrate studieren, um das mathematische Modell und die ursprünglichen Ideen der Methode aufzubauen, die zur Entwicklung der PLS-Methode führen, um einige wichtige Eigenschaften der PLS-Methode herleiten zu können. Wir stellen die

Methoden der Hauptkomponentenanalyse und die Methode der kanonischen Korrelation als Minimierung des mittleren quadratischen Vorhersagefehlers vor. Zum Schluss betrachten wir Beziehungen zwischen den Methoden der Regression, Hauptkomponentenanalyse, kanonischen Korrelation und Partiellen kleinsten Quadrate.
Seien X und Y quadratintegrierbare reelle Zufallsvariablen, dann betrachten wir als Motivation das Optimierungsproblem:

$$\min_{a,b\in\mathbb{R}} E[(Y-b-aX)^2]. \tag{5.1}$$

Dieses besitzt die Lösung:

$$a^* = \frac{Cov(X,Y)}{Var(X)},\ b^* = E(Y) - a^*E(X). \tag{5.2}$$

Es gilt $Cov(a^*X,Y) = Var(a^*X) = \frac{[Cov(X,Y)]^2}{Var(X)}$. Der Minimalwert M^* in (5.1) heißt Residualvarianz und ergibt sich zu :

$$\begin{aligned} M^* &= E[(Y-a^*X-b^*)^2] \\ &= Var(Y-a^*X). \\ &= Var(Y) - Cov(a^*X,Y). \\ &= Var(Y)(1-Corr^2(X,Y)), \end{aligned}$$

5.1. Motivation

Wir betrachten in diesem Abschnitt das Modell 1.3.b mit $q = 1$, und treffen für das Modell folgende zusätzliche Annahmen

1. Erwartungswert ist null $E(\mathbf{X}) = 0$
2. Y ist nicht beobachtbar.

Kurz:

$$\begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \sim N_{p+1}(\mathbf{0}, \mathbf{V}),\ \text{mit}$$
$$\mathbf{V} = \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V}_{\mathbf{x}y} \\ \mathbf{V}_{y\mathbf{x}} & V_{yy} \end{pmatrix} \in PD(p+1)$$
Y nicht beobachtbar

Wir betrachten das folgende Problem der Prognose: basierend auf einem Beobachtungswert der $\mathbb{R}^p$-wertigen Zufallsvariable $\mathbf{X}$, möchte man eine beste lineare Vorhersage von Y möglichst im Sinne der Minimierung des mittleren quadratischen Fehlers herleiten. Also ist X eine beobachtete Zufallsvariable, die die Variable Y gut vorhersagen soll. Mit anderen Worten hat unsere Aufgabe häufig das Ziel, eine lineare Transformation $\mathbf{a}^T\mathbf{X}$, mit $\mathbf{a}^T\mathbf{a} = 1$ so bestimmt werden, dass $E[(Y-\mathbf{a}^T\mathbf{X})^2]$ minimal wird.

(I) Aus den Eigenschaften der bedingten Erwartung folgt:

$$\begin{aligned} \min_{g \in SP} E[(Y - g(\mathbf{X}))^2] &= E[(Y - E(Y|\mathbf{X}))^2] && (5.3) \\ &= Var(Y - \mathbf{V_{yx}}\mathbf{V_{xx}^{-1}}\mathbf{X}). && (5.4) \end{aligned}$$

Aus der Definition der bedingten Erwartung der Normalverteilung (1.13) bemerken wir, dass die optimale Lösung von $\min_{g \in SP} Var(Y - g(\mathbf{X}))$ mit der optimalen Lösung von $\min_{\text{l} \in \text{LP}} Var(Y - l(\mathbf{X}))$ unter Normalverteilgungsannahme übereinstimmt.

(II) Aus den Verteilungsresultaten für (mehrdimensionale) normalverteilte Zufallsvariablen wissen wir, dass

$$\begin{pmatrix} \mathbf{a}^T\mathbf{X} \\ Y \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} & \mathbf{a}^T\mathbf{V_{x}}y \\ \mathbf{V}_{y\mathbf{x}}\mathbf{a} & V_{yy} \end{pmatrix} \right)$$

ist. (siehe Fujikoshi, Ulyanov und Shimizu (2010, Aufgabe 11.7))

- Die beste lineare erwartungstreue Vorhersage für Y unter $\mathbf{a}^T\mathbf{X}$ ist gegeben durch:

$$\begin{aligned} \widehat{Y}_{\mathbf{a}} &= E(Y|\mathbf{a}^T\mathbf{X}) && (5.5) \\ &= Cov(Y, \mathbf{a}^T\mathbf{X})[Var(\mathbf{a}^T\mathbf{X})]^{-1}\mathbf{a}^T\mathbf{X}(\text{ nach (1.13)}) && (5.6) \\ &= \frac{\mathbf{V}_{y\mathbf{x}}\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{V_{xx}}\mathbf{a}}\mathbf{X} && (5.7) \end{aligned}$$

- Der mittlere quadratische Vorhersagefehler von $\widehat{Y}_{\mathbf{a}}$ ist gegeben durch:

$$\begin{aligned} R(\mathbf{a}) &= E(Y - \widehat{Y}_{\mathbf{a}})^2 \\ &= Var(Y - \widehat{Y}_{\mathbf{a}}) \\ &= Var(Y) - Var(\widehat{Y}_{\mathbf{a}}) \\ &= V_{yy} - \frac{\mathbf{a}^T\mathbf{V_{x}}y\mathbf{V_{x}}y^T\mathbf{a}}{\mathbf{a}^T\mathbf{V_{xx}}\mathbf{a}} \qquad \text{(nach Folgerung 3.5)} \end{aligned}$$

Gesucht wird die Linearkombination

$$\mathbf{a}^T\mathbf{X} = a_1X_1 + \cdots + a_pX_p, \mathbf{a} \neq 0$$

für die $R(\mathbf{a})$ minimum ist. Minimierung von $R(\mathbf{a})$ ist äquivalent zu Maximierung von $\frac{\mathbf{a}^T\mathbf{V_{x}}y\mathbf{V}_{y\mathbf{x}}\mathbf{a}}{\mathbf{a}^T\mathbf{V_{xx}}\mathbf{a}}$, und die optimale Lösung ist :

$$\mathbf{a}^* = e.v_{max}(\mathbf{V_{xx}^{-1}}\mathbf{V_{x}}y\mathbf{V_{x}}y^T) \qquad \text{(nach Satz A.3)}, \tag{5.8}$$

wobei $\mathbf{a}^*$ der Eigenvektor ist, der zum $\lambda_{\max}$ der Matrix $\mathbf{V_{xx}^{-1}}\mathbf{V_{x}}y\mathbf{V_{x}}y^T$ gehört. Nach Satz A.4 erhalten wir folgende Lösung:

$$\mathbf{a}^* = \frac{\mathbf{V_{xx}^{-1}}\mathbf{V_{x}}y}{\mathbf{V_{x}}y^T\mathbf{V_{xx}^{-1}}\mathbf{V_{x}}y}$$

- Es gilt also:

$$\begin{aligned} R(\mathbf{a}) &= V_{yy} - \frac{\mathbf{V_{yx}aa}^T\mathbf{V_{xy}}}{\mathbf{a}^T\mathbf{V_{xx}a}} \\ &= Var(Y) - Cov(Y, \widehat{Y}_{\mathbf{a}}) \\ &= Var(Y) - Var(\widehat{Y}_{\mathbf{a}}) \\ &= Var(Y)(1 - Corr^2(\widehat{Y}_{\mathbf{a}}, Y)). \end{aligned}$$

Beachte

- Hieraus ergibt sich, dass

$$\begin{aligned} \min_{\mathbf{a}} R(\mathbf{a}) &= \min_{\mathbf{a}} Cov(Y - \widehat{Y}_{\mathbf{a}}, Y) \\ &= \min_{\mathbf{a}} [Var(Y) - Var(\widehat{Y}_{\mathbf{a}})] \\ &= \min_{\mathbf{a}} Var(Y)(1 - Corr^2(\widehat{Y}_{\mathbf{a}}, Y)). \end{aligned}$$

  sind.

- Eine äquivalente Darstellung lautet:

$$\begin{aligned} \min_{\mathbf{a}\in R^p} R(\mathbf{a}) &\Leftrightarrow \max_{\mathbf{a}} Cov(\widehat{Y}_{\mathbf{a}}, Y) && (5.9) \\ &\Leftrightarrow \max_{\mathbf{a}} Var(\widehat{Y}_{\mathbf{a}}) && (5.10) \\ &\Leftrightarrow \max_{\mathbf{a}} Corr^2(\widehat{Y}_{\mathbf{a}}, Y). && (5.11) \end{aligned}$$

- Hieraus ergibt sich

$$\begin{aligned} \min_{\mathbf{a}\in R^p} Var(Y - \mathbf{a^T X}) &= \min_{\mathbf{a}\in R^p}[\min_{h\in SP_1} Var(Y - h(\mathbf{a^T X})] && (5.12) \\ &= \min_{\mathbf{a}\in R^p} Var(Y - E(Y|\mathbf{a}^T\mathbf{X})), && (5.13) \end{aligned}$$

  wobei $SP_1 = \{g|g : \mathbb{R} \longrightarrow \mathbb{R}, g \text{ messbar}, E(g(Z)^2) < \infty\}$ ist. In diesem Fall ist dazu äquivalent

$$\begin{aligned} \min_{\mathbf{a}\in R^p} Var(Y - \mathbf{a^T X}) &= \min_{\mathbf{a}\in R^p}[\min_{h\in L_1} Var(Y - h(\mathbf{a^T X})] && (5.14) \\ &= \min_{\mathbf{a}\in R^p} Var(Y - BLUP(Y, \mathbf{a^T X})), && (5.15) \end{aligned}$$

$$L_1 = \{g|g : \mathbb{R} \longrightarrow \mathbb{R}, g \text{ messbar}, g(z) = cz, c \in \mathbb{R}\}.$$

Dieses Optimierungsproblem für $R(\mathbf{a})$ werden wir in diesem Kapitel und in den anschließenden Aufgaben vertiefen.

5.2. Modellspezifikation und Problemstellung

Wir betrachten das Modell 1.3.a mit $q = 1$, unter der zusätzlichen Annahme, dass Y nicht beobachtbar ist. Kurz:

$$\begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \text{ eine } \mathbb{R}^{p+1} \text{ wertige ZV, mit}$$
$$\boldsymbol{\mu} \in \mathbb{R}^{p+1}, \mathbf{V} \in PD(p+1)$$
$$Y \text{ nicht beobachtbar.}$$

Wir hatten im Kapitel 3 die Aufgabe, basierend auf den Beobachtungswerten $x_1, \ldots, x_p$ der $\mathbb{R}^p$-wertigen Zufallsvariable $\mathbf{X}$ den Wert der $\mathbb{R}$-wertigen Zafallsvariable Y (im Sinne des mittleren quadratischen Fehlers) zu prognostizieren, untersucht.

$$\min_{g \in SP} E[(Y - g(\mathbf{X})^2]. \tag{5.16}$$

Die optimale Lösung von (5.16) war $m(\mathbf{X}) = E(Y|\mathbf{X})$. Zur Bestimmung von $E(Y|\mathbf{X})$ benötigen wir die gemeinsame Verteilung von Y und $\mathbf{X}$. Außerdem ist $E(Y|\mathbf{X})$ nicht einfach zu bestimmen. Aus diesem Grund beschränken wir SP auf die Klasse LP.
Nachteile:
Da $LP \subseteq SP$ gilt, dass

$$\min_{g \in SP} E[(Y - g(\mathbf{X}))^2] \leq \min_{l \in LP} E[(Y - l(\mathbf{X}))^2]$$

$$\max_{g \in SP} Corr^2(Y, g(\mathbf{X})) \geq \max_{l \in LP} Corr^2(Y, l(\mathbf{X}))$$

Vorteile:

1. Die Berechnung von $\min_{l \in LP} E[(Y - l(\mathbf{X}))^2]$ ist einfach.

2. $BLUP(Y, \mathbf{X})$ hängt nur von den ersten und zweiten Momenten ab

3. Unter Normalverteilungsannahme gilt:

$$\min_{g \in SP} E[(Y - g(\mathbf{X}))^2] = \min_{l \in LP} E[(Y - l(\mathbf{X}))^2].$$

Die Ausdrücke (5.9), (5.10) und (5.11) führen zu den folgenden Begriffsbildungen:

(i) Vorhersage durch Maximierung der Korrelation:
Es wurde bereits im Kapitel 3 gezeigt, dass

$$\arg\max_{g \in SP} Corr^2(Y, g(\mathbf{X})) = E(Y|\mathbf{X})$$
$$\arg\max_{\mathbf{a} \in \mathbb{R}^p} Corr^2(Y, \mathbf{a}^T\mathbf{X}) = \mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{V}_{\mathbf{x}y}.$$

(ii) Vorhersage durch Maximierung der Varianz oder der Kovarianz
Dazu betrachtet man

$$Var(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} \text{ und } \qquad Cov(\mathbf{a}^T\mathbf{X}, Y) = \mathbf{a}^T\mathbf{V_{x}}_y. \tag{5.17}$$

Diese Vorhersagen sind bestimmt durch Maximierung der Varianz der Linearkombination $\mathbf{a}^T\mathbf{X}$ oder durch Maximierung der Kovarianz zwischen $\mathbf{a}^T\mathbf{X}$ und Y. Die Maximierungsprobleme sind nicht eindeutig lösbar auf $\mathbf{a} \in \mathbb{R}^p$. Der Vektor $\mathbf{a}$ ist damit nur bis auf eine Konstante eindeutig bestimmt und eine zusätzliche Normierungsbedingung ist notwendig um Eindeutigkeit zu erreichen, wie zum Beispiel in der kanonischen Korrelation $Var(\mathbf{a}^T\mathbf{X}) = \mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} = 1$. Die Lösung von

$$\begin{aligned} &\max_{\mathbf{a}} Cov(\mathbf{a}^T\mathbf{X}, Y) \\ &\text{unter der Nebenbedingung } \mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} = 1, \end{aligned}$$

ist $\sqrt{\mathbf{V}_{\mathbf{x}y}^T\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{V}_{\mathbf{x}y}}$. Anstatt die Bedingung $\mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} = 1$ zu fordern, wählen wir in diesem Kapitel die Nebenbedingung $\mathbf{a}^T\mathbf{a} = 1$ (in der Praxis wie z.B. Data Mining). Die Maximierungsprobleme stellen sich damit in der Form

$$\begin{aligned} &\max_{\mathbf{a}} Var(\mathbf{a}^T\mathbf{X}) \text{ oder } \max_{a} Cov(\mathbf{a}^T\mathbf{X}, Y) \\ &\text{unter der Nebenbedingung } \mathbf{a}^T\mathbf{a} = 1. \end{aligned}$$

In diesem Kapitel und in den anschließenden Aufgaben werden wir das Optimierungsproblem

$$\min_{g \in LP_1} E(Y - g(\mathbf{X}))^2 \tag{5.18}$$

vertiefen, wobei die Menge

$$LP_1 = \{g | g : \mathbb{R}^p \to \mathbb{R}, g(\mathbf{x}) = \mu_y + \mathbf{a}^T(\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}),\ \mathbf{a} \in \mathbb{R}^p \text{ mit } \mathbf{a}^T\mathbf{a} = 1\} \ \subseteq LP$$

kompakt ist. Durch die Maximierung von Kovarianz-Varianz lassen sich verschiedene Vorhersagen für Y unter $\mathbf{X}$ in der Klasse LP_1 bestimmen. Ziel ist es, die Qualität einer Vorhersage für Y zu messen, wenn der Prognosefehler $Y - \widehat{Y}$ nicht orthogonal zu allen in die Prognosefunktion eingehenden Variablen $X_1, \ldots, X_p$ steht.

5.3. Die Vorhersage auf LP_1

5.3.1. Vorhersage durch Maximierung der Korrelation

Wir suchen einen Koeffizientenvektor $\mathbf{a} \in \mathbb{R}^p \setminus \{0\}$ mit $\mathbf{a}^T\mathbf{a} = 1$, wobei die Korrelation zwischen Y und $\mathbf{a^T X}$ maximal ist.

$$\begin{aligned}
\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Corr(Y, \mathbf{a}^T\mathbf{X}) &= \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Corr(Y, \mathbf{a}^T\mathbf{V}_{\mathbf{xx}}^{-1/2}\mathbf{X}), \quad \text{mit } \mathbf{V}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\mathbf{xx}}^{-1/2} = \mathbf{V}_{\mathbf{xx}}^{-1} \\
&\Leftrightarrow \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Cov(\mathbf{a}^T\mathbf{V}_{\mathbf{xx}}^{-1/2}\mathbf{X}, Y) \\
&\Leftrightarrow \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} \mathbf{a}^T\mathbf{V}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\mathbf{x}y} \\
&= \sqrt{\mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{V}_{y\mathbf{x}}^T} \qquad (\text{ Satz A.1}),
\end{aligned}$$

und das Maximum wird angenommen für:

$$\arg\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Corr(Y, \mathbf{a}^T\mathbf{X}) = \frac{\mathbf{V}_{\mathbf{xx}}^{-1/2}\mathbf{V}_{\mathbf{x}y}}{\sqrt{\mathbf{V}_{y\mathbf{x}}\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{V}_{yx}^T}}.$$

Also erhalten wir wieder die Regressionsfunktion als Lösung des Maximierungsproblems auf LP_1.

5.3.2. Vorhersage durch Maximierung der Varianz (PCA-Methode)

In diesem Abschnitt werden wir eine Methode zur Vorhersage vorstellen„ die als Hauptkomponentenanalyse bekannt ist, (engl. principal component analysis PCA). Sie ist eines der ältesten Ordinationsverfahren zur Dimensionsreduktion. Sie wurde in der ersten Hälfte des 20. Jahrhunderts im angloamerikanischen Raum entwickelt.

Definition 5.1 ***(PCA-Kombination, Koeffizientenvektor, PCA-Vorhersage erster Art)*** *Puntanen, Styan und Isotalo (2013, Kap.9.4)*

- *Die Linearkombination $\mathbf{a}_{pca}^T\mathbf{X}$ heißt PCA-Komponente von $\mathbf{X}$, falls sie die maximale Varianz besitzt.*

$$Var(\mathbf{a}_{pca}^T\mathbf{X}) = \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Var(\mathbf{a}^T\mathbf{X}). \tag{5.19}$$

- *Dabei heißt $\mathbf{a}_{pca}$ der Koeffizientenvektor (Richtungsvektor) der PCA-Komponente*

- *Die Vorhersage $g_{pca}(\mathbf{X}) = \mu_y + \mathbf{a}_{pca}^T(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{x}})$ heißt **PCA-Vorhersage erster Art** von Y unter $\mathbf{X}$ in der Klasse LP_1.*

Eigenschaften des PCA-Koeffizientenvektors

(i) **Satz 5.1** *Es gilt*

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}} Var(\mathbf{a}^T\mathbf{X}) = \lambda_{max}(\mathbf{V_{xx}}) \quad \textit{nach Satz A.2,}$$

und die optimale Lösung wird angenommen für $\mathbf{a}_{pca} = e.v_{max}(\mathbf{V_{xx}})$*, der zu dem größten Eigenwert* λ_{max} *der Matrix* $\mathbf{V_{xx}}$ *gehört.*

(ii) Mit Hilfe der Projektion können wir die Bestimmung des PCA-Koeffizientenvektors als Minimierungsproblem betrachten, wie der folgende Satz aussagt.

Satz 5.2 *Für* $\mathbf{a} \in \mathbb{R}^p$ *gilt:*

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Var(\mathbf{a}^T\mathbf{X}) \Longleftrightarrow \min_{\mathbf{a}} Spur(Cov[(\mathbf{X} - P_{\mathbf{a}}\mathbf{X})^T(\mathbf{X} - P_{\mathbf{a}}\mathbf{X})])$$
unter der Bedingung $\mathbf{a}^T\mathbf{a} = 1$,

wobei $P_{\mathbf{a}} = \mathbf{a}\mathbf{a}^T$ *die Projektionsmatrix ist.*

Beweis:
Man kann sich überlegen, dass die Projektion von $\mathbf{x}$ auf $\mathbf{a}$ gleich

$$P_{\mathbf{a}}\mathbf{x} = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}\mathbf{x} = \mathbf{a}\mathbf{a}^T\mathbf{x}$$

ist.

$$\begin{aligned}\min_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Spur(Cov[(\mathbf{X} - P_{\mathbf{a}}\mathbf{X})^T(\mathbf{X} - P_{\mathbf{a}}\mathbf{X})]) &= \min_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Spur(\mathbf{V}) - \mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} \\ &\Leftrightarrow \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Var(\mathbf{a}^T\mathbf{X})\end{aligned}$$

(iii) Unter Normalverteilgunsannahme mit Erwartungswert gleich 0 gilt

Satz 5.3 *Puntanen, Styan und Isotalo (2013, Kap.9.4)*
Für $\mathbf{a} \in \mathbb{R}^p$ *gilt*

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Var(\mathbf{a}^T\mathbf{X}) \Longleftrightarrow \min_{\mathbf{a}} MSEP(\mathbf{X}, E(\mathbf{X}|\mathbf{a}^T\mathbf{X}))$$
unter der Bedingung $\mathbf{a}^T\mathbf{a} = 1$.

Folgerung 5.1 *Für* $\mathbf{a} \in \mathbb{R}^p$ *gilt*

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Var(\mathbf{a}^T\mathbf{X}) \Longleftrightarrow \min_{\mathbf{a}} MSEP(\mathbf{X}, BLUP(\mathbf{X}, \mathbf{a}^T\mathbf{X}))$$
unter der Bedingungen $\mathbf{a}^T\mathbf{a} = 1$.

5.3.3. Vorhersage durch Maximierung der Kovarianz: (PLS-Methode)

Definition 5.2 ***(PLS-Kombination, PLS-Koeffizientenvektor, PLS-Vorhersage erster Art)*** *Liu und Rayens (2007) und Garthwaite (1994)*

- *Die Linearkombination* $\mathbf{a}_{pls}^T\mathbf{X}$ *heißt PLS-Komponente von* Y *unter* $\mathbf{X}$*, falls sie die maximale Kovarianz mit* Y *besitzt.*

$$Cov(Y, \mathbf{a}_{pls}^T\mathbf{X}) = \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Cov(Y, \mathbf{a}^T\mathbf{X}). \tag{5.20}$$

- *Dabei heißt* $\mathbf{a}_{pls}$ *der Koeffizientenvektor (Richtungsvektor) der PLS-Komponente*
- *Die Vorhersage* $g_{pls}(\mathbf{X}) = \mu_y + \mathbf{a}_{pls}^T(\mathbf{X} - \boldsymbol{\mu}_\mathbf{x})$ *heißt* ***PLS-Vorhersage erster Art*** *von* Y *unter* $\mathbf{X}$ *in der Klasse* LP_1*.*

Dann gilt der Satz:

Satz 5.4 *Die Optimallösung ist gegeben durch:*

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Cov(Y, \mathbf{a}^T\mathbf{X})) = \sqrt{\lambda_{max}(\mathbf{V}_{\mathbf{xy}}\mathbf{V}_{\mathbf{xy}}^T)} \quad \textit{(nach A1,A2)},$$

und sie wird angenommen für $\mathbf{a}_{pls} = e.v_{max}(\mathbf{V}_{\mathbf{xy}}\mathbf{V}_{\mathbf{xy}}^T)$*, der zu dem größten Eigenwert der Matrix* $\mathbf{V}_{\mathbf{xy}}\mathbf{V}_{\mathbf{xy}}^T$ *gehört.*

Bemerkung 5.1 *Da* $Rang(\mathbf{V}_{\mathbf{xy}}\mathbf{V}_{\mathbf{yx}}) = 1$ *ist, erhalten wir, dass*

$$\mathbf{a}_{pls} = c\mathbf{V}_{\mathbf{xy}} \textit{ mit } c = \parallel \mathbf{V}_{\mathbf{xy}}^T\mathbf{V}_{\mathbf{xy}} \parallel . \tag{5.21}$$

Außerdem sind die folgenden Eigenschaften des PLS-Koeffizientenvektors nützlich:

Eigenschaften des PLS-Koeffizientenvektors

(i) Frank und Friedmann (1993) haben die folgende Beziehung zwischen PLS und PCA-Koeffizient gefunden:

Theorem 5.1 *Frank und Friedmann (1993)*

$$\begin{aligned} \mathbf{a}_{pls} &= \arg\max_{\mathbf{a}:\mathbf{a}^\mathbf{T}\mathbf{a}=1} Corr^2(\mathbf{a}^T\mathbf{X}, Y)Var(\mathbf{a}^T\mathbf{X}) \\ &= \arg\max_{\mathbf{a}^T\mathbf{a}=1} Cov^2(\mathbf{a}^T\mathbf{X}, Y) \end{aligned}$$

(ii) Auf die gleiche Weise können wir beweisen, dass

Theorem 5.2

$$\mathbf{a}_{pls} = \arg\min_{\mathbf{a}^T\mathbf{a}=1} \left[Var(Y - \mathbf{a}^T\mathbf{X}) + Spur(Cov(\mathbf{X} - \mathbf{a}\mathbf{a}^T\mathbf{X}))\right] \tag{5.22}$$

Beweis: Man kann sich überlegen, dass

$$Var(Y - \mathbf{a}^T\mathbf{X}) = Var(Y) + \mathbf{a}^T\mathbf{V}_{\mathbf{xx}}\mathbf{a} - 2\ Cov(Y, \mathbf{a}^T\mathbf{X}) \tag{5.23}$$

und

$$\begin{aligned} Spur\big(Cov(\mathbf{X} - \mathbf{a}\mathbf{a}^T\mathbf{X})\big) &= Spur\big(Cov[(\mathbf{I} - \mathbf{a}\mathbf{a}^T)\mathbf{X}]\big) && (5.24)\\ &= Spur(\mathbf{V}_{\mathbf{xx}} - \mathbf{V}_{\mathbf{xx}}\mathbf{a}\mathbf{a}^T), \text{da} \quad \mathbf{a}^T\mathbf{a} = 1 && (5.25)\\ &= Spur(\mathbf{V}_{\mathbf{xx}}) - \mathbf{a}^T\mathbf{V}_{\mathbf{xx}}\mathbf{a} && (5.26) \end{aligned}$$

Hieraus folgt, dass

$$\arg\min_{\mathbf{a}^T\mathbf{a}=1} [Var(Y - \mathbf{a}^T\mathbf{X}) + Spur(Cov(\mathbf{X} - \mathbf{a}\mathbf{a}^T\mathbf{X}))] = \arg\max_{\mathbf{a}^T\mathbf{a}=1} Cov(\mathbf{a}^T\mathbf{X}, Y). \tag{5.27}$$

bzw.

$$\min_{\mathbf{a}^T\mathbf{a}=1} [Var(Y - \mathbf{a}^T\mathbf{X}) + Spur(Cov(\mathbf{X} - \mathbf{a}\mathbf{a}^T\mathbf{X}))] \Leftrightarrow \max_{\mathbf{a}^T\mathbf{a}=1} Cov(\mathbf{a}^T\mathbf{X}, Y). \tag{5.28}$$

5.4. Modifizierte partielle kleinste Quadrate

In diesem Abschnitt betrachten wir ein etwas allgemeineres Modell als das, was wir im Abschnitt 5.2 vorgestellt haben, nämlich das Modell 1.3.a mit $q > 1$, d.h.

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \text{ eine } \mathbb{R}^{p+q} \text{ wertige ZV, mit}$$
$$\boldsymbol{\mu} \in \mathbb{R}^{p+q}, \mathbf{V} \in PD(p+q)$$
$$\boldsymbol{\mu}, \mathbf{V} \text{ bekannt}$$
$$\mathbf{Y} \text{ nicht beobachtbar.}$$

Das Vorgehen der kanonischen Korrelation stellt eine Verallgemeinerung des Multiplen Korrelationskoeffizients dar. Eine Verallgemeinerung der PLS-Methode wurde von Liu und Rayens (2007) vorgeschlagen. Die Bestimmung des PLS-Koeffizientenvektors in diesem Modell ist gegeben durch das Maximierungsproblem:

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Cov(\mathbf{a}^T\mathbf{X}, \mathbf{Y})[Cov(\mathbf{a}^T\mathbf{X}, \mathbf{Y})]^T \text{ bzw.} \tag{5.29}$$

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} \mathbf{a}^T\mathbf{V}_{\mathbf{x}y}(\mathbf{a}^T\mathbf{V}_{\mathbf{x}y})^T \qquad \text{Liu und Rayens (2007)}$$

Im Fall $q = 1$ erhalten wir:

$$\max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Cov(Y, \mathbf{a}^T\mathbf{X}) \Leftrightarrow \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} [Cov(Y, \mathbf{a}^T\mathbf{X}))]^2.$$

Hieraus und basierend auf der Überlegung von Abschnitt 5.1 werden wir eine Verallgemeinerung der partiellen kleinsten Quadrate vorschlagen, die es so in der Literatur noch nicht gibt. Die Idee der modifizierten PLS-Methode liegt darin, dass wir statt des Optimierungsproblems (5.29) betrachten:

$$\max_{\mathbf{a}\in R^p\setminus\{0\}} Spur\big[Cov\big(BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X}),\mathbf{Y}\big)\big] \tag{5.30}$$

Beachte:
1) $BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X})$ ist $\mathbb{R}^q$-wertige Zufallsvariable und $Cov\big(BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X}),\mathbf{Y}\big) \in PD(q)$.
2) Die Lösung von (5.30) im Fall $q = 1$ ist gleich der BLUP

Definition 5.3 ***(PLS1-Kombination, PLS1-Koeffizientenvektor)***

- *Die Linearkombination* $\mathbf{a}_{pla1}^T\mathbf{X}$ *heißt PLS1-Komponente von* Y *unter* $\mathbf{X}$*, falls gilt*

$$\mathbf{a}_{pls1} = \arg\max_{\mathbf{a}\in R^p\setminus\{0\}} Spur[Cov(BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X}),\mathbf{Y})] \tag{5.31}$$

- *Dabei heißt* $\mathbf{a}_{pls1}$ *der Koeffizientenvektor (Richtungsvektor) der PLS1-Komponente*

Eigenschaften des PLS1-Koeffizientenvektors

(i) Unter der Annahme, dass $E(\mathbf{X}) = 0$ und $E(\mathbf{Y}) = 0$ sind, gilt:

$$\max_{\mathbf{a}\in\mathbb{R}^p\setminus\{0\}} Spur[Cov(BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X}),\mathbf{Y})] = \min_{\mathbf{a}\in\mathbb{R}^p\setminus\{0\}} MSEP(\mathbf{Y},BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X})).$$

Beweis:
Gemäß der Definition der besten erwartungstreuen Vorhersage erhalten wir:

$$BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X}) = \widehat{\mathbf{Y}}_\mathbf{a} = \frac{\mathbf{V_{yx}aa}^T}{\mathbf{a}^T\mathbf{V_{xx}a}}\mathbf{X} \qquad \text{und} \qquad Cov(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a}) = \frac{\mathbf{V_{yx}aa}^T\mathbf{V_{xy}}}{\mathbf{a}^T\mathbf{V_{xx}a}}.$$

Der mittlere quadratischen Fehler der Vorhersage im mehrdimensionalen Fall ist

$$\begin{aligned} MSEP(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a}) &= E((\mathbf{Y}-\widehat{\mathbf{Y}}_\mathbf{a})^T(\mathbf{Y}-\widehat{\mathbf{Y}}_\mathbf{a})) \\ &= Spur\ [Cov(\mathbf{Y}-\widehat{\mathbf{Y}}_\mathbf{a})] + [E(\mathbf{Y}-\widehat{\mathbf{Y}}_\mathbf{a})]^T[E(\mathbf{Y}-\widehat{\mathbf{Y}}_\mathbf{a})] \\ &= Spur[Cov(\mathbf{Y}) + Cov(\widehat{\mathbf{Y}}_\mathbf{a}) - Cov(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a}) - Cov(\widehat{\mathbf{Y}}_\mathbf{a},\mathbf{Y})] + 0 \\ &= Spur\Big[\mathbf{V_{yy}} - \frac{\mathbf{V_{yx}aa}^T\mathbf{V_{xy}}}{\mathbf{a}^T\mathbf{V_{xx}a}}\Big]. \end{aligned}$$

und daher

$$\min_\mathbf{a} Spur\Big[\mathbf{V_{yy}} - \frac{\mathbf{V_{yx}aa}^T\mathbf{V_{xy}}}{\mathbf{a}^T\mathbf{V_{xx}a}}\Big] \Leftrightarrow \max_\mathbf{a} \frac{\mathbf{a}^T\mathbf{V_{xy}V_{yx}a}}{\mathbf{a}^T\mathbf{V_{xx}a}}$$

$$\min_\mathbf{a} MSEP(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a}) \Leftrightarrow \max_\mathbf{a} Spur[Cov(\mathbf{Y},\widehat{\mathbf{Y}}_\mathbf{a})].$$

(ii) Satz 5.5 *Es gilt*

$$\max_{\mathbf{a}\in R^p\setminus\{0\}} Spur[Cov(BLUP(\mathbf{Y},\mathbf{a}^T\mathbf{X}),\mathbf{Y})] = \lambda_{max}(\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{V}_{\mathbf{xy}}\mathbf{V}_{\mathbf{xy}}^T) \text{ (nach Satz A.3)},$$

und die optimale Lösung wird angenommen für $\mathbf{a}_{pls1} = e.v_{max}(\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{V}_{\mathbf{xy}}\mathbf{V}_{\mathbf{xy}}^T)$*, der zu dem größten Eigenwert der Matrix* $\mathbf{V}_{\mathbf{xx}}^{-1}\mathbf{V}_{\mathbf{xy}}\mathbf{V}_{\mathbf{xy}}^T$ *gehört.*

Im sechsten Kapitel findet sich eine Simulationsstudie unter der multivariaten Normalverteilungsannahme mit einer autoregressiven Kovarianzmatrix, um die Güte der Methoden: modifizierte partielle kleinste Quadrate, partielle kleinste Quadrate, Hauptkomponenten und kanonische Korrelation, miteinander zu vergleichen, am Beispiel des Vorhersageintervalls. Die Simulationsstudie zeigt die Vorteile der modifizierten partiellen kleinsten Quadrate gegenüber den bekannten Methoden.

5.5. Algorithmus der PLS-Methode: NIPALS-Algorithmus

Wenn der Prognosefehler $Y-\widehat{Y}$ orthogonal zu allen in die Prognosefunktion eingehenden Variablen $\mathbf{X}$ steht, Können wir in diesem Sinn die Vorhersage nicht mehr verbessern. Auf der anderen Seite leiten wir folgenden Algorithmus her, um die Vorhersage zu verbessern. Dieser Algorithmus wurde ursprünglich von H. Wold 1975 (siehe Krämer (2007)) eingeführt.
Algorithmus: NIPALS-Algorithmus (siehe Krämer (2007), Li, Udén und v. Rosen (2013))

- **Input:**
  $\mathbf{X}$ eine $\mathbb{R}^p$-wertige ZV (beobachtbar)
  Y eine $\mathbb{R}$-wertige ZV (nicht beobachtbar)
  $\mathbf{X}, Y$ seien gemeinsam verteilt.

  $$\begin{pmatrix}\mathbf{X}\\ Y\end{pmatrix} \text{ eine } \mathbb{R}^p\times\mathbb{R} \text{ wertige ZV, mit}$$
  $$E\begin{pmatrix}\mathbf{X}\\ Y\end{pmatrix} = \begin{pmatrix}0\\0\end{pmatrix}, Cov\begin{pmatrix}\mathbf{X}\\ Y\end{pmatrix} = \begin{pmatrix}\mathbf{V}_{\mathbf{xx}} & \mathbf{V}_{\mathbf{xy}}\\ \mathbf{V}_{y\mathbf{x}} & V_{yy}\end{pmatrix} \in PD(p+1) \text{ bekannt},$$

  Die PLS-Vorhersage erster Art von Y: $g_{pls}(\mathbf{X}) = \mathbf{a}_{pls}^T\mathbf{X}$.

- **Ziel:** Verbessrung von $g_{pls}(\mathbf{X})$

- $\mathbf{X}_1 = \mathbf{X}$

  For $i = 1,\ldots,m < p$ do

  1: $\mathbf{a}_i = \arg\max_{\mathbf{a}^T\mathbf{a}=1} Cov(\mathbf{a}^T\mathbf{X}_i, Y)Cov(Y,\mathbf{a}^T\mathbf{X}_i)$

  2: $T_i = \mathbf{a}_i^T\mathbf{X}_i$ latente Komponente

  3: $\mathbf{X}_{i+1} = \mathbf{X}_i - BLUP(\mathbf{X}_i, T_i)$

end For

- **Output:**
  $\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_m)^T$ eine $\mathbb{R}^m$-wertige ZV (latente Variable)
  $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_m) \in \mathbb{R}^{p \times m}$ (Gewichtsvektoren nach PLS)
  $\widehat{Y}_{pls} = BLUP(Y, \mathbf{T})$ (Vorhersage von Y) Helland (1988).

Dieser PLS-Algorithmus läuft nun in drei Phasen ab. In der ersten Phase werden wir eine Vorhersage bestimmen, die als PLS-Vorhersage (erster Art) heißt. In der zweiten Phase werden wir die latente Variable $\mathbf{T}$ herleiten. In der dritten Phase werden wir die beste lineare Vorhersage von Y unter $\mathbf{T}$ bestimmen.

Eigenschaften der Komponenten T_i

1. $\mathbf{X}_{i+1}$ und T_i sind orthogonal, da

$$\begin{aligned} Cov(\mathbf{X}_{i+1}, T_i) &= Cov(\mathbf{X}_i - BLUP(\mathbf{X}_i, T_i), T_i) \\ &= \mathbf{0}. \end{aligned}$$

2. Hieraus folgt auch, dass

$$\begin{aligned} Cov(T_{i+1}, T_i) &= Cov(\mathbf{a}_{i+1}^T \mathbf{X}_{i+1}, T_i) \\ &= \mathbf{a}_{i+1}^T \, Cov(\mathbf{X}_{i+1}, T_i) \\ &= \mathbf{a}_{i+1}^T \, Cov(\mathbf{X}_i - BLUP(\mathbf{X}_i, T_i), T_i) \\ &= \mathbf{0} \qquad \text{Theorem 3.1.} \end{aligned}$$

Beachte

- Es folgt aus Folgerung 3.4, dass

$$Cov(\mathbf{X}_i - BLUP(\mathbf{X}_i, T_i)) + Cov(BLUP(\mathbf{X}_i, T_i))) = Cov(\mathbf{X}_i)$$

  ist. Hieraus ergibt sich $Cov(\mathbf{X}_{i+1}) \leq Cov(\mathbf{X}_i)$

- wenn $g(\mathbf{X}) = \mathbf{a}_{pls}^T \mathbf{X}$ eine BLUP für Y ist, folgt

$$\begin{aligned} \widehat{\mathbf{Y}}_{pls} &= BLUP(Y, \mathbf{a}_{pls}^T \mathbf{X}) \\ &= BLUP(Y, BLUP(\mathbf{Y}, \mathbf{a}_{pls}^T \mathbf{X})) \\ &= BLUP(Y, \mathbf{X}) \end{aligned}$$

- MSEP von $\widehat{Y}_{pls}$ ist gegeben durch:

$$MSEP(Y, \widehat{Y}_{pls}) = V_{yy} - \mathbf{V}_{y\mathbf{x}} \mathbf{A} (\mathbf{A}^T \mathbf{V}_{\mathbf{xx}} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{V}_{\mathbf{x}y}$$

- Im Fall m=1 erhalten wir:
  (i) MSEP von $\widehat{Y}_{pls}$ ist gegeben durch:

$$MSEP(Y, \widehat{Y}_{pls}) = V_{yy} - \frac{\mathbf{a}_{pls}^T \mathbf{V}_{\mathbf{x}y} \mathbf{V}_{y\mathbf{x}} \mathbf{a}_{pls}}{\mathbf{a}_{pls}^T \mathbf{V}_{\mathbf{xx}} \mathbf{a}_{pls}}$$

  (ii) Der multiple Korrelationskoeffizient von $\widehat{Y}_{pls}$ ist gegeben durch:

$$Corr^2(Y, \widehat{Y}_{pls}) = \frac{\mathbf{a}_{pls}^T \mathbf{V}_{\mathbf{x}y} \mathbf{V}_{y\mathbf{x}} \mathbf{a}_{pls}}{V_{yy}\ \mathbf{a}_{pls}^T \mathbf{V}_{\mathbf{xx}} \mathbf{a}_{pls}}$$

5.6. Schlussbemerkung und Interpretation

In diesem Abschnitt betrachten wir das Modell 1.3.b $\mathbb{R}^3$-wertige normal-verteilte Zufallsvariable mit autoregressiver Varianz-Kovarianz-Matrix $\mathbf{V}$:

$$\begin{pmatrix} X_1 \\ X_2 \\ Y \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \right) \quad \text{mit } \rho \text{ bekannt und} \quad |\rho| < 1.$$

- Der Koeffizientenvektor und **die PLS-Vorhersage erster Art** von Y auf $\mathbf{X}$ sind gegeben durch:

$$\mathbf{a}_{pls} = \frac{1}{\sqrt{1+\rho^2}} \begin{pmatrix} \rho \\ 1 \end{pmatrix}$$

$$g_{pls}(\mathbf{X}) = \frac{1}{\sqrt{1+\rho^2}} (\rho X_1 + X_2) \text{ (nach Satz A.1)}$$

- Der Koeffizientenvektor und **die PCA-Vorhersage erster Art** von Y auf $\mathbf{X}$ sind gegeben durch:

$$\rho > 0: \qquad \mathbf{a}_{pca} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad g_{pca}(\mathbf{X}) = \frac{1}{\sqrt{2}} (X_1 + X_2)$$

$$\rho < 0: \qquad \mathbf{a}_{pca} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \qquad g_{pca}(\mathbf{X}) = \frac{1}{\sqrt{2}} (X_1 - X_2)$$

- Die beste lineare Vorhersage von Y auf $\mathbf{X}$ ist gegeben durch:

$$BLUP(Y, \mathbf{X}) = \widehat{Y} = \rho X_2$$

- Die PLS-Vorhersage von Y auf $\mathbf{X}$ ist gegeben durch:

$$\begin{aligned} \widehat{Y}_{pls} &= BLUP(Y, \mathbf{a}_{pls}^T \mathbf{X}) \\ &= \frac{\mathbf{V}_{y\mathbf{x}} \mathbf{a}_{pls} \mathbf{a}_{pls}^T}{\mathbf{a}_{pls}^T \mathbf{V}_{\mathbf{xx}} \mathbf{a}_{pls}} \mathbf{X} \\ &= \frac{1+\rho^2}{3\rho^2+1} (\rho^2 X_1 + \rho X_2) \end{aligned}$$

- Die PCA-Vorhersage von Y auf $\mathbf{X}$ ist gegeben durch:

$$\begin{aligned}\widehat{Y}_{pca} &= BLUP(Y, \mathbf{a}_{pca}^T\mathbf{X}) \\ &= \frac{\mathbf{V}_{y\mathbf{x}}\mathbf{a}_{pca}\mathbf{a}_{pca}^T}{\mathbf{a}_{pca}^T\mathbf{V}_{\mathbf{xx}}\mathbf{a}_{pca}}\mathbf{X}\end{aligned}$$

und

$$\rho > 0: \quad \widehat{Y}_{pca} = \frac{1}{2}(\rho X_1 + \rho X_2), \qquad \rho < 0: \quad \widehat{Y}_{pca} = \frac{1}{2}(\rho X_1 - \rho X_2)$$

- Und somit ergibt sich:

$$\begin{aligned}Corr^2(\widehat{Y}, Y) &= \rho^2, \widehat{Y} = BLUP(Y, \mathbf{X}). \\ Corr^2(\widehat{Y}_{pls}, Y) &= \frac{\rho^2(1+\rho^2)^2}{3\rho^2+1} \\ Corr^2(\widehat{Y}_{pca}, Y) &= 0.5\ \rho^2(1+\rho) \qquad \rho > 0 \\ Corr^2(\widehat{Y}_{pca}, Y) &= 0.5\ \rho^2(1-\rho) \qquad \rho < 0\end{aligned}$$

- Auf die gleiche Weise ergibt sich der mittlere quadratische Fehler:

$$\begin{aligned}MSEP(\widehat{Y}, Y) &= 1-\rho^2 \\ MSEP(\widehat{Y}_{pls}, Y) &= 1 - \frac{\rho^2(1+\rho^2)^2}{3\rho^2+1} \\ MSEP(\widehat{Y}_{pca}, Y) &= \frac{\mathbf{V}_{y\mathbf{x}}\mathbf{a}_{pca}\mathbf{a}_{pca}^T\mathbf{V}_{y\mathbf{x}}^T}{\mathbf{a}_{pca}^T\mathbf{V}_{\mathbf{xx}}\mathbf{a}_{pca}}\end{aligned}$$

$$MSEP(\widehat{Y}_{pca}, Y) = 1 - 0.5\ \rho^2(1+\rho), \quad \rho > 0$$
$$MSEP(\widehat{Y}_{pca}, Y) = 1 - 0.5\ \rho^2(1-\rho), \quad \rho < 0.$$

Insbesondere folgt:

(a) Ist $|\rho| < 1$, folgt

$$MSEP(\widehat{Y}, Y) \leq MSEP(\widehat{Y}_{pls}, Y) \leq MSEP(\widehat{Y}_{pca}, Y)$$

$$Corr^2(\widehat{Y}, Y) \geq Corr^2(\widehat{Y}_{pls}, Y) \geq Corr^2(\widehat{Y}_{pca}, Y)$$

(b) Wenn ρ nahe bei 1 liegt, folgt

$$MSEP(\widehat{Y}, Y) = MSEP(\widehat{Y}_{pls}, Y) = MSEP(\widehat{Y}_{pca}, Y) \approx 0$$

$$Corr^2(\widehat{Y}, Y) = Corr^2(\widehat{Y}_{pls}, Y) = Corr^2(\widehat{Y}_{pca}, Y) \approx 1$$

(c) Wenn ρ nahe bei -1 liegt, folgt

$$MSEP(\widehat{Y}, Y) = MSEP(\widehat{Y}_{pls}, Y) = MSEP(\widehat{Y}_{pca}, Y) \approx 0$$

$$Corr^2(\widehat{Y}, Y) = Corr^2(\widehat{Y}_{pls}, Y) = Corr^2(\widehat{Y}_{pca}, Y) \approx 1.$$

(d) Im Fall $\rho = 0$ folgt, dass

$$MSEP(\widehat{Y}, Y) = MSEP(\widehat{Y}_{pls}, Y) = MSEP(\widehat{Y}_{pca}, Y) = 1$$

$$Corr^2(\widehat{Y}, Y) = Corr^2(\widehat{Y}_{pls}, Y) = Corr^2(\widehat{Y}_{pca}, Y) = 0.$$

Zur Erklärung ist die konkrete Situation in den Abbildungen (5.1) und (5.2) dargestellt

(5.1) Vergleich des mittleren quadratischen Fehlers in Abhängigkeit von ρ für die Vorhersagen $\widehat{Y}, \widehat{Y}_{pls}$ und $\widehat{Y}_{pca}$.

(5.2) Vergleich der Güte der Vorhersagen von Y in Abhängigkeit von ρ für die Vorhersagen $\widehat{Y}, \widehat{Y}_{pls}$ und $\widehat{Y}_{pca}$.

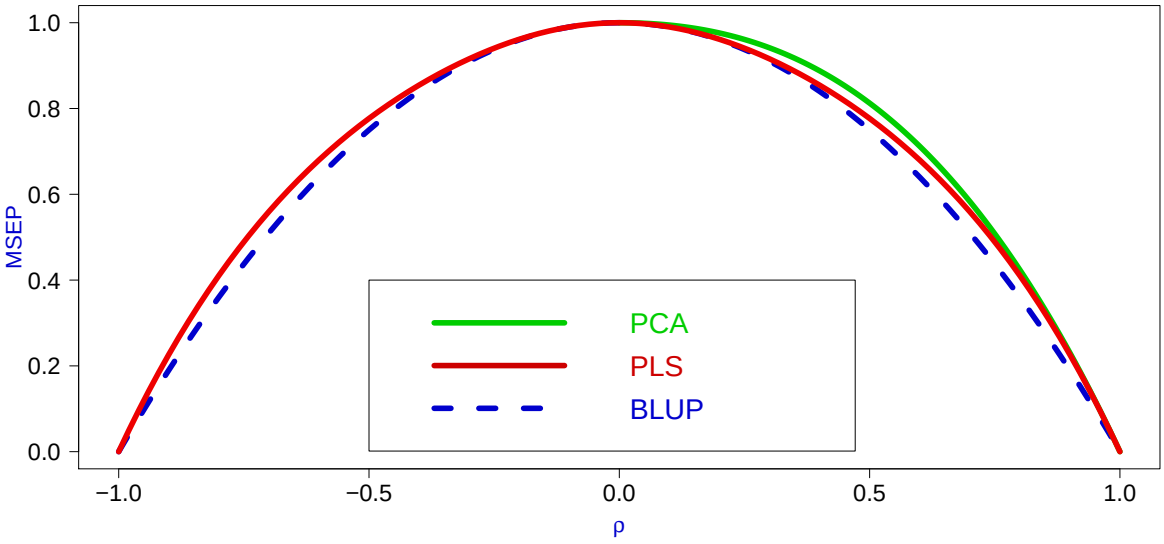

Abbildung 5.1.: Vergleich der MSEP

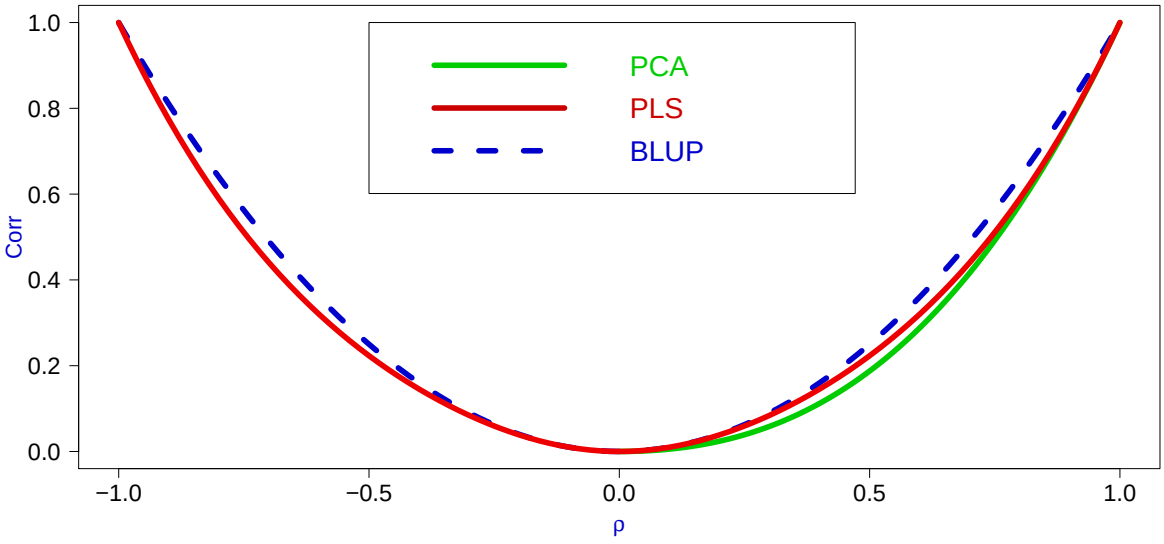

Abbildung 5.2.: Vergleich der Korrelationen

6. Simulationsstudie

In diesem Kapitel werden wir die Güte der Methoden: partielle kleinste Quadrate, Hauptkomponenten und Kanonische Korrelation, modifizierte partielle kleinste Quadrate, am Beispiel der Vorhersageintervalle miteinander durch Simulation vergleichen.

6.1. Modellspezifikation

Wir betrachten wieder das Modell 1.3.b : $\mathbb{R}^{p+q}$-wertige, normal-verteilte Zufallsvariablen,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{yx}} & \mathbf{V_{yy}} \end{pmatrix}\right),$$

wobei

- $\mathbf{X} \sim N_p(\mathbf{0}, \mathbf{V_{xx}})$ (beobachtbar)
- $\mathbf{Y}$ eine $\mathbb{R}^q$-wertige Zufallsvariablen (nicht beobachtbar)
- der Parameter
$$\begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{yx}} & \mathbf{V_{yy}} \end{pmatrix} \in PD(p+q),$$
bekannt ist.

6.2. Dimensionsreduktion

Was unter einer statistischen Datenreduktion zu verstehen ist, soll in diesem Abschnitt formalisiert und damit präzisiert werden, d.h. wie man für das Modell

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N_{p+q}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V_{xx}} & \mathbf{V_{xy}} \\ \mathbf{V_{yx}} & \mathbf{V_{yy}} \end{pmatrix}\right), \tag{6.1}$$

ein Modell

$$\begin{pmatrix} \mathbf{a}^T\mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim N_{1+q}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{a}^T\mathbf{V_{xx}}\mathbf{a} & \mathbf{a}^T\mathbf{V_{xy}} \\ \mathbf{V_{yx}}\mathbf{a} & \mathbf{V_{yy}} \end{pmatrix}\right),$$

durch eine Lineartransformation von $\mathbf{X}$ möglichst gut reduziert, mit anderen Worten hat unsere Aufgabe häufig das Ziel, eine niedrigere Dimensionalität zu erreichen, indem eine Lineartransformationen, benutzt wird.

Vorhersage von Y im reduzierten Modell

$$\begin{aligned}\widehat{\mathbf{Y}}_{\mathbf{a}} &= BLUP(\mathbf{Y}, Z), \text{ mit } Z = \mathbf{a}^T\mathbf{X} \\ &= E(\mathbf{Y}|Z) \\ &= \frac{\mathbf{V_{yx}aa}^T}{\mathbf{a}^T\mathbf{V_{xx}a}}\mathbf{X} \text{ nach (5.6)}\end{aligned}$$

Eigenschaften von $\widehat{\mathbf{Y}}_{\mathbf{a}}$:
Aus den Verteilungsresultaten für (mehrdimensinale) normalverteilte Zufallsvariablen erhalten wir, dass

1. $\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathbf{a}} \sim N(\mathbf{0}, \mathbf{C_a})$ das heißt ${\mathbf{C_a}}^{-1/2}(\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathbf{a}}) \sim N(\mathbf{0}, \mathbf{I}_q)$, wobei

$$\mathbf{C_a} = Cov(\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathbf{a}}) = \mathbf{V_{yy}} - \frac{\mathbf{V_{yx}aa}^T\mathbf{V_{xy}}}{\mathbf{a}^T\mathbf{V_{xx}a}} \in PD(q). \tag{6.2}$$

2. Hieraus ergibt sich, dass

$$(\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathbf{a}})^T\mathbf{C}_{\mathbf{a}}^{-1}(\mathbf{Y} - \widehat{\mathbf{Y}}_a) \sim \chi_q^2$$

3. $\frac{Y_i - \widehat{Y}_{\mathbf{a},i}}{\sqrt{C_{\mathbf{a},ii}}} \sim N(\mathbf{0}, 1), \quad C_{\mathbf{a},ii} = \mathbf{e}_i^T\mathbf{C_a}\mathbf{e}_i$

- Ein $(1-\alpha)$ Vorhersageellipsoid für $\mathbf{Y}$ ist gegeben durch:

$$\mathcal{I}_{Sch}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^q : (\mathbf{y} - \widehat{\mathbf{Y}}_{\mathbf{a}}(\mathbf{x}))^T C_{\mathbf{a}}^{-1}(\mathbf{y} - \widehat{\mathbf{Y}}_{\mathbf{a}}(\mathbf{x})) \leq \chi^2_{q,1-\alpha}\}, \tag{6.3}$$

  wobei $\mathcal{I}(\mathbf{x})$ ein Ellipsoid in $\mathbb{R}^q$ mit dem Mittelpunkt $\widehat{\mathbf{Y}}_{\mathbf{a}}(\mathbf{x})$.

Methode zur Bestimmung der DR-Richtung a

(i) Methode der Hauptkomponente
$\mathbf{a}_{pca}$ heißt Koeffizientenvektor der Hauptkomponente wenn

$$\mathbf{a}_{pca} = \arg \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Var(\mathbf{a}^T\mathbf{X}))$$

gilt. Die Lösung ist

$$\mathbf{a}_{pca} = e.v_{max}(\mathbf{V_{xx}}) \text{ (nach Satz A2)} \tag{6.4}$$

(ii) Methode der partiellen kleinsten Quadrate
$\mathbf{a}_{pls}$ heißt PLS-Koeffizientenvektor wenn

$$\mathbf{a}_{pls} = \arg \max_{\mathbf{a}:\mathbf{a}^T\mathbf{a}=1} Cov(\mathbf{a}^T\mathbf{X}, \mathbf{Y})[Cov(\mathbf{a}^T\mathbf{X}, \mathbf{Y})]^T$$

gilt. Die Lösung ist

$$\mathbf{a}_{pls} = e.v_{max}(\mathbf{V_{xy}V_{yx}}) \text{ (nach Satz A2)} \tag{6.5}$$

(iii) Methode der modifizierten partiellen kleinsten Quadrate
$\mathbf{a}_{pls1}$ heißt modifizierter PLS-Koeffizientenvektor wenn

$$\mathbf{a}_{pls1} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{V_{xy}} \mathbf{V_{xy}^T} \mathbf{a}}{\mathbf{a^T V_{xx} a}}.$$

Die Lösung ist gegeben durch:

$$\mathbf{a}_{pls1} = e.v_{max}(\mathbf{V_{xx}^{-1} V_{xy} V_{yx}}) \text{ (nach Satz A3)} \tag{6.6}$$

(v) Methode der Kanonischen Korrelation
$\mathbf{a}_{cc}$ heißt CC-Koeffizientenvektor wenn

$$\mathbf{a}_{cc} = \arg\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{V_{xy}} \mathbf{V_{yy}^{-1}} \mathbf{V_{xy}^T} \mathbf{a}}{\mathbf{a^T V_{xx} a}} \text{ (Satz A.3).}$$

gilt. Die Lösung ist

$$\mathbf{a}_{cc} = e.v_{max}(\mathbf{V_{xx}^{-1} V_{xy} V_{yy}^{-1} V_{yx}}), \tag{6.7}$$

der zum größten Eigenwert gehört.

6.3. Simulationsdesign und Resultate

In diesem Abschnitt werden das Design der Simulationsstudie, die verwendeten Modelle sowie die erhaltenen Resultate zusammenfassend beschrieben.
Das Ziel ist:

- Vergleich des mittleren quadratischen Fehlers zweier Vorhersagen (MSEP) $\widehat{\mathbf{Y}}$ und $\widehat{\mathbf{Y}}_{\mathbf{a}}$ für $\mathbf{Y}$,

  $\widehat{\mathbf{Y}}$ BUP im vollen Modell $\widehat{\mathbf{Y}}_{\mathbf{a}}$ BUP im reduzierten Modell.

- Vergleich zweier Vorhersageintervalle für $\mathbf{Y}$

Datenerzeugung

$\mathbf{X} \sim N_2(\mathbf{0}, \mathbf{V_{xx}})$ (beobachtbar) $\mathbf{Y} \sim N_2(\mathbf{0}, \mathbf{V_{yy}})$ (nicht beobachtbar)

(a) **Resultate für den Fall: AR(1)-Modell**

$$\mathbf{V_{xx}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \mathbf{V_{yy}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \mathbf{V_{xy}} = \begin{pmatrix} \rho^2 & \rho^3 \\ \rho & \rho^2 \end{pmatrix} \quad \text{mit } \rho \text{ bekannt.}$$

1) Die beste lineare Vorhersage von Y unter X im vollen Modell und der zugehörige mittlere quadratische Fehler sind:

$$\widehat{Y} = (\rho, \rho^2)^T X_2, \qquad MSEP(Y, \widehat{Y}) = 2 - \rho^4 - \rho^2$$

2) Die beste lineare Vorhersage von Y unter X im reduzierten Modell und der zugehörige mittlere quadratische Fehler sind:

$$\widehat{\mathbf{Y}}_{\mathbf{a}} = \frac{\mathbf{V_{yx}aa}^T}{\mathbf{a}^T\mathbf{V_{xx}a}^T}\mathbf{X}, \qquad MSEP(\mathbf{Y}, \widehat{\mathbf{Y}}_{\mathbf{a}}) = \text{Spur}\Big(\mathbf{V_{yy}} - \frac{\mathbf{V_{yx}aa}^T\mathbf{V_{xy}}}{\mathbf{a}^T\mathbf{V_{xx}a}^T}\Big).$$

Vergleich MSEP der Methoden PCA, PLS, PLS1 und CC:

ρ	BLUP	PCA	PLS	PLS1	CC
.9	0.533900	0.607205	0.599682	0.533900	0.533900
.7	1.269900	1.379415	1.343767	1.269900	1.269900
.3	1.901900	1.936235	1.908226	1.901900	1.901900
.1	1.989900	1.994445	1.989997	1.989900	1.989900

(b) **Resultate für den Fall: Zufällige Kovarianz-Matrizen**

Vergleich MSEP der Methoden PCA, PLS, PLS1 und CC:

BLUP	PCA	PLS	PLS1	CC
4.627538	5.292095	4.737850	4.707023	4.708533
1.622197	2.010749	1.781372	1.624496	1.624543
0.9217632	1.0693525	0.9836368	0.9743091	0.9827330
1.474965	2.592202	1.575064	1.547850	1.547962
2.866695	2.904931	2.867041	2.866729	2.866739
3.818612	4.045535	3.898841	3.826117	3.830563

(c) **Beurteilung der Resultate**

Zur Gütebeurteilung der Methode der modifizierten partiellen kleinsten Quadrate werden die Methode der kanonischen Korrelation und die beste erwartungstreue Vorhersage verwendet. Insgesamt stellt die Methode der modifizierten partiellen kleinsten Quadrate in den betrachteten Situationen eine Verbesserung der PLS-Methode dar.

Die folgenden graphischen Darstellungen veranschaulichen die Resultate der Simulationsstudie des AR(1)-Modell.

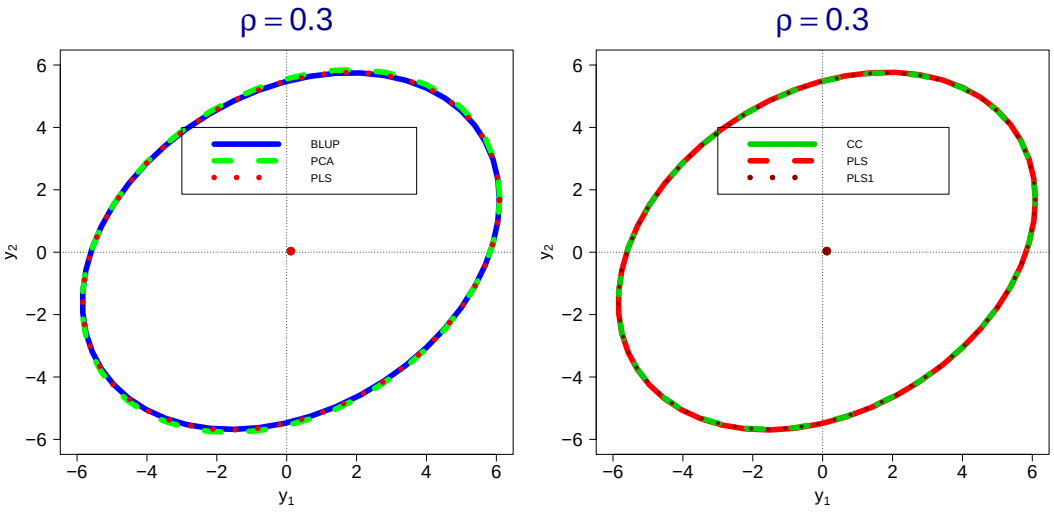

Abbildung 6.1.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$ im AR(1)-Modell mit $\rho = .3, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

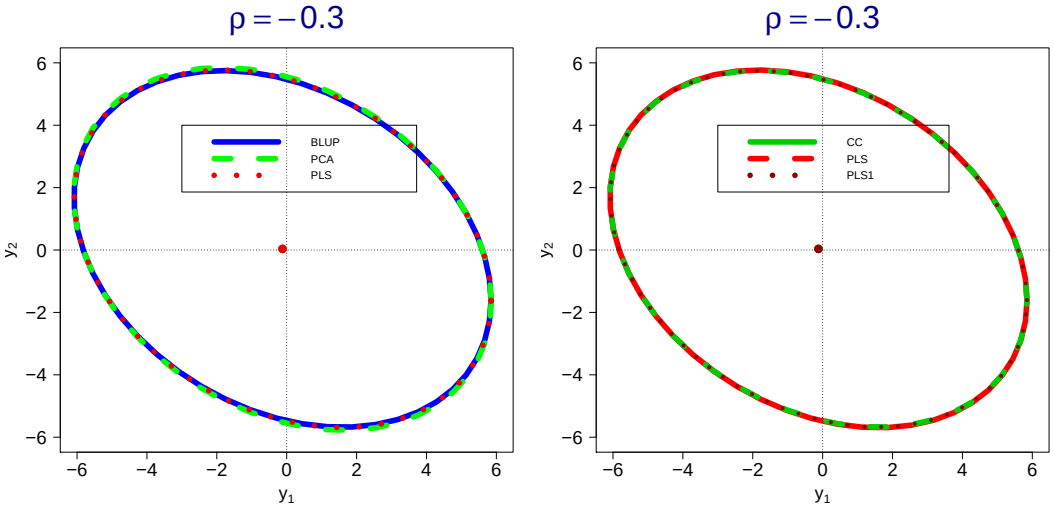

Abbildung 6.2.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$ mit $\rho = -0.3, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

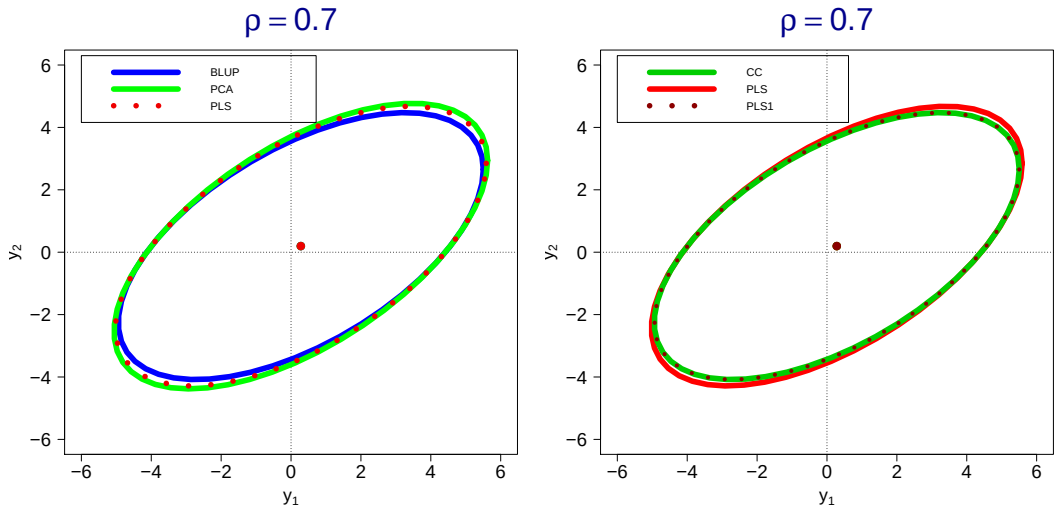

Abbildung 6.3.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$mit $\rho = .7, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

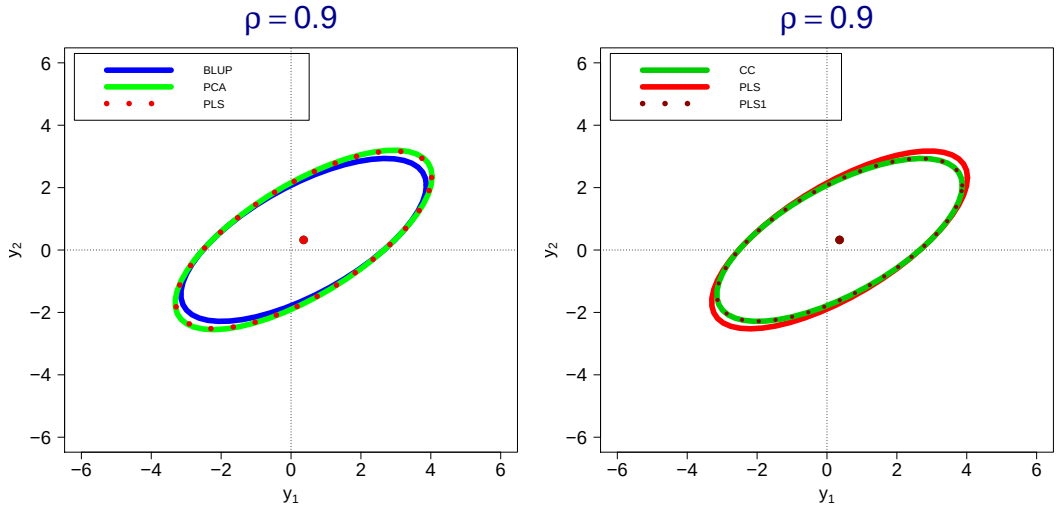

Abbildung 6.4.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$ mit $\rho = 0.9, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

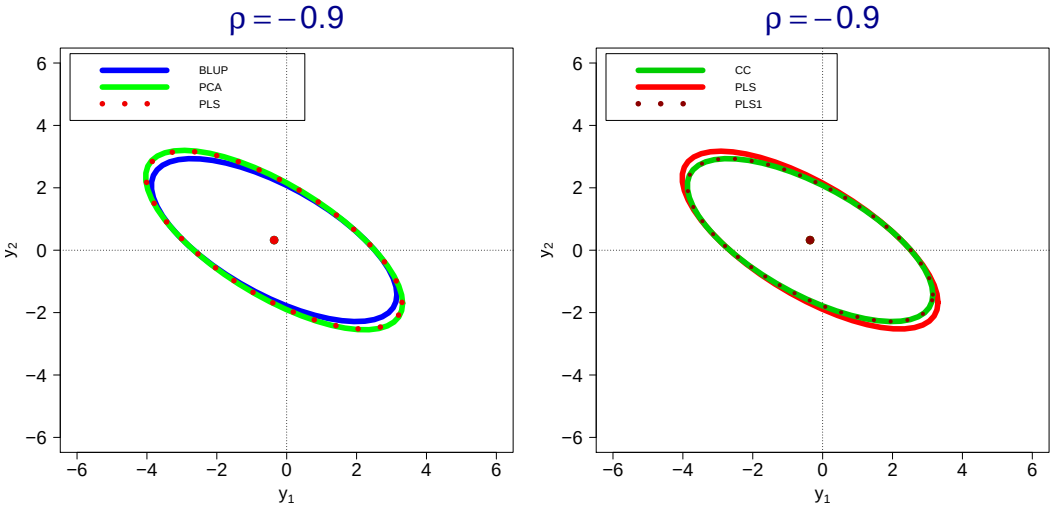

Abbildung 6.5.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$ mit $\rho = -0.9, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

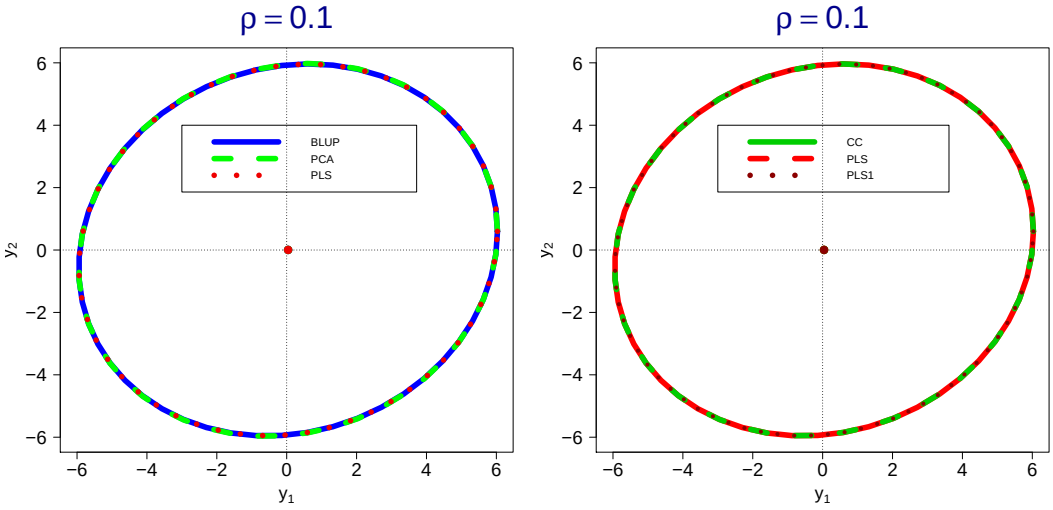

Abbildung 6.6.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$mit $\rho = 0.1, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

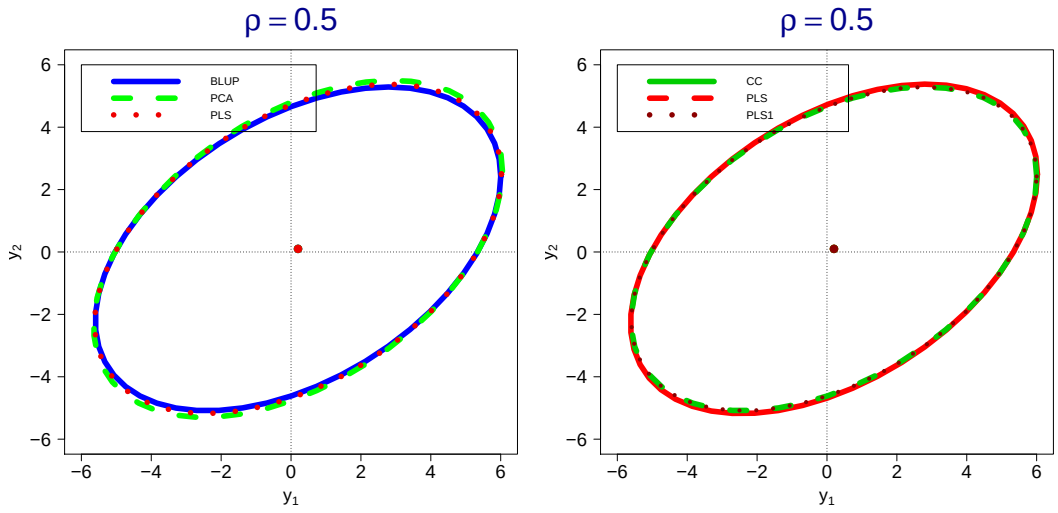

Abbildung 6.7.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$ mit $\rho = 0.5, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

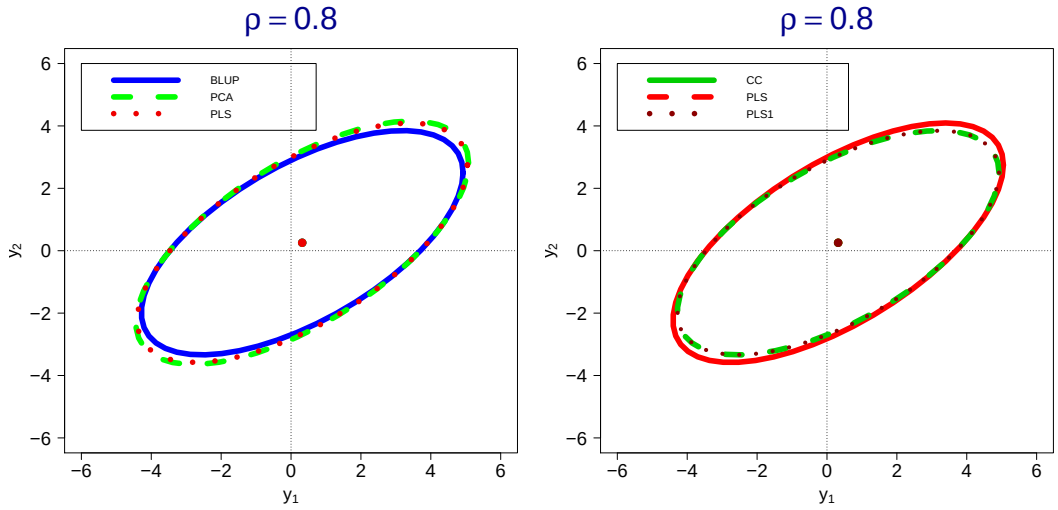

Abbildung 6.8.: Vergleich der Vorhersageellipsoide für $\mathbf{Y}$ und $\mathbf{Y_a}$ mit $\rho = 0.8, \alpha = .05$. Links(blau: BLUP, grün: PCA, rot: PLS). Rechts(grün: CC, dunkelrot: modifizierte-PLS, rot: PLS)

7. Ausblick

Aufgrund der theoretischen Resultate und der numerischen Vergleiche können wir sagen, dass die PLS-Methode somit gewisse Ähnlichkeiten mit der Hauptkomponentenanalyse (PCA) und der multiplen Korrelation (MK) hat. Bei der PCA-Methode werden ebenfalls zueinander orthogonale $\mathbf{a}^T\mathbf{a} = 1$ neue Variablen als Linearkombination aus den ursprünglichen Variablen bestimmt. Im Unterschied zur PLS werden bei der PCA nicht die Kovarianz (Korrelation), sondern nur die Varianzen maximiert, d.h. bei der PLS wird im Gegensatz zur PCA hier der Zusammenhang mit den Komponenten berücksichtigt. Der mittlere quadratische Fehler der modifizierten PLS ist kleiner als der mittlere quadratische Fehler der normalen PLS. Ebenfalls ist der mittlere quadratische Fehler der modifizierten PLS kleiner als der mittlere quadratische Fehler der CC. Zusammenfassend können wir hinsichtlich der Güte der modifizierten PLS-Methode feststellen, dass die modifizierte PLS-Methode besser als die normale PLS-Methode ist. Ein Nachteil der modifizierten PLS-Methode ist, dass der mittlere quadratische Fehler der modifizierten PLS-Methode groß im Gegensatz zur besten erwartungstruen Vorhersage sein kann. Ein wesentlicher Vorteil von PLS ist , dass viele Variablen verarbeitet werden können. Es ist sogar möglich mit weniger Versuchen auszukommen, als Variablen vorhanden sind.
Insgesamt kann gefolgert werden, dass alle Methoden (BLUP, PLS, PCA, PLS1, CC) gewisse Ähnlichkeiten baben, sodass alle den mittleren quadratischen Fehler der Vorhersage verringern. Natürlich bleiben Ansatzpunkte für weitere Forschung. Eine Aufgabe für die Zukunft besteht zum Beispiel darin, die Theorie der Vorhersage für SUR-Modell (Seemingly Unrelated Regressions) und Growth curve-Modell zu entwickeln. Schließlich kann als Anwendung in der Schätztheorie versucht werden, verbesserte Schätzer im Sinne der Verringerung der Varianz zu finden.

A. Lösungen einiger Maximierungsprobleme

In diesem Anhang geben wir eine kompakte Einführung in Optimierungsprobleme, die auf Matrizen basiert. Für Beweise und weitere Eigenschaften verweisen wir auf die Literatur, z.B. Puntanen, Styan und Isotalo (2013, Kap.22.18,22.25), Fujikoshi, Ulyanov und Shimizu (2010, Anhang. A2).

Satz A.1 *Seien* $\mathbf{B}$ *eine positiv definite symmetrische* $n \times n$*-Matrix und* $\mathbf{x} \in \mathbb{R}^n$

$$(a) \quad \max_{\mathbf{a}^T\mathbf{B}\mathbf{a}=1} \mathbf{a}^T\mathbf{x} = \sqrt{\mathbf{x}^T\mathbf{B}^{-1}\mathbf{x}}$$

$$(b) \quad \arg\max_{\mathbf{a}^T\mathbf{B}\mathbf{a}=1} \mathbf{a}^T\mathbf{x} = \frac{\mathbf{B}^{-1}\mathbf{x}}{\sqrt{\mathbf{x}^T\mathbf{B}^{-1}\mathbf{x}}}.$$

Satz A.2 *Sei* $\mathbf{B}$ *eine symmetrische* $n \times n$*-Matrix, dann folgt*

$$\max_{\mathbf{a}\in R^n\setminus\{0\}} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = \lambda_{max}(\mathbf{B}) \quad \textit{und} \quad \arg\max_{\mathbf{a}\in R^n\setminus\{0\}} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{a}} = e.v_{max}(\mathbf{B})$$

Satz A.3 *(Extrema quadratischer Formen unter qudratischer Nebenbedingung) Seien* $\mathbf{A}$ *eine symmetrische* $n \times n$*-Matrix und* B *eine positiv definite* $n \times n$*-Matrix. Sei* λ_n *der größte Eigenwert von* $\mathbf{B}^{-1}\mathbf{A}$*. Dann gilt*

$$\max_{\mathbf{a}\in R^n\setminus\{0\}} \frac{\mathbf{a}^T\mathbf{A}\mathbf{a}}{\mathbf{a}^T\mathbf{B}\mathbf{a}} = \lambda_n \text{ und } \max_{\mathbf{a}:\mathbf{a}^T\mathbf{B}\mathbf{a}=1} \mathbf{a}^T\mathbf{A}\mathbf{a} = \lambda_n,$$

und die Maximums werden jeweils in einem Eigenvektor der Matrix $\mathbf{B}^{-1}\mathbf{A}$ *zu dem Eigenwert* λ_n *erreicht.*

Auf ähnliche Weise, wie im Satz A.1 erhaltem wir

Satz A.4 *Seien* $\mathbf{B}$ *eine positiv definite* $n \times n$*-Matrix und* $\mathbf{x} \in \mathbb{R}^n$*. Dann ist*

$$(a) \quad \max_{\mathbf{a}\in R^n\setminus\{0\}} \frac{(\mathbf{a}^T\mathbf{x})^2}{\mathbf{a}^T\mathbf{B}\mathbf{a}} = \mathbf{x}^T\mathbf{B}^{-1}\mathbf{x}$$

$$(b) \quad \arg\max_{\mathbf{a}\in R^n\setminus\{0\}} \frac{(\mathbf{a}^T\mathbf{x})^2}{\mathbf{a}^T\mathbf{B}\mathbf{a}} = \frac{\mathbf{B}^{-1}\mathbf{x}}{\mathbf{x}^T\mathbf{B}^{-1}\mathbf{x}}.$$

Literaturverzeichnis

[1] Alvin, R.C. (2008): Linear models in statistics. Wiley New York.

[2] Christensen, R. (2011): Plane answers to complex questions, Fourth edition. Springer New York.

[3] Christensen, R., Lin,Y. (2013): Linear models that allow perfect estimation, stat Papers 54:695-708.

[4] Frank, I., Friedman J. A. (1993): Statistical view of some chemometrics regression tools, Technometrics 35:109-148.

[5] Cook, R. D. (1994): On the Interpretation of Regression Plots, journal of the American Statistical Assocition, 89:177-189.

[6] Fujikoshi,Y., Ulyanov, V. V., Shimizu, R. (2010): Multivariate Statistics: High Dimensional and Large-Sample Approximations. Wiley New York.

[7] Garthwaite, P. H. (1994): An interpretation of partial least squares. Journal of American Statistical Association, 89:122-127.

[8] Grüning, M. (2005): Untersuchungen zur Diskriminanzanalyse mit hochdimensionalen Daten, Magdeburg, Univ., Fak. für Mathematik, Diss.

[9] Helland, I. S. (1988): On the structure of partial least squares regression. Comm. Statist. Simulation Comput. 17(2):581-607.

[10] Helland, I. S. (1990): Partial Least Squares Regression and Statistical Models. Scand. J. Statist.17:97- 114.

[11] Hocking, R. R. (2005): Methods and Applications of Linear Models : Regression and the Analysis of Variance, second edition. Wiley New York.

[12] Isotalo, J., Puntanen, S. (2009): A note on the equality of the OLSE and the BLUE of the parametric function in the general Gauss-Markov model. Stat Papers 50:185:193.

[13] Krämer, N. (2007): An overview on the shrinkage properties of partial least squares regression. Jounal of Computational. 22(2): 249-273

[14] Kreiß, J. P., Neuhaus, G. (2006): Einführung in die Zeitreihenanalyse, Springer Berlin.

[15] Li, K-C. (1991): Siliced Inverse Regression for Dimension Reduction, journal of the American Statistical Assocition, 86:316-342.

[16] Li, Y., Udén, P., v. Rosen, D. (2013) : A two-step PLS inspired method for linear prediction with group effect. Sankhya A. 75(1): 96-117

[17] Liu,Y., Rayens, W., (2007): PLS and dimension reduction for classifcation. Jounal of Computational. 22(2): 189-208

[18] Puntanen, S., Styan G. P.H., Isotalo, J. (2011): Matrix Tricks for Linear Statistical Models. Springer-Verlag Berlin.

[19] Puntanen, S., Styan G. P. H., Isotalo, J. (2013): Formulas Useful for Linear Regression Analysis and Related Matrix Theory. Springer-Verlag Berlin.

[20] Rao, C. R., Shalabh, Toutenburg, H., Heumann. C. (2008): Linear Models and Generalizations : least squares and alternatives, third edition. Springer Berlin.

[21] Rao, C. R. (1973): Linear Statistical Inference and its Applications, second edition. Wiley New York.

[22] Schlittgen, R. (2009): Multivariate Statistik. Oldenbourg München

[23] Schmidt, D. K. (1988): Prediction in the linear model: A direct approach. Metrika. 48(2): 141-147

[24] Seber, G. A. (1984): Multivariate Observations. Wiley New York.

[25] Sengupta, D. S., Jammalamadaka, R. S. (2003): Linear Models An Integrated Approach. World Scientific Publishing Company London.

[26] Shao, J. (2003): Mathematical statistics, second edition. Springer New York.

[27] Witting, H. (1985): Mathematische Statistik I. Teubner Stuttgart.

[28] Wold, H. (1966): Estimation of principal components and related models by iterative least squares.In P.R. Krishnaiah (ed.) Multivariate Analysis, 391-420. New York: Academic Press.