Statistische Methoden der Parameteridentifikation in strukturmechanischen Modellen

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium

(Dr. rer. nat.)

von Diplom-Mathematiker Robert Offinger

geb. am 13.06.1970 in Monheim/Schwaben

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

Gutachter: Prof. Dr. Norbert Gaffke
Prof. Dr. Thomas Müller-Gronbach

eingereicht am: 01. Juli 2009
Verteidigung am: 28. September 2009

Inhaltsverzeichnis

Symbol- und Abkürzungsverzeichnis iii

Zusammenfassung v

1. Nichtlineare Normalverteilungsmodelle 1
- 1.1. Das klassische Modell . . . 1
- 1.2. Bayes-Modell mit Vorinformation . . . 1
- 1.3. Nichtlineare Modellfunktionen und Kovarianzmatrizen . . . 1
  - 1.3.1. Ein sehr einfaches Beispiel . . . 2
  - 1.3.2. Das lineare Modell als Spezialfall . . . 2
  - 1.3.3. Eigenwerte und Eigenvektoren eines strukturellen Eigenwertproblems . . . 2
- 1.4. A-priori-Verteilungen . . . 3

2. Kleinste-Quadrate-Schätzung im Modell ohne a-priori-Information 4
- 2.1. Eigenschaften des Kleinste-Quadrate-Schätzers . . . 4
  - 2.1.1. Eigenschaften im linearen Modell . . . 4
  - 2.1.2. Eigenschaften im nichtlinearen Modell . . . 9
- 2.2. Tests und Konfidenzbereiche basierend auf dem Kleinste-Quadrate-Schätzer 29
  - 2.2.1. Lineares Modell . . . 29
  - 2.2.2. Nichtlineares Modell . . . 35

3. Kleinste-Quadrate-Schätzung: numerische Methoden 39
- 3.1. Lineares Modell . . . 39
  - 3.1.1. Singulärwertzerlegung, Konditionszahl und Sensitivitätsanalysen . 39
  - 3.1.2. Verfahren basierend auf der Cholesky-Zerlegung . . . 52
  - 3.1.3. Verfahren basierend auf der LR-Zerlegung mittels Gaußscher Elimination . . . 54
  - 3.1.4. Verfahren basierend auf der QR-Zerlegung . . . 55
  - 3.1.5. Verfahren basierend auf der Singulärwertzerlegung . . . 57
- 3.2. Nichtlineares Modell . . . 61
  - 3.2.1. Newton-Verfahren . . . 70
  - 3.2.2. Gauß-Newton-Verfahren . . . 77
  - 3.2.3. Quasi-Newton-Verfahren: BFGS und weitere Verfahren . . . 89
  - 3.2.4. Abbruchkriterien . . . 96

4. Schätzung bei a-priori-Information: Theorie 99

4.1. Bayes-Risiko und a-posteriori-Maximierung . . . 99
4.2. Bayes-Schätzung im linearen Modell . . . 105
4.3. Der Modalwert-Schätzer im nichtlinearen Modell . . . 107
4.4. Empirische Bayes-Schätzung . . . 111

5. Schätzung bei a-priori-Information: numerische Verfahren 113
5.1. Numerische Verfahren für den Modalwert-Schätzer . . . 113
5.1.1. Der Algorithmus von Collins u.a. . . . 114
5.1.2. Der Algorithmus von Friswell . . . 117
5.1.3. Ein Algorithmus basierend auf dem Gauß-Newton-Verfahren . . . 119
5.1.4. Der Algorithmus von Pázman . . . 120
5.1.5. Weitere Verfahren und Abbruchkriterien . . . 122
5.1.6. Vergleich der Verfahren in einem nichtlinearen Fall . . . 124
5.2. Numerische Verfahren zur Empirischen Bayes-Schätzung . . . 128
5.2.1. Approximative Lösung durch Linearisierung . . . 129
5.2.2. Approximative Lösung durch Laplace-Approximation . . . 130
5.2.3. Vergleich der Approximationen an einem Beispiel . . . 133

6. Anwendungen 135
6.1. Kriechversuche . . . 135
6.2. Finite-Elemente-Modelle für Eigenwertprobleme . . . 139
6.2.1. Eigenwertprobleme bei partiellen Differentialgleichungen . . . 139
6.2.2. Lineare Schwingungen . . . 140
6.2.3. Verteilungsannahmen für die Messungen der Eigenfrequenzen und der Eigenvektoren . . . 142
6.2.4. Ableitungen der Eigenwert- und Eigenvektorfunktionen . . . 144
6.2.5. Räumliche Modelle für a-priori-Verteilungen . . . 148

A. Beweise 152
A.1. Beispiel 2.5 . . . 152
A.2. Beispiel 2.32 . . . 153
A.3. Beispiel 2.38 . . . 154
A.4. Rechenoperationen bei der QR-Zerlegung . . . 155
A.5. Rechenoperationen bei der Singulärwertzerlegung . . . 159

B. Grundlagen aus der Differentialgeometrie 165

Literaturverzeichnis 174

Symbol- und Abkürzungsverzeichnis

$\overline{\mathbb{R}}$	$= \mathbb{R} \cup \{\pm\infty\}$ kompaktifizierter reeller Raum
$\mathbb{R}^n$	n-dimensionaler reeller Raum
$\mathbb{R}^{N\times N}$	Raum der reellen $N \times N$-Matrizen
$\mathrm{PD}(p)$	Menge der positiv definiten Matrizen aus $\mathbb{R}^{p\times p}$
$\mathrm{PSD}(p)$	Menge der positiv semidefiniten Matrizen aus $\mathbb{R}^{p\times p}$
$\mathbf{0}_n$	Null-Vektor $(0,\ldots,0)^{\mathrm{T}}$ im $\mathbb{R}^n$
$\mathbf{1}_n$	Einser-Vektor $(1,\ldots,1)^{\mathrm{T}}$ im $\mathbb{R}^n$
I_N	$N \times N$ Identitätsmatrix
$\mathbf{0}_{m\times n}$	$m \times n$-Nullmatrix
e_j	j-ter Einheitsvektor
$\boldsymbol{i}$	imaginäre Einheit
$\mathbf{1}_A$	Indikatorfunktion der Menge A
$\#A$	Anzahl der Elemente in der Menge A
$\mathcal{B}^n$	Borel-σ-Algebra über $\mathbb{R}^n$
$\mathcal{B}^n$	Borelsche σ-Algebra im $\mathbb{R}^n$
$\mathcal{B}^n_\Theta$	Borelsche σ-Algebra im $\mathbb{R}^n$ eingeschränkt auf Teilmengen von $\Theta \in \mathcal{B}^n$
$\mathcal{N}(\beta,\Sigma)$	multivariate Normalverteilung mit Erwartungswertvektor β und Kovarianzmatrix Σ
χ^2_n	χ^2-Verteilung mit n Freiheitsgraden
$\chi^2_{n;1-\alpha}$	$(1-\alpha)$-Quantil der χ^2_n-Verteilung
$F_{p,q}$	F-Verteilung mit p Zählerfreiheitsgraden und q Nennerfreiheitsgraden
t_n	t-Verteilung mit n Freiheitsgraden
$F_{p,q;1-\alpha}$	$(1-\alpha)$-Quantil der $F_{p,q}$-Verteilung
$t_{n;1-\alpha}$	$(1-\alpha)$-Quantil der t_n-Verteilung
$f_{c,A}$	Dichtefunktion der $\mathcal{N}(c,A)$-Verteilung
$P^{\boldsymbol{Y}\mid\boldsymbol{\theta}=\vartheta}$	bedingte Verteilung von $\boldsymbol{Y}$ gegeben $\boldsymbol{\theta} = \vartheta$
$\boldsymbol{\lambda}^N$	N-dimensionales Lebesgue-Maß auf $(\mathbb{R}^N, \mathcal{B}^N)$
$\xrightarrow[\mathcal{L}]{}$	Verteilungskonvergenz
$< x, y >_\Sigma$	$= x^{\mathrm{T}}\Sigma^{-1}y$; von Σ induziertes Skalarprodukt
$\|v\|^2_\Sigma$	$= v^{\mathrm{T}}\Sigma^{-1}v$
$\mathrm{rg}(X)$	Rang der Matrix X
$\Sigma^{1/2}$	Quadratwurzel der positiv definiten Matrix Σ
A^+	Moore-Penrose-Inverse der Matrix A
$\|x\|$	Euklidische Norm des Vektors x

$\|C\|$	Spektralnorm der $m \times n$ Matrix C
$\|C\|_{\text{Frob}}$	$= \left(\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}^2\right)^{1/2}$; Frobenius-Norm der $m \times n$ Matrix C
$\rho(C)$	Spektralradius der $n \times n$ Matrix C (betragsgrößter Eigenwert)
$\text{Bild}(X)$	Bild der Matrix X
$\text{Null}(X)$	Nullraum der Matrix X
$\mathcal{L}^{\perp}$	Orthogonalraum zum linearen Raum $\mathcal{L}$
$\text{E}_{\vartheta,\sigma}(\boldsymbol{Y})$	Erwartungswert des Zufallsvektors $\boldsymbol{Y}$ unter den Parameterwerten ϑ und σ (wenn die Parameterwerte aus dem Zusammenhang klar sind, werden sie unterdrückt)
$\text{E}[P]$	Erwartungswert der Wahrscheinlichkeitsverteilung P
$\text{D}_{\vartheta,\sigma}(\boldsymbol{Y})$	Kovarianzmatrix des Zufallsvektors $\boldsymbol{Y}$ unter den Parameterwerten ϑ und σ
$\text{D}[P]$	Varianz der Wahrscheinlichkeitsverteilung P
$\text{Cov}(\boldsymbol{X}, \boldsymbol{Y})$	$= \text{E}((\boldsymbol{X} - \text{E}(\boldsymbol{X}))(\boldsymbol{Y} - \text{E}(\boldsymbol{Y}))^{\text{T}})$; Kovarianz von $\boldsymbol{X}$ und $\boldsymbol{Y}$
$\leq_{\mathfrak{L}}$	Loewner-Halbordnung, $A \leq_{\mathfrak{L}} B$, wenn $B - A$ positiv semidefinit
$C^k(U)$	k-mal stetig (partiell) differenzierbare ($\mathbb{R}^N$-wertige) Funktionen auf U
$\text{int}(\Theta)$	Inneres von Θ
$\text{cl}(\Theta)$	Abschluss von Θ
$\partial\Theta$	$= \text{cl}(\Theta) \setminus \text{int}(\Theta)$; topologischer Rand von Θ
$J_\eta(\vartheta)$	$N \times p$-Jacobi-Matrix von η, wobei $\vartheta \in \mathbb{R}^p$ und $\eta : \mathbb{R}^p \to \mathbb{R}^N$
$S^{\eta,y}(\vartheta)$	$= \|y - \eta(\vartheta)\|_{\Sigma}^2$; Residuenquadratsumme bei Beobachtung y und Parameterwert ϑ
$H_\eta(\vartheta)$	$= \sum_{j=1}^{N} \left(\Sigma^{-1}\left(\eta(\vartheta) - y\right)\right)_j \text{Hess}\, \eta_j(\vartheta)$; Term, der im Gauß-Newton-Verfahren vernachlässigt wird
$\mathcal{P}_{X,\Sigma}$	$= X(X^{\text{T}}\Sigma^{-1}X)^{+}X^{\text{T}}\Sigma^{-1}$; Projektion auf $\text{Bild}(X)$ bzgl. $<, >_{\Sigma}$
$f\|_U$	Einschränkung einer Funktion f auf U
$\text{grad}\, f$	Gradient der Funktion $f : \mathbb{R}^p \to \mathbb{R}$ (Zeilenvektor)
$\text{Hess}\, f$	Hesse-Matrix der Funktion $f : \mathbb{R}^p \to \mathbb{R}$
$\text{pr}_{\mathbb{R}^n}^{\mathbb{R}^N}$	Projektion vom $\mathbb{R}^N$ in den $\mathbb{R}^n$ durch Abbildung auf die ersten n Komponenten (für $n < N$)
g^+	$= \max(0, g)$; Positivteil einer reellwertigen Funktion g
g^-	$= \max(0, -g)$; Negativteil einer reellwertigen Funktion g
$\text{sgn}(x)$	Vorzeichen von x
$\bar{y}$	$= \frac{1}{N}\sum_{i=1}^{N} y_i$; empirischer Mittelwert von $y \in \mathbb{R}^N$
S^{n-1}	$= \{x \in \mathbb{R}^n : \|x\| = 1\}$; Einheitssphäre im $\mathbb{R}^n$
$\dot{x}$	1. Ableitung des Ortsvektors $x = x(t)$ nach t
$\ddot{x}$	2. Ableitung des Ortsvektors $x = x(t)$ nach t

Zusammenfassung

Ziel der Arbeit ist die Entwicklung und Analyse von Verfahren zur Parameterschätzung in nichtlinearen Normalverteilungsmodellen, wie sie bei Finite-Elemente-Modellen der strukturmechanischen Dynamik auftreten. Das Augenmerk liegt hier auch auf dem Fall, dass man noch Zusatzinformation über den Parameter besitzt, die etwa aus Vorversuchen oder Erfahrungswerten stammt, und die man durch eine a-priori-Verteilung spezifiziert. Hierbei wird auch auf die in der Literatur sehr selten behandelte Situation eingegangen, dass die a-priori-Verteilung unbekannte Parameter enthält.

Im ersten Kapitel werden nichtlineare Normalverteilungsmodelle mit und ohne Vorinformation vorgestellt. Im zweiten Kapitel wird ein Abriss der Theorie zur Parameterschätzung in linearen und nichtlinearen Normalverteilungsmodellen ohne Vorinformation gegeben, wobei wir uns tiefergehend mit der Existenz und Eindeutigkeit der Lösung des Kleinste-Quadrate-Problems beschäftigen. Das dritte Kapitel behandelt zum einen die numerischen Eigenschaften von verschiedenen Verfahren zur Lösung des linearen Kleinste-Quadrate-Problems inklusive einer ausführlichen Sensitivitätsanalyse des Problems. Zum anderen stellt es Verfahren zur Lösung des nichtlinearen Kleinste-Quadrate-Problems vor und betrachtet ihre theoretischen Konvergenzeigenschaften. Das vierte Kapitel ist theoretischen Grundlagen der Parameterschätzung, wenn Vorinformation gegeben ist, gewidmet. Hierauf werden im fünften Kapitel sowohl einige aus der Literatur bekannte als auch neu entwickelte Algorithmen zur Ermittlung der Parameterschätzung in Modellen mit Vorinformation vorgestellt und bewertet. Im sechsten Kapitel finden sich praktische Anwendungsbeispiele aus den Ingenieurwissenschaften und es werden die vorgestellten Algorithmen an diesen praktischen Beispielen untersucht. Abschließend werden noch Ausblicke auf mögliche Zielrichtungen für weitere Untersuchungen gegeben. Im Anhang finden sich einige Beweise eher technischer Natur, die im Haupttext den Lesefluss gestört hätten, und eine kurze Einführung in die Differentialgeometrie, da wir unterwegs einige Begriffe und Resultate daraus benötigen.

Computercodes, mit denen die dargestellten Problemstellungen gelöst wurden, werden auf Wunsch gerne zur Verfügung gestellt.

Summary

Aim of this thesis is the development and analysis of parameter estimation procedures in nonlinear Gaussian models that are used in finite element models in structural dynamics. We especially focus on the case of prior information on the parameters, which may result from former experiments or from expert knowledge. The prior information is modelled by a prior distribution. In this context we also address the issue that the prior distribution contains unkown parameters, which has been neglected in the literature.

The first chapter features an introduction to nonlinear Gaussian models with or without prior information. In the second chapter we give a summary of parameter estimation theory in linear and nonlinear Gaussian models without prior information. Here, we present an in-depth study of the existence and uniqueness of the solution of the least squares estimation problem. We discuss the numerical properties of procedures for the solution of the least squares problem including a detailed sensitivity analysis of the problem in the third chapter. Thereupon we list several algorithms for the solution of the nonlinear least squares problem and examine their theoretical convergence properties. In the fourth chapter we address the foundations of parameter estimation theory in models with prior information. Then, in the fifth chapter we present and analyse some algorithms known from the literature and also new algorithms for parameter estimation in models with prior information. Practical applications from engineering are presented in the sixth chapter and the algorithms are tested for the solution of practical examples. Finally, we give some recommendations for future work. For the sake of readability you will find some technical proofs in the appendix. Further we give a short introduction to differential geometry, since we are using some concepts and results from this field.

On request we are pleased to provide source codes that were used for the solution of the given examples and applications.

1. Nichtlineare Normalverteilungsmodelle

1.1. Das klassische Modell

Die Überlegungen im Folgenden sind auf allgemeine nichtlineare Modelle von der Form

$$\boldsymbol{Y} \sim \mathcal{N}(\eta(\vartheta), \Sigma) \tag{1.1}$$

mit unbekanntem Parameter $\vartheta \in \Theta \subseteq \mathbb{R}^p$, $\Theta \neq \emptyset$, und **Modellfunktion**

$$\eta : \Theta \to \mathbb{R}^N$$

anwendbar. Dabei bezeichnet $\mathcal{N}(\eta(\vartheta), \Sigma)$ die multivariate Normalverteilung mit Erwartungswertvektor $\eta(\vartheta) \in \mathbb{R}^N$ und positiv definiter Kovarianzmatrix $\Sigma \in \mathbb{R}^{N\times N}$. Die Modellfunktion η ist im Allgemeinen nichtlinear, die Kovarianzmatrix Σ kann bekannt oder unbekannt sein. Etwas vereinfacht wird die statistische Analyse, wenn die Kovarianzmatrix bis auf einen Faktor bekannt ist, also von der Form $\Sigma = \sigma^2\Sigma_0$ mit bekannter Matrix $\Sigma_0 \in \mathbb{R}^{N\times N}$ und unbekanntem Parameter $\sigma > 0$.
Das Modell (1.1) werden wir im Folgenden als das „klassische Modell“ bezeichnen.

Alternativen zur Normalverteilungsannahme werden im Abschnitt 6.2.3 besprochen.

1.2. Bayes-Modell mit Vorinformation

Oft hat man noch Vorinformationen über den unbekannten Parameter und betrachtet ihn als Zufallsvariable $\boldsymbol{\theta}$, um die Unsicherheit der Vorinformation statistisch zu modellieren. So verwendet man etwa das Modell

$$P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} = \mathcal{N}(\eta(\vartheta), \Sigma) \tag{1.2}$$

und die **a-priori-Annahme**

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu(\xi), B(\rho)) \tag{1.3}$$

mit $\mu : \mathbb{R}^w \to \mathbb{R}^p$, $B : \mathbb{R}^v \to \mathbb{R}^{p\times p}$ und $\Theta = \mathbb{R}^p$. Der Mittelwert bzw. die Streuung der a-priori-Verteilung in (1.3) hängen eventuell noch von unbekannten Parametern $\xi \in \mathbb{R}^w$ bzw. $\rho \in \mathbb{R}^v$ ab.

1.3. Nichtlineare Modellfunktionen und Kovarianzmatrizen

In diesem Abschnitt stellen wir einige Modellfunktionen η und Kovarianzmatrizen Σ vor, die in dieser Arbeit eine wichtige Rolle spielen werden.

1.3.1. Ein sehr einfaches Beispiel

Ein sehr einfaches Beispiel ist durch die Modellfunktion

$$\eta : \mathbb{R} \to \mathbb{R}^N,\ \eta(\vartheta) = \frac{1}{2}\,\vartheta^2 \mathbf{1}_N = \left(\frac{1}{2}\,\vartheta^2, \ldots, \frac{1}{2}\,\vartheta^2\right)^{\mathrm{T}} \tag{1.4}$$

und die Kovarianzmatrix $\Sigma = \sigma_0^2 I_N$ mit bekannter Streuung σ_0^2 gegeben. In diesem Modell ohne a-priori-Annahmen liegen also N Beobachtungen $\boldsymbol{Y}_i = \frac{1}{2}\vartheta^2 + \mathbf{e}_i$ mit unabhängigen und identisch gemäß $\mathcal{N}(0, \sigma_0^2)$-verteilten Messfehlern $\mathbf{e}_1, \ldots, \mathbf{e}_N$ vor.

1.3.2. Das lineare Modell als Spezialfall

Gelegentlich werden wir als Spezialfall das lineare Modell

$$\eta : \Theta \to \mathbb{R}^N,\ \eta(\vartheta) = X\vartheta$$

oder das affin-lineare Modell

$$\eta : \Theta \to \mathbb{R}^N,\ \eta(\vartheta) = y_0 + X \cdot (\vartheta - \vartheta_0)$$

mit $y_0 \in \mathbb{R}^N$, $X \in \mathbb{R}^{N\times p}$ und $\vartheta_0 \in \mathbb{R}^p$ betrachten, um Eigenschaften von Algorithmen zu untersuchen bzw. die Konstruktion von Algorithmen zu motivieren.

1.3.3. Eigenwerte und Eigenvektoren eines strukturellen Eigenwertproblems

Durch Anpassung von Parametern eines Finite-Elemente-Modells auf der Basis von Experimenten versucht man ein geeignetes mathematisches Modell für das Schwingungsverhalten eines physikalischen Systems (z.B. einer Platte oder eines Motorblocks) zu erhalten. Wir stellen hier nur wesentliche Bestandteile vor und verweisen auf den Abschnitt 6.2, wo solche Finite-Elemente-Modelle genauer erläutert und untersucht werden.

Bezeichnet man mit $x_i(t) \in \mathbb{R}$, $i = 1, \ldots, n$, den Ort der i-ten Komponente des Systems zur Zeit t, so betrachtet man unter Vernachlässigung von Reibungseffekten ein **ungedämpftes dynamisches System** mit der Bewegungsgleichung

$$M(\vartheta)\ddot{x} + K(\vartheta)x = \mathbf{0}_n$$

mit strukturellem Parameter $\vartheta \in \mathbb{R}^p$, Massenmatrix $M(\vartheta) \in \mathbb{R}^{n\times n}$ und Steifigkeitsmatrix $K(\vartheta) \in \mathbb{R}^{n\times n}$, die jeweils positiv definit sind. Durch den Ansatz $x(t) = ve^{\boldsymbol{i}\omega t}$ erhält man mit $\lambda = \omega^2$ das **strukturelle Eigenwertproblem**

$$K(\vartheta)v = \lambda\, M(\vartheta)v. \tag{1.5}$$

Dabei nennt man das zu einem Eigenwert $\lambda > 0$ gehörige $\omega = \sqrt{\lambda}$ eine Eigenfrequenz des ungedämpften dynamischen Systems.

Mit Hilfe der experimentellen Modalanalyse – vergleiche hierzu Natke (1992) – erhält man Messungen der Eigenfrequenzen aus einem bestimmten Frequenzbereich – typischerweise die kleinsten Eigenfrequenzen – und einige Komponenten der zugehörigen Eigenvektoren. Zur Darstellung der zugehörigen Modellfunktion η bezeichnen wir mit

$$\lambda_1(\vartheta) \leq \ldots \leq \lambda_n(\vartheta)$$

die sortierten Eigenwerte des strukturellen Eigenwertproblems (1.5) und mit

$$v_1(\vartheta), \ldots, v_n(\vartheta)$$

zugehörige Eigenvektoren. Es gilt also $K(\vartheta)v_i(\vartheta) = \lambda_i(\vartheta)M(\vartheta)v_i(\vartheta)$ für $i = 1, \ldots, n$. Zu gegebenen Indexmengen $I, L \subseteq \{1, \ldots, n\}$ seien ferner $\lambda_I(\vartheta) = (\lambda_i(\vartheta))_{i\in I}$ der Teilvektor der Komponenten, die zur Indexmenge I gehören, und analog $v_{LI}(\vartheta) = (v_{l,i}(\vartheta))_{l\in L, i\in I}$ die durch L bestimmte Auswahl von Komponenten der zugehörigen Eigenvektoren $v_i(\vartheta) = (v_{1,i}(\vartheta), \ldots, v_{n,i}(\vartheta))^{\mathrm{T}}$, $i \in I$. Dann ist die Modellfunktion gegeben durch

$$\eta(\vartheta) = \begin{pmatrix} \lambda_I(\vartheta) \\ v_{LI}(\vartheta) \end{pmatrix}.$$

1.4. A-priori-Verteilungen

Im Beispiel 1.3.1 ist eine mögliche a-priori-Annahme durch

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu, \tau^2) \tag{1.6}$$

gegeben. Der a-priori-Mittelwert $\mu \in \mathbb{R}$ und die a-priori-Standardabweichung $\tau > 0$ können bekannt oder unbekannt sein.
Im strukturellem Eigenwertproblem 1.5 kann man sich als a-priori-Annahme analog

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu, \tau^2 I_p) \tag{1.7}$$

mit $\mu \in \mathbb{R}^p$ und $\tau > 0$ vorstellen. Interessanter erscheinen uns diejenigen a-priori-Verteilungen, die räumliche Abhängigkeiten zwischen den Parameterkomponenten einbeziehen, etwa

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu, \tau^2 B(\rho)) \tag{1.8}$$

mit $\mu \in \mathbb{R}^p$, $\tau > 0$, $\rho \in (-1, 1)$ und

$$B(\rho) = \big(b_{ij}(\rho)\big)_{1\leq i,j\leq p}$$

mit

$$b_{ij}(\rho) = \rho^{|i-j|} \quad \text{für } i, j = 1, \ldots, p.$$

Solche Ansätze werden in Abschnitt 6.2.5 vorgestellt und eingehend diskutiert.

2. Kleinste-Quadrate-Schätzung im Modell ohne a-priori-Information

Wir geben zuerst eine kurze Zusammenschau der wichtigsten Resultate zur Parameterschätzung im linearen Modell und verweisen für Beweise und weitere Ergebnisse auf die Literatur, z.B. Searle (1982), Rao u. Rao (1998) zur Matrix-Theorie und Seber (1977) und Christensen (2002) zur Theorie im linearen Modell. Anschließend wenden wir uns der Parameterschätzung im nichtlinearen Modell zu. Ergänzungen zur hier vorgestellten Theorie im nichtlinearen Modell finden sich in Pázman (1993), Seber u. Wild (1989) und Bates u. Watts (1988) und viele praktische Beispiele in Fahrmeir u. a. (2007), Ryan (1997), Draper u. Smith (1998) und Box u. Draper (1987). Zuletzt beleuchten wir kurz Tests und Konfidenzbereiche im linearen und nichtlinearen Modell.

2.1. Eigenschaften des Kleinste-Quadrate-Schätzers

2.1.1. Eigenschaften im linearen Modell

Wir betrachten in diesem Abschnitt zu der bekannten Design-Matrix $X \in \mathbb{R}^{N\times p}$, der positiv definiten Kovarianzmatrix $\Sigma \in \mathbb{R}^{N\times N}$ und bekannten Vektoren $y_0 \in \mathbb{R}^N$ und $\vartheta_0 \in \mathbb{R}^p$ das affin-lineare Modell

$$\begin{aligned} \boldsymbol{Y} &\sim \mathcal{N}(\eta(\vartheta), \Sigma) \quad \text{mit} \\ \eta : \mathbb{R}^p \to \mathbb{R}^N,\ \eta(\vartheta) &= y_0 + X \cdot (\vartheta - \vartheta_0). \end{aligned} \tag{2.1}$$

Als Spezialfall ergibt sich für $y_0 = X\vartheta_0$ das lineare Modell

$$\begin{aligned} \boldsymbol{Y} &\sim \mathcal{N}(\eta(\vartheta), \Sigma) \quad \text{mit} \\ \eta : \mathbb{R}^p \to \mathbb{R}^N,\ \eta(\vartheta) &= X\vartheta. \end{aligned} \tag{2.2}$$

Wir studieren hinsichtlich der Kovarianzmatrix $\Sigma \in \mathbb{R}^{N\times N}$ die folgenden beiden Fälle:

A. Σ ist bekannt.

B. $\Sigma = \sigma^2\Sigma_0$ mit einer bekannten positiv definiten Matrix $\Sigma_0 \in \mathbb{R}^{N\times N}$, $\sigma > 0$ ist unbekannt.

Diese Unterscheidung wird im Abschnitt 2.2 zu Tests und Konfidenzbereichen wesentlich sein. Hier betrachten wir zunächst die Schätzung von $\vartheta \in \mathbb{R}^p$.

Bemerkung 2.1 (Modelltransformation)
Mit den affin-linearen Transformationen

$$\tilde{\vartheta} = \vartheta - \vartheta_0, \quad \tilde{\boldsymbol{Y}} = \Sigma_0^{-1/2}(\boldsymbol{Y} - y_0) \quad \text{und} \quad \tilde{X} = \Sigma_0^{-1/2} X$$

erhält man aus dem Modell (2.1) mit bis auf σ bekannter Kovarianzmatrix (Fall $\boldsymbol{B}$) einen Spezialfall von Modell (2.2), nämlich das Modell

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\tilde{X}\tilde{\vartheta}, \sigma^2 I_N), \tag{2.3}$$

wie es meist in der Literatur zu linearen Modellen betrachtet wird. Die Komponenten $\mathbf{e}_i = (\tilde{\boldsymbol{Y}} - \tilde{X}\tilde{\vartheta})_i$, $i = 1, \ldots, N$, des Fehlervektors $\mathbf{e} = (\tilde{\boldsymbol{Y}} - \tilde{X}\tilde{\vartheta})$ sind dann unabhängig und identisch verteilt. Aus den entsprechenden Ergebnissen in der Literatur erhält man durch Rücktransformation die unten aufgeführten Resultate.
Im Fall $\boldsymbol{A}$ der vollständig bekannten Kovarianzmatrix Σ erhält man mit den Transformationen

$$\tilde{\vartheta} = \vartheta - \vartheta_0, \quad \tilde{\boldsymbol{Y}} = \Sigma^{-1/2}(\boldsymbol{Y} - y_0) \quad \text{und} \quad \tilde{X} = \Sigma^{-1/2} X$$

sogar das Modell

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\tilde{X}\tilde{\vartheta}, I_N)$$

und kann ebenfalls wieder die entsprechenden Resultate aus der Literatur übertragen.

Im Fall $\boldsymbol{A}$ der vollständig bekannten Kovarianzmatrix Σ in (2.1) zeigt sich, dass zur Maximierung der Likelihood-Funktion für $\vartheta \in \mathbb{R}^p$ zur Beobachtung $y \in \mathbb{R}^N$ das (gewichtete) Kleinste-Quadrate-Problem

$$\min_{\vartheta \in \mathbb{R}^p} \|y - \eta(\vartheta)\|_\Sigma^2 \tag{2.4}$$

mit

$$\|y - \eta(\vartheta)\|_\Sigma^2 = (y - \eta(\vartheta))^{\mathrm{T}} \Sigma^{-1} (y - \eta(\vartheta))$$

zu lösen ist.[1]
Die Lösungen ϑ^* von (2.4) sind genau die Lösungen der **Normalgleichungen**

$$X^{\mathrm{T}} \Sigma^{-1} X (\vartheta^* - \vartheta_0) = X^{\mathrm{T}} \Sigma^{-1} (y - y_0). \tag{2.5}$$

[1] Im Fall $\boldsymbol{B}$ ist die Maximierung der Likelihood-Funktion für $(\vartheta, \sigma) \in \mathbb{R}^p \times (0, \infty)$ nicht möglich, falls $y \in \eta(\mathbb{R}^p)$. Im Fall $\mathrm{rg}(X) < N$, etwa für $p < N$, ist $\eta(\mathbb{R}^p)$ ein echter affin-linearer Teilraum des $\mathbb{R}^N$ und damit eine Nullmenge unter der Normal-Verteilung. Sieht man von dieser Nullmenge ab, dann stimmen Maximum-Likelihood-Schätzer und Kleinste-Quadrate-Schätzer überein, und sind im Fall $\mathrm{rg}(X) = p$ gegeben durch

$$\hat{\vartheta} = \vartheta_0 + (X^{\mathrm{T}} \Sigma^{-1} X)^{-1} X^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{Y} - y_0) = \vartheta_0 + (X^{\mathrm{T}} \Sigma_0^{-1} X)^{-1} X^{\mathrm{T}} \Sigma_0^{-1} (\boldsymbol{Y} - y_0).$$

Ist $\mathrm{rg}(X) = N \leq p$, so existiert keine Maximum-Likelihood-Schätzung für (ϑ, σ). Alle Lösungen $\hat{\vartheta}(y)$ des Kleinste-Quadrate-Ansatzes,

$$\min_{\vartheta \in \mathbb{R}^p} \|y - \eta(\vartheta)\|_{\Sigma_0}^2 ,$$

erfüllen dann sogar $\eta(\hat{\vartheta}(y)) = y$.

Ist $\mathrm{rg}(X) = p$, dann gibt es genau eine Lösung $\hat{\vartheta}(y) = \vartheta^*$, die durch

$$\hat{\vartheta}(y) = \vartheta_0 + (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}(y - y_0) \tag{2.6}$$

gegeben ist, und

$$\hat{\vartheta} = \hat{\vartheta}(\boldsymbol{Y}) = \vartheta_0 + (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - y_0) \tag{2.7}$$

wird als der (gewichtete) **Kleinste-Quadrate-Schätzer** bezeichnet.
Im Fall $\mathrm{rg}(X) < p$ ist (2.4) nicht eindeutig lösbar; die Menge der Lösungen ϑ^* von (2.4) bzw. (2.5) ist dann gegeben durch

$$\mathcal{L}(y) = \{\vartheta_0 + (X^{\mathrm{T}}\Sigma^{-1}X)^{+}X^{\mathrm{T}}\Sigma^{-1}(y - y_0) + (I_p - (X^{\mathrm{T}}\Sigma^{-1}X)^{+}X^{\mathrm{T}}\Sigma^{-1}X)v : v \in \mathbb{R}^p\}, \tag{2.8}$$

wobei $(X^{\mathrm{T}}\Sigma^{-1}X)^{+}$ die Moore-Penrose-Inverse von $X^{\mathrm{T}}\Sigma^{-1}X$ bezeichnet. Wählt man

$$\hat{\vartheta}_{MP}(y) = \vartheta_0 + (X^{\mathrm{T}}\Sigma^{-1}X)^{+}X^{\mathrm{T}}\Sigma^{-1}(y - y_0) \in \mathcal{L}(y), \tag{2.9}$$

um einen Schätzer festzulegen, so entscheidet man sich unter den Lösungen $\vartheta^* \in \mathcal{L}(y)$ von (2.4) für diejenige, die zusätzlich $\|\vartheta^* - \vartheta_0\|$ minimiert, d.h. in der Euklidischen Norm den geringsten Abstand von ϑ_0 hat.

Wir betrachten noch die allgemeinere Situation, dass man $L\vartheta + a$ schätzen will, wobei $L \in \mathbb{R}^{q\times p}$ und $a \in \mathbb{R}^q$. Dabei heißt $L\vartheta + a$ schätzbar, wenn es für diesen Aspekt einen erwartungstreuen linearen Schätzer gibt. Dies ist genau dann der Fall, wenn $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$ gilt, und beides ist äquivalent zu $L\vartheta_1^* = L\vartheta_2^*$ für alle $\vartheta_1^*, \vartheta_2^* \in \mathcal{L}(y)$ und alle $y \in \mathbb{R}^N$.
In diesem Fall erhält man durch Anwendung des Aspekts auf den Schätzer $\hat{\vartheta}_{MP}$ den **Gauß-Markov-Schätzer**

$$\widehat{L\vartheta + a} = \widehat{L\vartheta + a}(\boldsymbol{Y}) = L\vartheta_0 + L(X^{\mathrm{T}}\Sigma^{-1}X)^{+}X^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - y_0) + a \tag{2.10}$$

für $L\vartheta + a$. Es gilt $\widehat{L\vartheta + a} = \widehat{L\vartheta} + a$ und im Fall $\mathrm{rg}(X) = p$ ferner $\widehat{L\vartheta + a} = L\hat{\vartheta} + a$. Ein Anwendungsfall hiervon ist, dass man $\eta(\vartheta)$ schätzen will, dass also $L = X$ und $a = y_0 - X\vartheta_0$ gilt. Die Bedingung bezüglich der Bildräume ist dann trivialerweise erfüllt und es ergibt sich als Gauß-Markov-Schätzer

$$\widehat{\eta(\vartheta)} = \widehat{\eta(\vartheta)}(\boldsymbol{Y}) = X(X^{\mathrm{T}}\Sigma^{-1}X)^{+}X^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - y_0) + y_0. \tag{2.11}$$

Ferner folgt aus der hinreichenden und notwendigen Bedingung der Bildinklusion, dass im Fall $\mathrm{rg}(X) < p$ der Schätzer $\hat{\vartheta}_{MP} = \hat{\vartheta}_{MP}(\boldsymbol{Y})$ nicht erwartungstreu für ϑ ist.

Für die beiden folgenden Resultate benötigen wir von der Normalverteilungsannahme in Modell (2.1) lediglich die Annahmen über die ersten beiden Momente,[2] d.h.

$$\mathrm{E}(\boldsymbol{Y}) = \eta(\vartheta) \quad \text{und}$$
$$\mathrm{D}(\boldsymbol{Y}) = \Sigma.$$

[2]Hier und in der Folge bezeichnen wir mit $\mathrm{E}_\vartheta(g(\boldsymbol{Y}))$ (bzw. $\mathrm{E}_{\vartheta,\sigma}(g(\boldsymbol{Y}))$ im Fall $\boldsymbol{B}$) den Erwartungswert und mit $\mathrm{D}_\vartheta(g(\boldsymbol{Y}))$ die Varianz der Zufallsvariablen $g(\boldsymbol{Y})$ im Modell $\boldsymbol{Y} \sim \mathcal{N}(\eta(\vartheta), \Sigma)$. Ist der Parameterwert ϑ aus dem Zusammenhang klar, wird er unterdrückt und wir schreiben z.B. $\mathrm{E}(g(\boldsymbol{Y}))$.

Ferner betrachten wir die Klasse Δ der erwartungstreuen affin-linearen Schätzer für $L\vartheta + a$, d.h. im Fall $\boldsymbol{A}$

$$\Delta = \{U\boldsymbol{Y} + b : U \in \mathbb{R}^{q\times N}, b \in \mathbb{R}^q \text{ und } \mathrm{E}_\vartheta(U\boldsymbol{Y} + b) = L\vartheta + a \text{ für alle } \vartheta \in \mathbb{R}^p\}$$

bzw. im Fall $\boldsymbol{B}$

$$\Delta = \{U\boldsymbol{Y} + b : U \in \mathbb{R}^{q\times N}, b \in \mathbb{R}^q \text{ und } \mathrm{E}_{\vartheta,\sigma}(U\boldsymbol{Y} + b) = L\vartheta + a \text{ für alle } \vartheta \in \mathbb{R}^p, \sigma > 0\}.$$

Mit $\leq_{\mathfrak{L}}$ bezeichnen wir die Loewner-Halbordnung auf den symmetrischen Matrizen, d.h. für Matrizen $A, B \in \mathbb{R}^{k\times k}$ gilt $A \leq_{\mathfrak{L}} B$, wenn $B - A$ positiv semidefinit ist.

Satz 2.2 (Satz von Gauß-Markov)
Es gelte $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$. *Dann ist der Gauß-Markov-Schätzer aus* (2.10) *erwartungstreu für* $L\vartheta + a$ *und erfüllt*

$$\mathrm{D}(\widehat{L\vartheta + a}) \leq_{\mathfrak{L}} \mathrm{D}(\delta) \quad \textit{für alle } \delta \in \Delta.$$

Ferner gilt

$$\mathrm{D}(\widehat{L\vartheta + a}) = L(X^{\mathrm{T}}\Sigma^{-1}X)^{+}L^{\mathrm{T}}.$$

Will man für $c \in \mathrm{Bild}(X^{\mathrm{T}})$ einen reellen Aspekt $c^{\mathrm{T}}\vartheta$ schätzen, so hat also der Gauß-Markov-Schätzer $\widehat{c^{\mathrm{T}}\vartheta}$ die kleinste Varianz unter allen affin-linearen erwartungstreuen Schätzern für $c^{\mathrm{T}}\vartheta$.

Unter der Normalverteilungsannahme in (2.1) für $\boldsymbol{Y}$ ist der Gauß-Markov-Schätzer als affin-linearer Schätzer in $\boldsymbol{Y}$ normalverteilt. Es gilt folglich[3]

$$\widehat{L\vartheta + a} \sim \mathcal{N}\left(L\vartheta + a, L(X^{\mathrm{T}}\Sigma^{-1}X)^{+}L^{\mathrm{T}}\right).$$

Korollar 2.3 (Eigenschaften des Kleinste-Quadrate-Schätzers)
Es gelte $\mathrm{rg}(X) = p$.

(i) Der Kleinste-Quadrate-Schätzer $\hat{\vartheta}$ *ist erwartungstreu, d.h. es gilt* $\mathrm{E}_\vartheta(\hat{\vartheta}) = \vartheta$ *im Fall* $\boldsymbol{A}$ *bzw.* $\mathrm{E}_{\vartheta,\sigma}(\hat{\vartheta}) = \vartheta$ *im Fall* $\boldsymbol{B}$.

(ii) Für die Kovarianzmatrix von $\hat{\vartheta}$ *gilt* $\mathrm{D}(\hat{\vartheta}) = (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}$.

(iii) Unter der Normalverteilungsannahme in (2.1) *ist der Kleinste-Quadrate-Schätzer normalverteilt, d.h. es gilt*

$$\hat{\vartheta} \sim \mathcal{N}\left(\vartheta, (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}\right).$$

[3] Im Fall $\mathrm{rg}(L) < q$ ist die Normalverteilung entartet.

Die Verteilungsfamilie $(\mathcal{N}(\eta(\vartheta),\Sigma))_{\vartheta\in\mathbb{R}^p}$ des Modells (2.1) bei vollständig bekannter Kovarianzmatrix Σ (Fall $\boldsymbol{A}$) ist eine p-parametrige Exponentialfamilie in

$$T:(\mathbb{R}^N,\mathcal{B}^N)\to(\mathbb{R}^p,\mathcal{B}^p),\ T(y)=X^{\mathrm{T}}\Sigma^{-1}y$$

mit ϑ als natürlichem Parameter.
Ist die Kovarianzmatrix $\Sigma=\sigma^2\Sigma_0$ nur bis auf σ bekannt (Fall $\boldsymbol{B}$), dann ist die Verteilungsfamilie $\big(\mathcal{N}(\eta(\vartheta),\sigma^2\Sigma_0)\big)_{(\vartheta,\sigma)\in\mathbb{R}^p\times(0,\infty)}$ eine $(p+1)$-parametrige Exponentialfamilie in

$$T:(\mathbb{R}^N,\mathcal{B}^N)\to(\mathbb{R}^{p+1},\mathcal{B}^{p+1}),\ T(y)=\begin{pmatrix}X^{\mathrm{T}}\Sigma_0^{-1}y\\(-y_0+X\vartheta_0+y)^{\mathrm{T}}\Sigma_0^{-1}y\end{pmatrix}$$

mit natürlichem Parameter

$$\beta(\vartheta,\sigma)=\begin{pmatrix}\frac{1}{\sigma^2}\,\vartheta\\-\frac{1}{\sigma^2}\end{pmatrix}.$$

In beiden Fällen ist die Statistik $T(\boldsymbol{Y})$ suffizient und vollständig. Folglich sind der Kleinste-Quadrate-Schätzer $\hat{\vartheta}$ im Fall $\mathrm{rg}(X)=p$ bzw. allgemeiner der Gauß-Markov-Schätzer $\widehat{L\vartheta+a}$ im Fall $\mathrm{Bild}(L^{\mathrm{T}})\subseteq\mathrm{Bild}(X^{\mathrm{T}})$ als erwartungstreue Schätzer basierend auf $T(\boldsymbol{Y})$ nach dem Satz von Lehmann-Scheffé optimal bezüglich jeder konvexen Verlustfunktion in der Klasse der erwartungstreuen Schätzer für ϑ bzw. $L\vartheta+a$. Insbesondere gilt bei quadratischem Verlust:

$$\mathrm{D}(\hat{\vartheta})\leq_{\mathfrak{L}}\mathrm{D}(\delta(\boldsymbol{Y}))$$

für alle $\delta:(\mathbb{R}^N,\mathcal{B}^N)\to(\mathbb{R}^p,\mathcal{B}^p)$ mit $\mathrm{E}_\vartheta(\delta(\boldsymbol{Y}))=\vartheta$ für jedes $\vartheta\in\mathbb{R}^p$ im Fall $\boldsymbol{A}$ bzw. mit $\mathrm{E}_{\vartheta,\sigma}(\delta(\boldsymbol{Y}))=\vartheta$ für jedes $\vartheta\in\mathbb{R}^p$ im Fall $\boldsymbol{B}$ und

$$\mathrm{D}(\widehat{L\vartheta+a})\leq_{\mathfrak{L}}\mathrm{D}(\delta(\boldsymbol{Y}))$$

für alle $\delta:(\mathbb{R}^N,\mathcal{B}^N)\to(\mathbb{R}^q,\mathcal{B}^q)$ mit $\mathrm{E}_\vartheta(\delta(\boldsymbol{Y}))=L\vartheta+a$ für jedes $\vartheta\in\mathbb{R}^p$ im Fall $\boldsymbol{A}$ bzw. mit $\mathrm{E}_{\vartheta,\sigma}(\delta(\boldsymbol{Y}))=L\vartheta+a$ für jedes $\vartheta\in\mathbb{R}^p$ im Fall $\boldsymbol{B}$.

Abschließend diskutieren wir noch eine geometrische Interpretation der Kleinste-Quadrate-Schätzung, vergleiche auch Abbildung 2.1. Die Matrix

$$\mathcal{P}_{X,\Sigma}=X(X^{\mathrm{T}}\Sigma^{-1}X)^{+}X^{\mathrm{T}}\Sigma^{-1}$$

ist orthogonaler Projektor auf $\mathrm{Bild}(X)$ bezüglich des von Σ induzierten Skalarprodukts, d.h. es gilt $\mathrm{Bild}(\mathcal{P}_{X,\Sigma})=\mathrm{Bild}(X)$, $\mathcal{P}_{X,\Sigma}^2=\mathcal{P}_{X,\Sigma}$ und $<x,\mathcal{P}_{X,\Sigma}y>_\Sigma=<\mathcal{P}_{X,\Sigma}x,y>_\Sigma$ für $x,y\in\mathbb{R}^N$, wobei $<x,y>_\Sigma=x^{\mathrm{T}}\Sigma^{-1}y$. Abkürzend schreiben wir hierbei

$$\mathcal{P}_X=\mathcal{P}_{X,I_N}.$$

Nach (2.11) ist daher $\widehat{\eta(\vartheta)}(y)-y_0=\mathcal{P}_{X,\Sigma}(y-y_0)$ die $<,>_\Sigma$-orthogonale Projektion von $y-y_0$ auf $\mathrm{Bild}(X)$.
Man kann sich also die Kleinste-Quadrate-Schätzung als zweistufigen Prozess vorstellen:

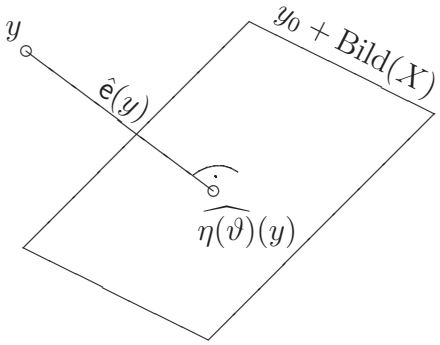


Abbildung 2.1.: Orthogonale Projektion auf die Erwartungswertmenge $y_0 + \mathrm{Bild}(X)$

Zuerst wird $y - y_0$ mittels $\mathcal{P}_{X,\Sigma}$ ($<,>_\Sigma$-orthogonal) auf $\mathrm{Bild}(X)$ projiziert.
Im Fall $\mathrm{rg}(X) = p$ ist dann im zweiten Schritt das zu dieser Projektion $\mathcal{P}_{X,\Sigma}(y - y_0)$ zugehörige ϑ, die Kleinste-Quadrate-Schätzung, eindeutig bestimmt, d.h. $\mathcal{P}_{X,\Sigma}(y-y_0) = X(\vartheta - \vartheta_0)$ ist eindeutig lösbar nach ϑ.
Der **Residuenvektor** zu dieser Projektion

$$\hat{\mathsf{e}}(y) = y - \widehat{\eta(\vartheta)}(y) = y - y_0 - \mathcal{P}_{X,\Sigma}(y - y_0) = (I_N - \mathcal{P}_{X,\Sigma})(y - y_0) \tag{2.12}$$

erfüllt

$$\|\hat{\mathsf{e}}(y)\|_\Sigma = \min_{z \in \mathrm{Bild}(X)} \|y - (z + y_0)\|_\Sigma = \min_{z \in y_0 + \mathrm{Bild}(X)} \|y - z\|_\Sigma .$$

Dabei ist $I_N - \mathcal{P}_{X,\Sigma}$ ein $<,>_\Sigma$-orthogonaler Projektor auf

$$\mathrm{Bild}(X)^{\perp_\Sigma} = \{y \in \mathbb{R}^N :< x, y >_\Sigma = 0 \text{ für alle } x \in \mathrm{Bild}(X)\},$$

den Orthogonalraum zu $\mathrm{Bild}(X)$ bezüglich des von Σ induzierten Skalarprodukts.

2.1.2. Eigenschaften im nichtlinearen Modell

In diesem Abschnitt betrachten wir die Schätzung von ϑ im Modell (1.1), also

$$\boldsymbol{Y} \sim \mathcal{N}(\eta(\vartheta), \Sigma)$$

mit $\vartheta \in \Theta \subseteq \mathbb{R}^p$, $\Theta \neq \emptyset$, und (nichtlinearer) Modellfunktion

$$\eta : \Theta \to \mathbb{R}^N,$$

und wir betrachten bei der Kovarianzmatrix $\Sigma \in \mathbb{R}^{N \times N}$ wieder die beiden Fälle:

A. Σ ist bekannt.

B. $\Sigma = \sigma^2 \Sigma_0$ mit einer bekannten positiv definiten Matrix $\Sigma_0 \in \mathbb{R}^{N \times N}$, $\sigma > 0$ ist unbekannt.

Bemerkung 2.4 (Modelltransformation)
Mit den linearen Transformationen

$$\tilde{\boldsymbol{Y}} = \Sigma_0^{-1/2}\boldsymbol{Y} \quad \text{und} \quad \tilde{\eta} = \Sigma_0^{-1/2}\eta$$

erhält man im Fall $\boldsymbol{B}$ das Modell

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\tilde{\eta}(\vartheta), \sigma^2 I_N) \tag{2.13}$$

mit unabhängigen und identisch-verteilten Fehlern $\mathbf{e}_i = \tilde{\boldsymbol{Y}}_i - \tilde{\eta}_i(\vartheta)$, $i = 1, \ldots, N$, wie es meist in der Literatur betrachtet wird.
Im Fall $\boldsymbol{A}$ der vollständig bekannten Kovarianzmatrix erhält man mit den linearen Transformationen

$$\tilde{\boldsymbol{Y}} = \Sigma^{-1/2}\boldsymbol{Y} \quad \text{und} \quad \tilde{\eta} = \Sigma^{-1/2}\eta$$

sogar das Modell

$$\tilde{\boldsymbol{Y}} \sim \mathcal{N}(\tilde{\eta}(\vartheta), I_N). \tag{2.14}$$

Die Verteilungsfamilie $(\mathcal{N}(\eta(\vartheta), \Sigma))_{\vartheta\in\Theta}$ unseres Modells (1.1) ist bei vollständig bekannter Kovarianzmatrix Σ, also im Fall $\boldsymbol{A}$, eine N-parametrige Exponentialfamilie in der Statistik

$$T : (\mathbb{R}^N, \mathcal{B}^N) \to (\mathbb{R}^N, \mathcal{B}^N),\ T(y) = \Sigma^{-1}y,$$

wobei der natürliche Parameter durch $\beta(\vartheta) = \eta(\vartheta)$ gegeben ist.
Ist im Fall $\boldsymbol{B}$ die Kovarianzmatrix $\Sigma = \sigma^2\Sigma_0$ nur bis auf σ bekannt, dann ist die Verteilungsfamilie $\big(\mathcal{N}(\eta(\vartheta), \sigma^2\Sigma_0)\big)_{(\vartheta,\sigma)\in\Theta\times(0,\infty)}$ eine $(N+1)$-parametrige Exponentialfamilie in der Statistik

$$T : (\mathbb{R}^N, \mathcal{B}^N) \to (\mathbb{R}^{N+1}, \mathcal{B}^{N+1}),\ T(y) = \begin{pmatrix} \Sigma_0^{-1}y \\ y^{\mathrm{T}}\Sigma_0^{-1}y \end{pmatrix},$$

wobei der natürliche Parameter durch

$$\beta(\vartheta, \sigma) = \begin{pmatrix} \frac{1}{\sigma^2}\,\eta(\vartheta) \\ -\frac{1}{\sigma^2} \end{pmatrix}$$

gegeben ist.

Im Allgemeinen sind im nichtlinearen Modell die Voraussetzungen für den Satz von Lehmann-Scheffé nicht erfüllt. Zwei Probleme treten auf:
Zwar ist die Statistik $T(\boldsymbol{Y})$ suffizient, aber im Allgemeinen nicht mehr vollständig. Man beachte hierbei, dass die hinreichende Bedingung, dass $\beta(\Theta)$ bzw. $\beta(\Theta \times (0,\infty))$ nichtleeres Inneres besitzt, im Allgemeinen nicht mehr gegeben ist.
Weiterhin existiert im Allgemeinen kein erwartungstreuer Schätzer für ϑ, wie wir im folgenden Beispiel erörtern.

Beispiel 2.5

a) Sei $\Theta = [0, \pi]$, $\eta : \Theta \to \mathbb{R}^2$, $\eta(\vartheta) = (\cos(\vartheta), \sin(\vartheta))^{\mathrm{T}}$ und $\Sigma = I_2$.
Naheliegend ist, sich hier auf diejenigen Schätzer $g(\boldsymbol{Y})$ für $\vartheta \in \Theta$ einzuschränken, bei denen $g(\mathbb{R}^2) \subseteq [0, \pi]$ gilt. In dieser Klasse gibt es keinen erwartungstreuen Schätzer für $\vartheta \in \Theta$:
Denn damit $g(\boldsymbol{Y})$ erwartungstreu für ϑ ist, muss $\mathrm{E}_\vartheta(g(\boldsymbol{Y})) = 0$ insbesondere für $\vartheta = 0$ gelten und, da $g \geq 0$, folglich $g = 0$ Lebesgue-fast-überall gelten. Damit wäre aber $\mathrm{E}_\vartheta(g(\boldsymbol{Y})) \neq \vartheta$ für $\vartheta \neq 0$.[4]
Mathematisch unbefriedigend an diesem Beispiel ist die Einschränkung der Klasse der erlaubten Schätzer, was aber in praktischen Anwendungen oft notwendig ist, um physikalisch sinnvolle Schätzungen zu erhalten. Einige kurze Überlegungen dazu, ob es ohne die Einschränkung der Klasse der Schätzer einen erwartungstreuen Schätzer gibt, finden sich im Anhang A.1.[5]

b) Sei $\Theta = \mathbb{R} \setminus \{0\}$, $\eta : \Theta \to \mathbb{R}^N$, $\eta(\vartheta) = 1/\vartheta \cdot \mathbf{1}_N$ und $\Sigma = I_N$. Dann gibt es keinen erwartungstreuen Schätzer für ϑ, wie wir in Anhang A.1 zeigen. Der Grund ist der Pol der Modellfunktion. Betrachtet man stattdessen $\Theta = (0, \infty)$, dann gibt es erwartungstreue Schätzer, vergleiche ebenfalls Anhang A.1.

Im Fall $\boldsymbol{A}$ der vollständig bekannten Kovarianzmatrix ist im Modell (1.1) die Maximierung der Likelihood-Funktion für $\vartheta \in \Theta$ zur Beobachtung y äquivalent zum (gewichteten) Kleinste-Quadrate-Problem

$$\min_{\vartheta \in \Theta} \|y - \eta(\vartheta)\|_\Sigma^2 . \tag{2.15}$$

Im Fall $\boldsymbol{B}$ ist die Maximierung der Likelihood-Funktion auf $\Theta \times (0, \infty)$ für $y \in \eta(\Theta)$ nicht möglich. Im Allgemeinen ist die Bedingung $p < N$ nicht mehr hinreichend dafür, dass die Menge $\eta(\Theta)$ eine Lebesgue-Nullmenge ist. Ein Gegenbeispiel wäre etwa bei einer Bijektion $\eta : \mathbb{R}^p \to \mathbb{R}^N$ gegeben. Eine hinreichende Bedingung liefert:

Satz 2.6
Sei $U \subseteq \mathbb{R}^n$ offen und $f : U \to \mathbb{R}^n$ stetig differenzierbar auf U. Ist $A \subseteq U$ eine Lebesgue-Nullmenge im R^n, so ist auch $f(A)$ eine Lebesgue-Nullmenge im R^n.

Beweis: siehe Forster (1984, § 7, Satz 6) □

Dies können wir nun wie folgt anwenden: Ist $\Theta \subseteq \mathbb{R}^p$ offen mit $p < N$ und $\eta : \Theta \to \mathbb{R}^N$ stetig differenzierbar, dann ist auf $f : \Theta \times \mathbb{R}^{N-p} \to \mathbb{R}^N$, $f(\vartheta, z) = \eta(\vartheta)$ und $A = \Theta \times \{\mathbf{0}_{N-p}\}$ der Satz anwendbar und man kann folgern, dass $\eta(\Theta) = f(A)$ eine Lebesgue-Nullmenge ist.

[4] Diese Überlegungen sind nicht spezifisch für die betrachtete Modellfunktion und übertragen sich auf andere Fälle mit eingeschränktem Parameterbereich und der entsprechend eingeschränkten Klasse von Schätzern, z.B. für $\Theta = [0, 1]$, $\tilde{\eta} : \Theta \to \mathbb{R}^2$, $\eta(\vartheta) = \left(\vartheta, \vartheta^2\right)^{\mathrm{T}}$.

[5] Im Modell aus Fußnote 4 ist der Sachverhalt einfach: Dort ist mit $g : \mathbb{R}^2 \to \mathbb{R}$, $g(x_1, x_2) = x_1$ ein erwartungstreuer Schätzer $g(\boldsymbol{Y})$ gegeben, wenn man auf die Einschränkung verzichtet.

Ist Θ nicht offen, so kann man oft η stetig differenzierbar auf eine offene Obermenge von Θ fortsetzen oder muss den Rand von Θ gesondert behandeln.
Bei $p < N$ ist daher in praktischen Anwendungen $\eta(\Theta)$ meist eine Lebesgue-Nullmenge und folglich eine Nullmenge unter der Normalverteilung. In diesem Fall ist daher ein Schätzer definiert durch Lösen des Kleinste-Quadrate-Problems

$$\min_{\vartheta \in \Theta} \|y - \eta(\vartheta)\|^2_{\Sigma_0} \tag{2.16}$$

(wie in (2.15) nur mit Σ_0 anstelle von Σ), abgesehen von obiger Nullmenge, auch ein Maximum-Likelihood-Schätzer.

Wir wenden uns nun dem Fall $p \geq N$ zu. Hier und im Folgenden bezeichnen wir zu einer in einem Punkt $\vartheta \in \Theta$ differenzierbaren Abbildung $\eta : \Theta \to \mathbb{R}^N$ mit $J_\eta(\vartheta) \in \mathbb{R}^{N \times p}$ die Jacobi-Matrix von η in $\vartheta \in \Theta$.

Lemma 2.7
Sei $\vartheta \in \Theta$ und $p \geq N$. Ferner sei $\Theta_1 \subseteq \Theta$ eine offene Umgebung von ϑ und η stetig differenzierbar[6] auf Θ_1 und $\operatorname{rg}(J_\eta(\vartheta)) = N$, d.h. $J_\eta(\vartheta)$ ist surjektiv. Dann gibt es $U \subseteq \Theta_1$ mit $\vartheta \in U$, so dass $\eta(U)$ offen im $\mathbb{R}^N$ ist.

Beweis: Da $\operatorname{rg}(J_\eta(\vartheta)) = N$ gibt es Indizes $i_1, \ldots, i_N$, so dass $\left(\frac{\partial \eta}{\partial \vartheta_{i_1}}(\vartheta), \ldots, \frac{\partial \eta}{\partial \vartheta_{i_N}}(\vartheta)\right)$ invertierbar ist. Zur Vereinfachung der Notation nehmen wir o.B.d.A. an, dass $i_1 = 1, \ldots, i_N = N$. Ferner sei im Fall $p > N$ entsprechend $\vartheta = (x, z)$ mit $x \in \mathbb{R}^N$, $z \in \mathbb{R}^{p-N}$. Dann ist $\tilde{\Theta}_1 = \{u \in \mathbb{R}^N : (u, z) \in \Theta_1\}$ offen und die Abbildung $\tilde{\eta} : \tilde{\Theta}_1 \to \mathbb{R}^N$, $\tilde{\eta}(u) = \eta(u, z)$ ist stetig differenzierbar auf $\tilde{\Theta}_1$ und $J_{\tilde{\eta}}(x)$ ist invertierbar. (Im Fall von $p = N$ wählt man $\tilde{\Theta}_1 = \Theta_1$ und $\tilde{\eta} = \eta$.) Nach dem Satz über Umkehrabbildungen, vgl. z.B. Rudin (1998), gibt es dann eine offene Menge $\tilde{U} \subseteq \tilde{\Theta}_1$ mit $x \in \tilde{U}$, so dass $\tilde{\eta}$ injektiv auf $\tilde{U}$ und $\tilde{\eta}(\tilde{U}) \subseteq \mathbb{R}^N$ offen ist. Im Fall $p > N$ erhält man mit $U = \tilde{U} \times \{z\}$ und im Fall $p = N$ mit $U = \tilde{U}$ die Behauptung. □

Im Fall $p \geq N$ ist daher in der Regel $\eta(\Theta)$ keine Lebesgue-Nullmenge und der Maximum-Likelihood-Schätzer auf einer Menge mit positiver Wahrscheinlichkeit nicht definiert. Anders gesagt sind Lösungen des Kleinste-Quadrate-Problems (2.16) mit positiver Wahrscheinlichkeit keine Maximum-Likelihood-Schätzungen.

Wir werden im weiteren Verlauf unsere Betrachtungen am Fall ***A*** der vollständig bekannten Kovarianzmatrix Σ ausrichten. Im Fall ***B*** wäre im Folgenden entsprechend Σ durch Σ_0 zu ersetzen.

Wir wenden uns nun der Frage zu, für welche η und welche Beobachtungen y es überhaupt eine Lösung des Kleinste-Quadrate-Problems (2.15) gibt bzw. ob diese eindeutig ist.

[6] Man kann die Voraussetzungen noch weiter abschwächen, indem man nur fordert, dass die im Beweis definierte Funktion $\tilde{\eta}$ stetig differenzierbar ist. Auch die Stetigkeit der Ableitung benötigt man nur in einem Punkt, vgl. Nijenhuis (1974).

Man kann analog zur geometrischen Interpretation im linearen Fall (siehe S. 8) auch im nichtlinearen Modell die Lösung des Kleinste-Quadrate-Problems (2.16) in zwei Schritte zerlegen. Im ersten Schritt bestimmt man eine Lösung z^* von

$$\min_{z\in\mathcal{E}} \|y-z\|_\Sigma^2 \tag{2.17}$$

mit der sogenannten „Erwartungswertmenge“

$$\mathcal{E} = \eta(\Theta).$$

Eine solche Lösung nennt man auch ein **Proximum** an y in $\mathcal{E}$ bezüglich der durch Σ induzierten Norm $\|\cdot\|_\Sigma$. Zu dieser Lösung z^* ist dann im zweiten Schritt ein entsprechendes ϑ^* mit $\eta(\vartheta^*) = z^*$ zu finden.

Ist z^* eindeutig, so ist die Injektivität[7] von η eine hinreichende Bedingung für die Eindeutigkeit von ϑ^*. Ferner ist die Existenz von ϑ^* nach Definition von $\mathcal{E}$ gewährleistet. Damit bleibt noch die Untersuchung der Existenz und Eindeutigkeit einer Lösung von (2.17).

Zuerst zur Existenz.

Lemma 2.8
Ist $\mathcal{E}$ abgeschlossen, dann existiert für jedes $y \in \mathbb{R}^N$ ein Proximum in (2.17).

Beweis: Wir betrachten eine Folge $(z_n)_{n\in\mathbb{N}}$ in $\mathcal{E}$ mit

$$\lim_{n\to\infty} \|y-z_n\|_\Sigma = \inf_{z\in\mathcal{E}} \|y-z\|_\Sigma =: E_{\mathcal{E}}(y).$$

Dann gibt es ein $n_0 \in \mathbb{N}$ mit

$$\|y-z_n\|_\Sigma \le E_{\mathcal{E}}(y) + 1$$

für $n \ge n_0$. Folglich ist

$$\|z_n\|_\Sigma \le \|y-z_n\|_\Sigma + \|y\|_\Sigma \le E_{\mathcal{E}}(y) + 1 + \|y\|_\Sigma =: K$$

für $n \ge n_0$. Daher gilt mit $B_K = \{z \in \mathbb{R}^N : \|z\|_\Sigma \le K\}$, dass

$$\inf_{z\in\mathcal{E}} \|y-z\|_\Sigma^2 = \inf_{z\in\mathcal{E}\cap B_K} \|y-z\|_\Sigma^2 .$$

Da die Abbildung $S : \mathcal{E}\cap B_K \to \mathbb{R}$, $S(z) = \|y-z\|_\Sigma$ stetig ist und $\mathcal{E}\cap B_K$ kompakt ist, folgt die Existenz eines Proximums. □

Korollar 2.9
Falls Θ kompakt ist und η stetig, dann existiert für jedes $y \in \mathbb{R}^N$ ein Proximum in (2.17).

[7] Ist η von komplizierter Bauart und z.B. nur implizit über die Lösung einer Differentialgleichung oder eines Eigenwertproblems gegeben, so kann eine Prüfung auf Injektivität schwer oder unmöglich sein. Einen numerischen Ansatz, um auf Injektivität zu prüfen, liefern Winterfors u. Curtis (2008).

Beweis: Unter den Voraussetzungen ist dann $\mathcal{E} = \eta(\Theta)$ kompakt, also insbesondere abgeschlossen und wir können Lemma 2.8 anwenden. □

Statt der Abgeschlossenheit von $\mathcal{E}$ in Lemma 2.8 genügt auch, dass es ein $c > 0$ gibt, so dass

$$\mathcal{E}_c(y) = \{z \in \mathcal{E} : \|y - z\|_\Sigma \leq c\}$$

nichtleer und abgeschlossen ist, denn dann gilt

$$\arg\min_{z \in \mathcal{E}} \|y - z\|_\Sigma^2 = \arg\min_{z \in \mathcal{E}_c(y)} \|y - z\|_\Sigma^2 .$$

Hier beschränkt man sich bei der Erwartungswertmenge auf relevante Punkte. Beschränkt man sich hingegen bei der Parametermenge Θ auf relevante Punkte, so führt dies zum Begriff des Existenzlevels, vgl. Demidenko (1989, 2008).

Wir führen diesen Begriff allgemein ein, obwohl wir ihn in diesem Kapitel nur bei gegebenem $y \in \mathbb{R}^N$ für die Funktion

$$S^{\eta,y} : \Theta \to \mathbb{R},\ S^{\eta,y}(\vartheta) = \|y - \eta(\vartheta)\|_\Sigma^2 .$$

benötigen.

Definition 2.10
Für $f : \Theta \to \mathbb{R}$ bezeichne

$$f_E = \lim_{r \to \infty} \inf_{\substack{\|\vartheta\| > r \\ \vartheta \in \Theta}} f(\vartheta),$$

wobei $\inf \emptyset = \infty$. Ist Θ abgeschlossen, dann nennt man f_E das (obere) **Existenzlevel** zu f in Θ.

Wir weisen darauf hin, dass diese Definition des Existenzlevels nur im Fall $\Theta = \mathbb{R}^p$ mit der Definition von Demidenko (1989, 2008) übereinstimmt, vgl. die Diskussion vor Definition 2.13.

Lemma 2.11
Ist Θ abgeschlossen, $f : \Theta \to \mathbb{R}$ stetig und gibt es ein $\vartheta_0 \in \Theta$ mit

$$f(\vartheta_0) < f_E,$$

so existiert ein Minimumspunkt von f auf Θ.

Beweis: Da $f(\vartheta_0) < f_E$, existiert ein $r_0 > 0$ mit

$$f(\vartheta_0) < \inf_{\substack{\|\vartheta\| > r_0 \\ \vartheta \in \Theta}} f(\vartheta)$$

Daher ist

$$\inf_{\vartheta \in \Theta} f(\vartheta) = \inf_{\vartheta \in \Theta^*} f(\vartheta)$$

mit

$$\Theta^* = \Theta \cap \{\vartheta \in \mathbb{R}^p : \|\vartheta\| \leq r_0\},$$

einer beschränkten und abgeschlossenen und damit kompakten Teilmenge des $\mathbb{R}^p$. Somit existiert ein Minimum von f auf Θ. □

Korollar 2.12
Ist Θ abgeschlossen, η stetig, $y \in \mathbb{R}^N$ und gibt es ein $\vartheta_0 \in \Theta$ mit

$$\|y - \eta(\vartheta_0)\|_\Sigma^2 < S_E^{\eta,y},$$

so existiert ein Proximum in $\mathcal{E}$ an y.

Beweis: Ist η stetig, so ist auch $S^{\eta,y}$ stetig, und man kann Lemma 2.11 anwenden und die Existenz eines Minimums von $S^{\eta,y}$ in einem Punkt ϑ^* und damit eines Proximums $z = \eta(\vartheta^*)$ folgern. □

Demidenko (1989, 2006, 2008) definiert das (obere) Existenzlevel für Funktionen $f : \Theta \to \mathbb{R}$ wie folgt:

$$f_E^* = \inf_{\substack{(\vartheta_k)_{k\in\mathbb{N}} \subseteq \Theta \\ \vartheta_k \overset{k\to\infty}{\longrightarrow} \partial\Theta}} \liminf_{k\to\infty} f(\vartheta_k).$$

Die Begriffsbildung ist hierbei unvollständig. Versteht man den Rand $\partial\Theta$ in topologischem Sinne, also $\partial\Theta = \mathrm{cl}(\Theta) \setminus \mathrm{int}(\Theta)$, dann gilt eine Lemma 2.11 entsprechende Aussage nicht mehr, wie man an der Funktion $S^{\eta,y}$ zur Modellfunktion $\eta : [0;\infty) \to \mathbb{R}$, $\eta(\vartheta) = e^{-\vartheta}$ und einem beliebigem $y \leq 0$ sieht. Daher sollten bei der Infimum-Bildung auch Folgen in Θ mit uneigentlichen Grenzwerten betrachtet werden, wie dies in Mäkeläinen u. a. (1981) als Sprechweise vereinbart wird: Im Fall einer offenen Menge Θ heißt dort eine Folge $(\vartheta_k)_{k\in\mathbb{N}}$ konvergent gegen den Rand $\partial\Theta$ von Θ, wenn es für jede kompakte Menge $K \subseteq \Theta$ ein $k_0 \in \mathbb{N}$ gibt, so dass $\vartheta_k \notin K$ für alle $k \geq k_0$. Angelehnt daran erweitern wir Definition 2.10 auf den Fall, dass Θ nicht abgeschlossen ist, wobei wir den Rand $\partial\Theta = \mathrm{cl}(\Theta) \setminus \mathrm{int}(\Theta)$ im üblichen topologischen Sinne verstehen.

Definition 2.13
Sei $f : \Theta \to \mathbb{R}$. Dann heißt

$$f_{E,\partial} = \min\left\{ f_E, \inf_{\substack{(\vartheta_k)_{k\in\mathbb{N}} \subseteq \Theta \\ \lim_{k\to\infty} \vartheta_k \in \partial\Theta \setminus \Theta}} \liminf_{k\to\infty} f(\vartheta_k) \right\}$$

(wobei $\inf \emptyset = \infty$) das (obere) **Existenzlevel** zu f in Θ.

Ist Θ abgeschlossen, dann ist $\partial\Theta \setminus \Theta$ leer und es gilt $f_{E,\partial} = f_E$. Unter Verwendung von $f_{E,\partial}$ an Stelle von f_E gilt ferner die Aussage von Lemma 2.11 auch ohne die Voraussetzung, dass Θ abgeschlossen ist.

Lemma 2.14
Ist $f : \Theta \to \mathbb{R}$ stetig und gibt es ein $\vartheta_0 \in \Theta$ mit

$$f(\vartheta_0) < f_{E,\partial},$$

so existiert ein Minimumspunkt von f auf Θ.

Beweis: Setzt man

$$\Theta^* = \{\vartheta \in \Theta : f(\vartheta) \leq f(\vartheta_0)\},$$

so gilt

$$\inf_{\vartheta \in \Theta} f(\vartheta) = \inf_{\vartheta \in \Theta^*} f(\vartheta).$$

Zum einen ist Θ^* beschränkt:
Ansonsten gäbe es eine Folge $(\vartheta_k)_{k\in\mathbb{N}}$ in Θ mit $\|\vartheta_k\| \to \infty$ und $f(\vartheta_k) \leq f(\vartheta_0)$ für $k \in \mathbb{N}$. Insbesondere gilt dann $f_E \leq f(\vartheta_0)$, was aber im Widerspruch zu $f(\vartheta_0) < f_{E,\partial} \leq f_E$ steht.
Zum anderen ist Θ^* abgeschlossen:
Ansonsten gäbe es eine Folge $(\vartheta_k)_{k\in\mathbb{N}}$ in Θ mit $\lim_{k\to\infty} \vartheta_k \in \partial\Theta \setminus \Theta$ und $f(\vartheta_k) \leq f(\vartheta_0)$ für $k \in \mathbb{N}$. Dann gilt

$$\inf_{\substack{(\vartheta_k)_{k\in\mathbb{N}} \subseteq \Theta \\ \lim_{k\to\infty} \vartheta_k \in \partial\Theta\setminus\Theta}} \liminf_{k\to\infty} f(\vartheta_k) \leq f(\vartheta_0),$$

was im Widerspruch zu

$$f(\vartheta_0) < f_{E,\partial} \leq \inf_{\substack{(\vartheta_k)_{k\in\mathbb{N}} \subseteq \Theta \\ \lim_{k\to\infty} \vartheta_k \in \partial\Theta\setminus\Theta}} \liminf_{k\to\infty} f(\vartheta_k)$$

steht.
Daher ist Θ^* eine beschränkte und abgeschlossene und damit kompakte Teilmenge des $\mathbb{R}^p$ und es existiert ein Minimum von f auf Θ^* und daher auch auf Θ. □

Korollar 2.15
Sei $y \in \mathbb{R}^N$. Ist $\eta : \Theta \to \mathbb{R}^N$ stetig und gibt es ein $\vartheta_0 \in \Theta$ mit

$$\|y - \eta(\vartheta_0)\|_\Sigma^2 < S_{E,\partial}^{\eta,y},$$

so existiert ein Proximum in $\mathcal{E}$ an y.

Bemerkung 2.16
Um Lemma 2.14 (oder analog Korollar 2.15) anwenden zu können, genügt es trivialerweise, für das (obere) Existenzlevel eine Abschätzung $f_E^{\text{low}} \leq f_{E,\partial}$ nach unten und einen Punkt ϑ_0 mit $f(\vartheta_0) < f_E^{\text{low}}$ zu haben. Man ist also nicht unbedingt gezwungen, das Existenzlevel exakt zu bestimmen.

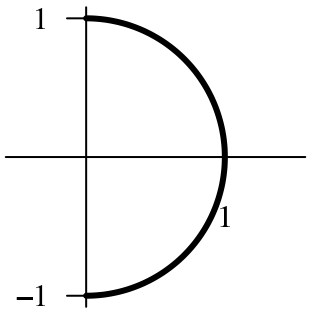


Abbildung 2.2.: Der Halbkreis $\{z \in \mathbb{R}^2 : \|z\| = 1,\ z_1 \geq 0\}$

Deterministische Minimierungsverfahren liefern in der Regel ausgehend von einem Startpunkt ϑ_0 eine Folge $(\vartheta_k)_{k\in\mathbb{N}_0}$ mit $f(\vartheta_{k+1}) \leq f(\vartheta_k)$ für $k \in \mathbb{N}_0$. Gilt im Startpunkt ϑ_0

$$f(\vartheta_0) < \min\left\{ f_E,\ \inf_{\substack{(\vartheta_k)_{k\in\mathbb{N}} \subseteq \Theta \\ \lim_{k\to\infty} \vartheta_k \in \partial\Theta}} \liminf_{k\to\infty} f(\vartheta_k) \right\}, \tag{2.18}$$

also betrachtet man in der rechten Infimumsbildung auch Folgen mit Grenzwerten in $\partial\Theta$ und nicht nur Folgen mit Grenzwerten in $\partial\Theta \setminus \Theta$, so kann man sicherstellen, dass die Folge im Inneren von Θ bleibt und nicht gegen einen Randpunkt konvergiert und entgeht so numerischen Problemen. Insbesondere wird unter der Bedingung (2.18) das globale Minimum in einem inneren Punkt von Θ angenommen.

Wir illustrieren die obigen Sachverhalte zur Existenz an einem Beispiel.

Beispiel 2.17

a) Sei $\Theta = [-\pi/2, \pi/2]$ und

$$\eta : \Theta \to \mathbb{R}^2,\ \eta(\vartheta) = \begin{pmatrix} \cos(\vartheta) \\ \sin(\vartheta) \end{pmatrix}$$

sowie $\Sigma = I_2$. Dann ist

$$\mathcal{E} = \eta(\Theta) = \{z \in \mathbb{R}^2 : \|z\| = 1,\ z_1 \geq 0\}$$

(vgl. Abbildung 2.2), und es gilt

$$\operatorname*{arg\,min}_{z\in\mathcal{E}} \|y - z\|_\Sigma = \begin{cases} \left\{ \frac{1}{\|y\|}\, y \right\} & \text{für } y_1 \geq 0, y \neq \mathbf{0}_2, \\ \mathcal{E} & \text{für } y = \mathbf{0}_2, \\ \left\{ \binom{0}{-1}, \binom{0}{1} \right\} & \text{für } y_1 < 0, y_2 = 0, \\ \left\{ \binom{0}{-1} \right\} & \text{für } y_1 < 0, y_2 < 0, \\ \left\{ \binom{0}{1} \right\} & \text{für } y_1 < 0, y_2 > 0 \end{cases}$$

Man erhält

$$\arg\min_{\vartheta\in\Theta} \|y-\eta(\vartheta)\|_\Sigma = \begin{cases} \left\{\arcsin\left(\frac{y_2}{\|y\|}\right)\right\} & \text{für } y_1 \geq 0,\ y \neq \mathbf{0}_2, \\ \Theta & \text{für } y = \mathbf{0}_2, \\ \{-\pi/2, \pi/2\} & \text{für } y_1 < 0,\ y_2 = 0, \\ \{-\pi/2\} & \text{für } y_1 < 0,\ y_2 < 0, \\ \{\pi/2\} & \text{für } y_1 < 0,\ y_2 > 0. \end{cases}$$

Hier ist Θ kompakt und $\mathcal{E}$ abgeschlossen und das Proximum an $y \in \mathbb{R}^2$ existiert in jedem Fall. Es ist allerdings auf einer Lebesgue-Nullmenge nicht eindeutig. Für das Existenzlevel von $S^{\eta,y}$ gilt $S_E^{\eta,y} = S_{E,\partial}^{\eta,y} = \infty$.

b) Sei $\Theta = \mathbb{R}$ und

$$\eta : \mathbb{R} \to \mathbb{R}^2,\ \eta(\vartheta) = \begin{pmatrix} \cos(\arctan(\vartheta)) \\ \sin(\arctan(\vartheta)) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1+\vartheta^2}} \\ \frac{\vartheta}{\sqrt{1+\vartheta^2}} \end{pmatrix}$$

sowie $\Sigma = I_2$. Zwar ist Θ abgeschlossen, aber

$$\mathcal{E} = \eta(\Theta) = \{z \in \mathbb{R}^2 : \|z\| = 1,\ z_1 > 0\}$$

ist nicht abgeschlossen. Ein Proximum in $\mathcal{E}$ existiert nicht für $y \in \mathbb{R}^2$ mit $y_1 \leq 0$, $y \neq \mathbf{0}_2$. Es gilt

$$\arg\min_{z\in\mathcal{E}} \|y-z\|_\Sigma = \begin{cases} \left\{\frac{1}{\|y\|}\, y\right\} & \text{für } y_1 > 0, \\ \mathcal{E} & \text{für } y = \mathbf{0}_2, \\ \emptyset & \text{für } y_1 \leq 0,\ y \neq \mathbf{0}_2. \end{cases}$$

Man erhält

$$\arg\min_{\vartheta\in\Theta} \|y-\eta(\vartheta)\|_\Sigma = \begin{cases} \left\{\frac{y_2}{y_1}\right\} & \text{für } y_1 > 0, \\ \Theta & \text{für } y = \mathbf{0}_2, \\ \emptyset & \text{für } y_1 \leq 0,\ y \neq \mathbf{0}_2; \end{cases}$$

es gibt also nur für $y = \mathbf{0}_2$ mehr als ein Proximum. Für das Existenzlevel von $S^{\eta,y}$ gilt

$$S_E^{\eta,y} = S_{E,\partial}^{\eta,y} = \min\{y_1^2 + (y_2-1)^2,\ y_1^2 + (y_2+1)^2\} = \begin{cases} y_1^2 + (y_2-1)^2 & \text{für } y_2 \geq 0, \\ y_1^2 + (y_2+1)^2 & \text{für } y_2 < 0. \end{cases}$$

Als Vorbereitung zur Frage der Eindeutigkeit der Kleinste-Quadrate-Schätzung untersuchen wir, wann $\mathcal{E}$ eine C^k-Mannigfaltigkeit ist. Für eine Erläuterung dieses Begriffs verweisen wir auf die Grundlagen aus der Differentialgeometrie in Anhang B.

Satz 2.18
Sei $p \leq N$, $\Theta \subseteq \mathbb{R}^p$ offen und $\eta : \Theta \to \mathbb{R}^N$ eine $C^k(\Theta)$-Abbildung mit $\operatorname{rg}(J_\eta(\vartheta)) = p$ für alle $\vartheta \in \Theta$. Dann gibt es zu jedem $\vartheta \in \Theta$ eine offene Umgebung $U_\vartheta \subseteq \Theta$, so dass $\eta(U_\vartheta)$ eine p-dimensionale C^k-Mannigfaltigkeit des $\mathbb{R}^N$ und $\eta|_{U_\vartheta} : U_\vartheta \to \eta(U_\vartheta)$ ein Homöomorphismus ist.

Beweis: siehe Forster (1984, § 14, Satz 3) □

Bemerkung 2.19
Unter den Voraussetzungen des Satzes 2.18 kann man nicht folgern, dass $\mathcal{E} = \eta(\Theta)$ eine C^k-Mannigfaltigkeit ist, sogar nicht einmal dann, wenn η injektiv ist, anders als in Pázman (1993, Proposition 4.3.2) behauptet, wie wir im nachfolgenden Beispiel 2.20 sehen werden.
Man kann aber folgern, dass $\mathcal{E}$ die Vereinigung endlich vieler oder abzählbar unendlich vieler C^k-Mannigfaltigkeiten ist: Das Mengensystem $(U_\vartheta)_{\vartheta\in\Theta}$ aus dem Satz bildet eine offene Überdeckung der offenen Teilmenge $\Theta \subseteq \mathbb{R}^p$. Da $\mathbb{R}^p$ sich als Vereinigung abzählbar vieler kompakter Quader schreiben lässt, genügen abzählbar viele Mengen U_ϑ, um Θ zu überdecken.

Beispiel 2.20

a) Wir betrachten die Modellfunktion

$$\eta_1 : \mathbb{R} \to \mathbb{R}^2,\ \eta_1(\vartheta) = (\vartheta^2 - 1, \vartheta^3 - \vartheta),$$

deren Bild in Abbildung 2.3 zu sehen ist und für die $\operatorname{rg}(J_{\eta_1}(\vartheta)) = \operatorname{rg}\left(\begin{pmatrix} 2\vartheta \\ 3\vartheta^2-1 \end{pmatrix}\right) = 1$ für alle $\vartheta \in \mathbb{R}$ gilt. Da $\eta_1(-1) = \eta_1(1) = \mathbf{0}_2$ gilt, ist η_1 nicht injektiv. Das Bild lässt sich als Vereinigung zweier eindimensionaler Mannigfaltigkeiten darstellen,

$$\eta_1(\mathbb{R}) = \eta_1\left(\left(-\infty, \tfrac{1}{2}\right)\right) \cup \eta_1\left(\left(-\tfrac{1}{2}, \infty\right)\right).$$

In keiner Umgebung um den Punkt $\mathbf{0}_2$ ist aber die Menge $\mathcal{E}$ der Graph einer reellen Funktion und nach Lemma B.7d) keine eindimensionale Mannigfaltigkeit.

b) Aber selbst Injektivität genügt nicht, damit das Bild $\mathcal{E} = \eta(\Theta)$ der Modellfunktion eine Mannigfaltigkeit ist. Wir betrachten hierzu die Funktion

$$\eta_2 : \mathbb{R} \to \mathbb{R}^2,\ \eta_2(\vartheta) = \eta_1(e^\vartheta - 1).$$

Diese Funktion ist injektiv und es gilt $\operatorname{rg}(J_{\eta_2}(\vartheta)) = 1$ für alle $\vartheta \in \mathbb{R}$. Trotzdem ist immer noch in keiner Umgebung um den Punkt $\mathbf{0}_2$ die Menge $\mathcal{E}$ der Graph einer reellen Funktion, wie in Lemma B.7d) gefordert, und damit ist $\mathcal{E} = \eta(\Theta)$ keine eindimensionale Mannigfaltigkeit, vgl. Abbildung 2.3.

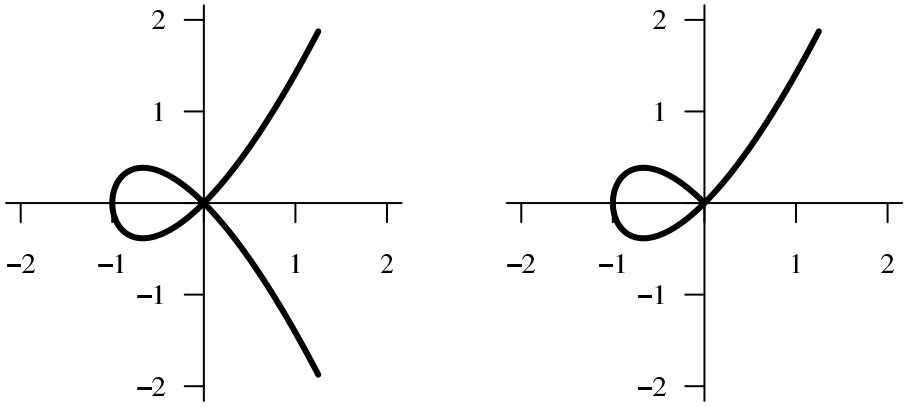


Abbildung 2.3.: Die Bilder von η_1 (links) und von η_2 (rechts)

Die Frage der Eindeutigkeit des Proximums (2.17) (und damit der Kleinste-Quadrate-Lösung) bei Injektivität der Modellfunktion) in der Situation von Satz 2.18 klärt der folgende Satz.

Satz 2.21
Sei I eine endliche oder abzählbar unendliche Indexmenge und $\mathcal{E} = \cup_{i\in I}\mathcal{E}_i$, wobei die Mengen $\mathcal{E}_i$ C^2-Mannigfaltigkeiten sind. Ferner sei Σ positiv definit. Dann ist

$$\{y \in \mathbb{R}^N : \# \arg\min_{z\in\mathcal{E}} \|y - z\|_\Sigma \geq 2\}$$

eine Lebesgue-Nullmenge im $\mathbb{R}^N$.

Beweis: siehe Pázman (1984) und Koutková (1992) □

Wir bemerken ferner, dass die C^2-Mannigfaltigkeiten $\mathcal{E}_i$ in Satz 2.21 nicht notwendig gleiche Dimension im R^N haben müssen und nicht notwendig disjunkt sein müssen.

Korollar 2.22
Sei $\Theta \subseteq \mathbb{R}^p$ offen und η eine $C^2(\Theta)$-Abbildung mit $\operatorname{rg}(J_\eta(\vartheta)) = p$ für alle $\vartheta \in \Theta$. Dann gibt es unter der Normalverteilungsannahme von Modell (1.1) *für P-fast alle $y \in \mathbb{R}^N$ höchstens ein Proximum an y in $\mathcal{E} = \eta(\Theta)$.*
Ist η zudem injektiv, so gibt es P-fast sicher höchstens eine Lösung des Kleinste-Quadrate-Problems (2.15).

Beweis: Wende Satz 2.18 und Satz 2.21 an. □

Ist Θ nicht offen oder ist nicht für alle $\vartheta \in \Theta$ die Annahme $\operatorname{rg}(J_\eta(\vartheta)) = p$ erfüllt, so kann man oft den Parameterbereich Θ geeignet in das Innere und den Rand von Θ, und gegebenenfalls weiter in Bereiche, auf denen der Rang $\operatorname{rg}(J_\eta(\vartheta))$ konstant ist,

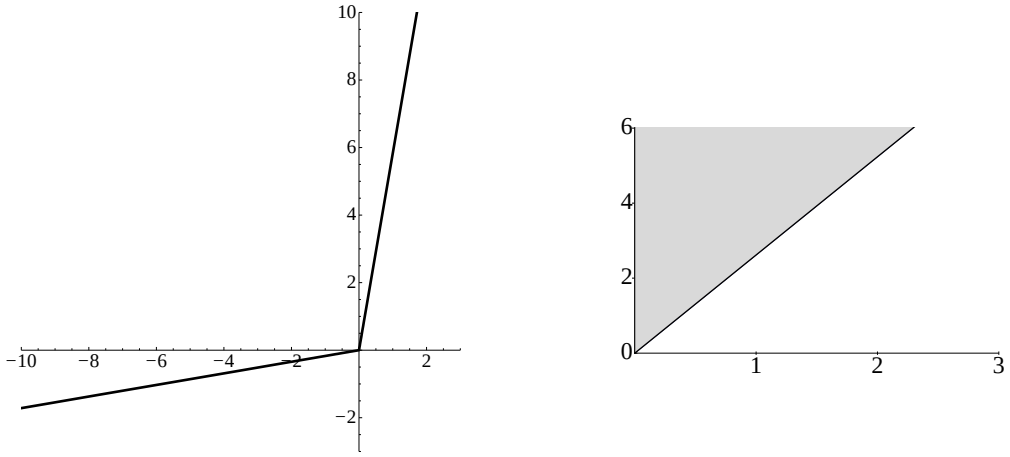


Abbildung 2.4.: Die Projektion von $\mathcal{E}$ auf die ersten beiden Komponenten aus Beispiel 2.23a) (links) und $\mathcal{E}$ aus Beispiel 2.23b) (rechts)

zerlegen und dann Satz 2.21 auf die sich ergebende Vereinigung von Mannigfaltigkeiten anwenden. Wir illustrieren dieses Vorgehen anhand der folgenden Beispiele.

Beispiel 2.23

a) Gegeben seien für $\vartheta \in \mathbb{R}$ die Matrix

$$A(\vartheta) = \begin{pmatrix} \vartheta & \vartheta \\ \vartheta & 3\vartheta \end{pmatrix}$$

und

$$\eta : \mathbb{R} \to \mathbb{R}^4,\ \eta(\vartheta) = (\lambda_{\min}(\vartheta), \lambda_{\max}(\vartheta), \lambda_{\min}(\vartheta), \lambda_{\max}(\vartheta))^{\mathrm{T}},$$

wobei $\lambda_{\min}(\vartheta) = \min\{(2-\sqrt{2})\vartheta, (2+\sqrt{2})\vartheta\}$ den minimalen Eigenwert von $A(\vartheta)$ und $\lambda_{\max}(\vartheta) = \max\{(2-\sqrt{2})\vartheta, (2+\sqrt{2})\vartheta\}$ den maximalen Eigenwert von $A(\vartheta)$ bezeichnen. Dann ist $\mathcal{E} = \eta(\mathbb{R} \setminus \{0\}) \cup \{\eta(0)\}$ die Vereinigung einer eindimensionalen C^2-Mannigfaltigkeit, bestehend aus zwei disjunkten Halbgeraden, und eines Punktes, also einer 0-dimensionalen Mannigfaltigkeit, vgl. Abbildung 2.4.

b) Gegeben sei für $\vartheta \in \mathbb{R}^2$ die Matrix

$$B(\vartheta) = \begin{pmatrix} \vartheta_2 & \vartheta_1 \\ \vartheta_1 & 2\vartheta_1 \end{pmatrix}.$$

Dann ist

$$\Theta = \{\vartheta \in \mathbb{R}^2 : \vartheta_1 > 0,\ \vartheta_2 > \vartheta_1/2\}$$

diejenige Menge, auf der $B(\vartheta)$ positiv definit ist. Wir betrachten die Modellfunktion

$$\eta : \Theta \to \mathbb{R}^2,\ \eta(\vartheta) = \begin{pmatrix} \vartheta_1 + \vartheta_2/2 - \frac{1}{2}\sqrt{8\vartheta_1^2 - 4\vartheta_1\vartheta_2 + \vartheta_2^2} \\ \vartheta_1 + \vartheta_2/2 + \frac{1}{2}\sqrt{8\vartheta_1^2 - 4\vartheta_1\vartheta_2 + \vartheta_2^2} \end{pmatrix},$$

wobei $\eta_1(\vartheta)$ der minimale und $\eta_2(\vartheta)$ der maximale Eigenwert von $B(\vartheta)$ ist. Dann ist $\mathcal{E} = \eta(\Theta) = \{(x,y)^{\mathrm{T}} \in \mathbb{R}^2 : x > 0,\ y \geq x(3+\sqrt{5})/2\}$, weil $\mathcal{E}$ offensichtlich ein Kegel ist (d.h. mit $z \in \mathcal{E}$ gilt auch $\mu z \in \mathcal{E}$ für $\mu > 0$), der wegzusammenhängend ist, und der Kegelbereich durch die extremalen Strahlen berandet wird, die sich aus $\vartheta_2 \to \vartheta_1/2$ bzw. $\vartheta_2 = 3\vartheta_1$ ergeben. Die Menge $\mathcal{E}$ kann man in ihr Inneres $\mathrm{int}(\mathcal{E})$ (eine offene Teilmenge des $\mathbb{R}^2$) und einen Halbstrahl zerlegen, vgl. Abbildung 2.4.

Wir haben uns bisher mit der Existenz und Eindeutigkeit der Lösung ϑ^* des Kleinste-Quadrate-Problems beschäftigt, in der das globale Minimum von

$$S^{\eta,y} : \Theta \to \mathbb{R},\ S^{\eta,y}(\vartheta) = \|y - \eta(\vartheta)\|_\Sigma^2$$

angenommen wird. Um eine solche Lösung numerisch zu berechnen, beachte man Folgendes:

Im Fall dass η stetig differenzierbar auf $\mathrm{int}(\Theta)$ ist und $\vartheta^* \in \mathrm{int}(\Theta)$ eine Lösung des Kleinste-Quadrate-Problems ist, d.h. $S^{\eta,y}$ das Minimum in einem inneren Punkt von Θ annimmt, muss bekanntermaßen

$$\mathrm{grad}\, S^{\eta,y}(\vartheta^*) = \mathbf{0}_p^{\mathrm{T}}$$

gelten. Wegen

$$\mathrm{grad}^{\,\mathrm{T}} S^{\eta,y}(\vartheta) = 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - y\right)$$

muss ϑ^* also die sogenannten **Normalgleichungen** erfüllen, d.h.

$$< J_\eta^i(\vartheta^*), \eta(\vartheta^*) - y >_\Sigma = 0 \tag{2.19}$$

für $i = 1, \ldots, p$ erfüllen, wobei $J_\eta^i(\vartheta^*) = \frac{\partial \eta}{\partial \vartheta_i}(\vartheta^*)$ den i-ten Spaltenvektor der Jacobi-Matrix $J_\eta(\vartheta^*) \in \mathbb{R}^{N\times p}$ bezeichne. Der Residuenvektor $y - \eta(\vartheta^*)$ steht also senkrecht (bzgl. des von Σ induzierten Skalarprodukts) auf den Vektoren $\frac{\partial \eta}{\partial \vartheta_i}(\vartheta^*)$, $i = 1, \ldots, p$, die den Tangentialraum in $\eta(\vartheta^*)$ aufspannen. In Abbildung 2.5 wird dies graphisch veranschaulicht, wobei zur besseren Illustration der Tangentialraum affin an den Vektor $\eta(\vartheta^*)$ angeheftet wurde.

Für die numerische Praxis ist es interessant, ob es bei nichtlinearen Kleinste-Quadrate-Problemen mehrere lokale Minima gibt. Dieser Frage haben sich unter anderem Mäkeläinen u. a. (1981) und Demidenko (1989, 2000) gewidmet. Zur Klärung sind zwei vorbereitende Definitionen angebracht.

Definition 2.24
Gegeben sei eine Modellfunktion $\eta : \mathbb{R}^p \to \mathbb{R}^N$. Man sagt, dass die Modellfunktion **infinite Tails** besitzt, falls

$$\|\eta(\vartheta_k)\| \overset{k\to\infty}{\longrightarrow} \infty \quad \text{für alle Folgen} \quad (\vartheta_k)_{k\in\mathbb{N}} \quad \text{mit} \quad \|\vartheta_k\| \overset{k\to\infty}{\longrightarrow} \infty.$$

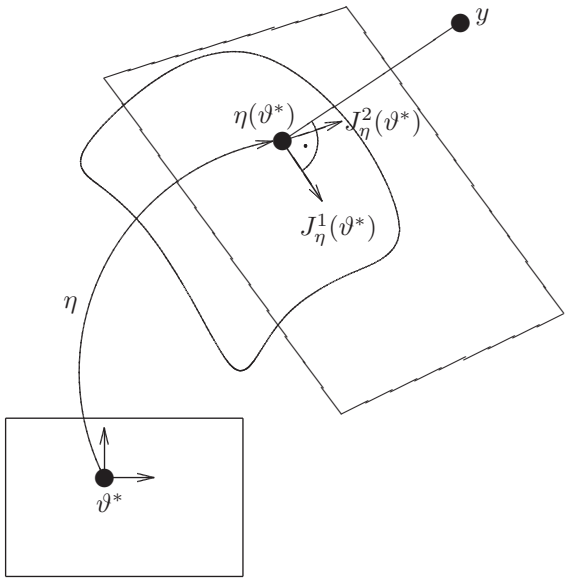


Abbildung 2.5.: Veranschaulichung der Normalgleichungen und des Tangentialraums

Bemerkung 2.25
Hat die stetige Modellfunktion $\eta : \mathbb{R}^p \to \mathbb{R}^N$ infinite Tails, dann ist für jedes $y \in \mathbb{R}^N$ $S_E^{\eta,y} = \infty$, so dass trivialerweise nach Korollar 2.15 für alle $y \in \mathbb{R}^N$ ein Proximum in $\mathcal{E}$ existiert. In der Sprache der Funktionalanalysis impliziert die Eigenschaft der infinite Tails, dass $S^{\eta,y}$ für alle $y \in \mathbb{R}^N$ koerzitiv ist.

Definition 2.26
Die zweimal stetig differenzierbare Modellfunktion $\eta : \mathbb{R}^p \to \mathbb{R}^N$ heißt **intrinsisch nichtlinear der 2. Ordnung**, wenn ein $\vartheta \in \mathbb{R}^p$ existiert, so dass die $N \times p(p+3)/2$-Matrix mit den Spalten

$$\frac{\partial \eta}{\partial \vartheta_i}(\vartheta) \text{ für } i = 1, \ldots, p \text{ und } \frac{\partial^2 \eta}{\partial \vartheta_k \partial \vartheta_j}(\vartheta) \text{ für } (j,k) \text{ mit } j,k = 1, \ldots, p \text{ und } j \geq k$$

maximalen Rang hat, also der Rang gleich $\min\{N, p(p+3)/2\}$ ist.

Ehe wir zur angekündigten Klärung schreiten, erläutern wir den Zusammenhang des Begriffs der intrinsischen Nichlinearität 2.Ordnung und des Begriffs der intrinsischen Linearität.

Definition 2.27
Sei $\Theta \subseteq \mathrm{cl}(\mathrm{int}(\Theta)) \subseteq \mathbb{R}^p$. Eine stetige Modellfunktion $\eta : \Theta \to \mathbb{R}^N$ heißt **intrinsisch linear**, falls $\eta(\mathrm{int}(\Theta))$ eine offene Teilmenge eines s-dimensionalen affinen Teilraums $\mathcal{L} \subseteq \mathbb{R}^N$ bezüglich der Relativtopologie von $\mathcal{L}$ ist mit $s \leq p$.

Die Bedingung $\Theta \subseteq \mathrm{cl}(\mathrm{int}(\Theta)) \subseteq \mathbb{R}^p$ dient dazu, dass Punkte auf dem Rand von Θ als Grenzwerte von Punkten aus dem Inneren darstellbar sind. Wenn eine Modellfunktion η

intrinsisch linear ist, dann kann man durch eine Reparametrisierung ein (affin)-lineares Modell erhalten, siehe Pázman (1993, S. 37).

Satz 2.28
Sei Θ offen und wegzusammenhängend, $\eta : \Theta \to \mathbb{R}^N$ zweimal stetig differenzierbar und $\operatorname{rg}(J_\eta(\vartheta)) = p$ für alle $\vartheta \in \Theta$. Dann ist die Modellfunktion η genau dann intrinsisch linear, falls für alle $i, j = 1, \dots, p$ die Gleichung

$$\frac{\partial^2 \eta}{\partial \vartheta_i \partial \vartheta_j}(\vartheta) = J_\eta(\vartheta) \left(J_\eta^{\mathrm{T}}(\vartheta) J_\eta(\vartheta) \right)^{-1} J_\eta^{\mathrm{T}}(\vartheta) \frac{\partial^2 \eta}{\partial \vartheta_i \partial \vartheta_j}(\vartheta) \tag{2.20}$$

gilt.

Beweis: siehe Pázman (1993, Proposition 2.2.1), dort fehlt jedoch die Voraussetzung des Wegzusammenhangs.
Im letzten Schritt des angegebenen Beweises wird gefolgert, dass der Raum linear ist, weil jede Geodätische in $\eta(\Theta)$ eine Gerade ist. Dieser Schluss ist ohne die Voraussetzung, dass Θ wegzusammenhängend ist, i.A. nicht korrekt, wie nachfolgendes Beispiel zeigt. □

Beispiel 2.29
Sei $\Theta = (0, 1) \cup (2, 3)$ und

$$\eta : \Theta \to \mathbb{R}^2,\ \eta(\vartheta) = \begin{cases} (\vartheta, \vartheta)^{\mathrm{T}} & \text{für } \vartheta \in (0, 1), \\ (\vartheta, \vartheta - 1)^{\mathrm{T}} & \text{für } \vartheta \in (2, 3). \end{cases}$$

Dann ist η offensichtlich nicht intrinsisch linear, erfüllt aber alle Voraussetzungen von Satz 2.28 bis auf den Wegzusammenhang.

Bemerkung 2.30
Sei nun η intrinsisch nichtlinear der 2. Ordnung und gelte $N > p$ und $\operatorname{rg}(J_\eta(\vartheta)) = p$. Wäre Gleichung (2.20) erfüllt, dann wäre $\frac{\partial^2 \eta}{\partial \vartheta_i \partial \vartheta_j}(\vartheta) \in \operatorname{Bild}(J_\eta(\vartheta))$ für alle $i, j = 1, \dots, p$ und daher der Rang der entsprechenden Matrix in Definition 2.26 lediglich p. Daher ist in diesem Fall η nicht intrinsisch linear.

Nach diesem kleinen Exkurs nun zur Antwort auf die Frage, ob es mehrere lokale Minima gibt.

Satz 2.31
Sei Σ positiv definit, $N \geq p(p+3)/2$ und sei $\eta : \mathbb{R}^p \to \mathbb{R}^N$ dreimal stetig differenzierbar, intrinsisch nichtlinear der 2. Ordnung und habe infinite Tails. Dann enthält die Menge

$$U = \{ y \in \mathbb{R}^N : S^{\eta, y} \text{ besitzt mindestens zwei lokale Minima} \}$$

eine nichtleere offene Teilmenge des $\mathbb{R}^N$ und hat daher positives Lebesgue-Maß.

Beweis: siehe Demidenko (2000) □

Für die numerischen Verfahren ist das etwas ernüchternd, da man unter den Voraussetzungen des Satzes nicht wissen kann, ob ein berechnetes lokales Minimum auch ein globales Minimum ist. Demidenko (2000) argumentiert, dass in „praktischen statistischen Problemen“ die Wahrscheinlichkeit $P(\boldsymbol{Y} \in U)$ klein sei, siehe aber hierzu die folgenden beiden Beispiele.

Beispiel 2.32
Gegeben sei die Modellfunktion

$$\eta : \mathbb{R} \to \mathbb{R}^2,\ \eta(\vartheta) = \begin{pmatrix} e^{\vartheta} \\ e^{-\vartheta} \end{pmatrix}.$$

Diese Modellfunktion hat offensichtlich unendliche Tails und ist intrinsisch nichtlinear der 2. Ordnung, weil die Matrix

$$\left(\frac{\partial \eta}{\partial \vartheta}(\vartheta),\ \frac{\partial^2 \eta}{\partial \vartheta^2}(\vartheta) \right) = \begin{pmatrix} e^{\vartheta} & e^{\vartheta} \\ -e^{-\vartheta} & e^{-\vartheta} \end{pmatrix}$$

für alle $\vartheta \in \mathbb{R}$ Rang 2 hat. Da $N = 2 = p(p+3)/2$ sind die Voraussetzungen des Satzes 2.31 erfüllt. Die Erwartungswertmenge $\mathcal{E} = \{(t, 1/t)^{\mathrm{T}} : t > 0\}$ ist eine Hyperbel, siehe die gestrichelte Kurve in Abbildung 2.6 (links). Wir werden im Folgenden für $\Sigma = I_2$ (oder allgemeiner $\Sigma = \sigma^2 I_2$ für $\sigma > 0$) die Menge U ermitteln. Details finden sich im Anhang A.1, wir erwähnen hier nur die wesentlichen Schritte. Für

$$S^{\eta,y} : \mathbb{R} \to \mathbb{R},\ S^{\eta,y}(\vartheta) = \|y - \eta(\vartheta)\|^2 = (y_1 - e^{\vartheta})^2 + (y_2 - e^{-\vartheta})^2$$

erhält man mit $t = e^{\vartheta}$ als Ableitung nach ϑ

$$\tfrac{\partial}{\partial \vartheta} S^{\eta,y} : \mathbb{R} \to \mathbb{R},\ \tfrac{\partial}{\partial \vartheta} S^{\eta,y}(\vartheta) = \frac{2}{t^2} p_y(t)$$

mit dem Polynom

$$p_y(t) = t^4 - y_1 t^3 + y_2 t - 1. \tag{2.21}$$

Wir untersuchen daher, wie viele Nullstellen $t > 0$ dieses Polynom 4. Grades in Abhängigkeit von y_1 und y_2 besitzt, wobei offensichtlich mindestens eine Nullstelle im Intervall $(-\infty, 0)$ liegt. Mit Hilfe von Bedingungen, wann ein Polynom 4. Grades vier reelle Nullstellen hat, und einer Kurvendiskussion ergibt sich die Menge der $y \in \mathbb{R}^2$, für die p_y drei reelle Nullstellen auf $(0, \infty)$ besitzt:

$$U = \{y \in \mathbb{R}^2 : y_1 > 0,\ y_2 > 0,\ 4(-y_1 y_2 + 4)^3 + 27(y_1^2 - y_2^2)^2 < 0\}.$$

Diese drei Nullstellen bedeuten dann zwei lokale Minima und ein lokales Maximum von $S^{\eta,y}$. Dieser Bereich U ist in Abbildung 2.6 (links) dargestellt. Er wird von oben von

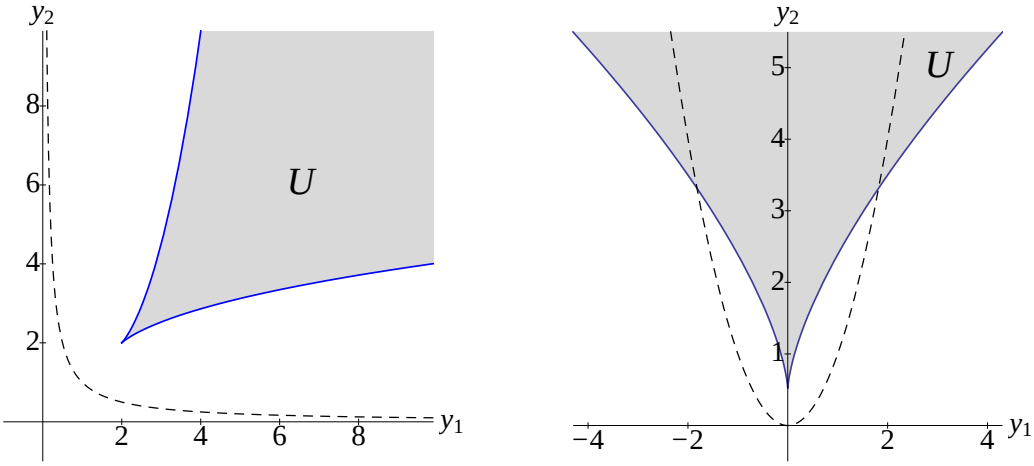


Abbildung 2.6.: Die Erwartungswertmenge $\mathcal{E}$ (gestrichelt) und der Bereich U, in dem für jeden Punkt $y \in U$ zwei lokale Minima von $S^{\eta,y}$ existieren, für das Beispiel 2.32 (links) und für das Beispiel aus Demidenko (2000) (rechts)

$$f_o : [2,\infty) \to \mathbb{R},\ f_o(y_1) =$$
$$\frac{y_1^3}{27} + \frac{1}{2}h(y_1) + \frac{1}{2}\sqrt{\frac{8y_1^6}{729} + \frac{8y_1^2}{27} - \frac{2}{9}g(y_1) - \frac{8\left(5y_1^4+64\right)}{9g(y_1)} + \frac{\frac{64y_1^9}{19683} + \frac{32y_1^5}{243} + \frac{512y_1}{9}}{4h(y_1)}}$$

und von unten von

$$f_u : [2,\infty) \to \mathbb{R},\ f_u(y_1) =$$
$$\frac{y_1^3}{27} + \frac{1}{2}h(y_1) - \frac{1}{2}\sqrt{\frac{8y_1^6}{729} + \frac{8y_1^2}{27} - \frac{2}{9}g(y_1) - \frac{8\left(5y_1^4+64\right)}{9g(y_1)} + \frac{\frac{64y_1^9}{19683} + \frac{32y_1^5}{243} + \frac{512y_1}{9}}{4h(y_1)}}$$

begrenzt, wobei zur Abkürzung die Teilformeln

$$g(y_1) =$$
$$\sqrt[3]{y_1^{10} + 40y_1^6 + 2560y_1^2 + \sqrt{y_1^{20} + 80y_1^{16} - 1280y_1^{12} - 102400y_1^8 + 2621440y_1^4 - 16777216}}$$

und

$$h(y_1) = \sqrt{\frac{4y_1^6}{729} + \frac{4y_1^2}{27} + \frac{2}{9}g(y_1) + \frac{8\left(5y_1^4+64\right)}{9g(y_1)}}$$

verwendet wurden.
Mit $\boldsymbol{Y} \sim \mathcal{N}(\eta(\vartheta), \sigma^2 I_2)$ ergibt sich für die Wahrscheinlichkeit, dass es mehrere lokale Minima gibt

$$P(\boldsymbol{Y} \in U) = \frac{1}{2\pi\sigma^2} \int_2^\infty \left(\int_{f_u(y_1)}^{f_o(y_1)} e^{-\frac{(y_2-1/t)^2}{2\sigma^2}}\, \mathrm{d}y_2 \right) e^{-\frac{(y_1-t)^2}{2\sigma^2}}\, \mathrm{d}y_1.$$

$P(\boldsymbol{Y} \in U)$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
$\sigma^2 = 0.25$	0.00009	0.00009	0.00002	0.00000	0.00000
$\sigma^2 = 0.5$	0.00136	0.00168	0.00100	0.00022	0.00003
$\sigma^2 = 1$	0.00711	0.00908	0.00899	0.00547	0.00242
$\sigma^2 = 2$	0.02013	0.02513	0.02972	0.02784	0.02127
$\sigma^2 = 4$	0.03995	0.04786	0.05860	0.06442	0.06321

Tabelle 2.1.: Die Wahrscheinlichkeiten $P(\boldsymbol{Y} \in U)$ im Beispiel 2.32

$P(\boldsymbol{Y} \in U)$	$\vartheta = 0$	$\vartheta = 1$	$\vartheta = 2$	$\vartheta = 3$
$\sigma^2 = 0.25$	0.0160	0.0557	0.7665	1.0000
$\sigma^2 = 0.5$	0.0332	0.1082	0.6999	1.0000
$\sigma^2 = 1$	0.0559	0.1496	0.6494	0.9998
$\sigma^2 = 2$	0.0822	0.1769	0.6084	0.9946
$\sigma^2 = 4$	0.1105	0.1960	0.5602	0.9652

Tabelle 2.2.: Wahrscheinlichkeiten $P(\boldsymbol{Y} \in U)$ im Beispiel aus Demidenko (2000)

In der Tabelle 2.1 ist diese Wahrscheinlichkeit für einige σ^2 und einige $t = e^{\vartheta}$ näherungsweise berechnet.

Beispiel 2.33
Demidenko (2000) untersucht die Modellfunktion

$$\eta : \mathbb{R} \to \mathbb{R}^2,\ \eta(\vartheta) = (\vartheta, \vartheta^2)^{\mathrm{T}},$$

d.h. die Erwartungswertmenge $\mathcal{E}$ ist eine Parabel. Er ermittelt hierfür (zu $\Sigma = \sigma I_2$ mit $\sigma > 0$)

$$U = \{y \in \mathbb{R}^2 : y_2 > 3|y_1|^{2/3}/\sqrt[3]{4} + 1/2\},$$

vgl. Abbildung 2.6 (rechts). Für die Wahrscheinlichkeit mehrerer lokaler Minima[8]

$$P(\boldsymbol{Y} \in U) = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \left(\int_{3|y_1|^{2/3}/\sqrt[3]{4}+\frac{1}{2}}^{\infty} e^{-\frac{\left(y_2-\vartheta^2\right)^2}{2\sigma^2}}\, \mathrm{d}y_2 \right) e^{-\frac{(y_1-\vartheta)^2}{2\sigma^2}}\, \mathrm{d}y_1$$

ergeben sich die Werte in Tabelle 2.2.

Die Wahrscheinlichkeit $P(\boldsymbol{Y} \in U)$ ist im ersten Beispiel in allen betrachteten Fällen kleiner als 0.07, während im zweiten Beispiel die Wahrscheinlichkeit in den meisten Fällen nicht zu vernachlässigen ist und sogar nahe bei 1 liegen kann. Anschaulich naheliegend

[8] Die in Demidenko (2000) angegebene Formel ist fehlerhaft.

ist, dass die größeren Wahrscheinlichkeiten im zweiten Beispiel an der deutlich größeren Krümmung der Erwartungswertmenge liegen. Die Aussage von Demidenko, dass in statistischen Problemen aus der Anwendungspraxis die Wahrscheinlichkeit $P(\boldsymbol{Y} \in U)$ klein ist, wird nicht durch das von ihm gegebene Beispiel gestützt.

Bemerkung 2.34
Wir werden nicht näher auf asymptotische Aussagen zum Kleinste-Quadrate-Schätzer eingehen, sondern verweisen auf die entsprechende Literatur, z.B. Pázman (1993) und Gallant (1987), und vermitteln hier nur einen kurzen Eindruck von der Art der Ergebnisse.
Sei Θ kompakt, seien $\eta_i : \Theta \to \mathbb{R}$ für $i \in \mathbb{N}$ und bezeichne $\eta^{(n)} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}} : \Theta \to \mathbb{R}^n$ für $n \in \mathbb{N}$. Gegeben sei eine Familie von Modellen für $\boldsymbol{Y}^{(n)}$, $n \in \mathbb{N}$, mit

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}^{(n)}) &= \eta^{(n)}(\vartheta) \quad \text{und} \\ \mathrm{D}(\boldsymbol{Y}^{(n)}) &= \sigma^2 I_n. \end{aligned}$$

Ferner existiere

$$M(\vartheta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \operatorname{grad}^{\mathrm{T}} \eta_i(\vartheta) \operatorname{grad} \eta_i(\vartheta) = \lim_{n\to\infty} \frac{1}{n} J^{\mathrm{T}}_{\eta^{(n)}}(\vartheta) J_{\eta^{(n)}}(\vartheta)$$

und sei invertierbar und sei $\hat{\vartheta}^{(n)} = \hat{\vartheta}^{(n)}(\boldsymbol{Y}^{(n)})$ ein Kleinste-Quadrate-Schätzer im n-ten Modell, d.h. für $y \in \mathbb{R}^n$ gilt

$$\hat{\vartheta}^{(n)}(y) \in \operatorname*{arg\,min}_{\vartheta\in\Theta} \left\| y - \eta^{(n)}(\vartheta) \right\|^2 .$$

Unter zusätzlichen Voraussetzungen an η_i, $i \in \mathbb{N}$, gilt dann für $\vartheta \in \operatorname{int}(\Theta)$ die Verteilungskonvergenz

$$\sqrt{n} \left(\hat{\vartheta}^{(n)} - \vartheta \right) \xrightarrow[\mathcal{L}]{n\to\infty} \mathcal{N}\left(0, \sigma^2 M^{-1}(\vartheta)\right).$$

Ferner gilt mit Wahrscheinlichkeit 1, dass

$$\lim_{n\to\infty} M(\hat{\vartheta}^{(n)}) = M(\vartheta).$$

Dies kann man wie folgt interpretieren: Bezeichnet ϑ_0 den wahren Parameterwert, so ist mit

$$\mathfrak{l}^{(n)} : \Theta \to \mathbb{R}^n,\ \mathfrak{l}^{(n)}(\vartheta) = J_{\eta^{(n)}}(\vartheta_0) \cdot (\vartheta - \vartheta_0) + \eta^{(n)}(\vartheta_0)$$

im Hinblick auf den Kleinste-Quadrate-Schätzer für großes n das lineare Modell

$$\tilde{\boldsymbol{Y}}^{(n)} \sim \mathcal{N}\left(\mathfrak{l}^{(n)}(\vartheta), \sigma^2 I_n\right)$$

eine gute Approximation des Modells für $\boldsymbol{Y}^{(n)}$. Da unter den obigen Voraussetzungen $\hat{\vartheta}^{(n)}(\boldsymbol{Y}^{(n)}) \to \vartheta_0$ mit Wahrscheinlichkeit 1 gilt, erscheint ferner zu einer Beobachtung $y \in \mathbb{R}^n$ eine Approximation durch das lineare Modell in $\vartheta_0 = \hat{\vartheta}^{(n)}(y)$ plausibel, z.B. insbesondere die Approximation

$$\mathrm{D}_\vartheta\left(\hat{\vartheta}^{(n)}\right) \approx \sigma^2 \left(J^{\mathrm{T}}_{\eta^{(n)}}(\hat{\vartheta}^{(n)}(y)) J_{\eta^{(n)}}(\hat{\vartheta}^{(n)}(y)) \right)^{-1}.$$

2.2. Tests und Konfidenzbereiche basierend auf dem Kleinste-Quadrate-Schätzer

2.2.1. Lineares Modell

Wir betrachten im affin-linearen Modell (2.1) zuerst den Fall $\boldsymbol{B}$, dass die Kovarianzmatrix von der Form $\Sigma = \sigma^2\Sigma_0$ mit unbekanntem $\sigma > 0$ und bekannter Matrix Σ_0 ist und formulieren vorbereitende Verteilungsresultate. Mit $\mathcal{P}_{X,\Sigma} = X(X^{\mathrm{T}}\Sigma^{-1}X)^+X^{\mathrm{T}}\Sigma^{-1} = X(X^{\mathrm{T}}\Sigma_0^{-1}X)^+X^{\mathrm{T}}\Sigma_0^{-1}$ bezeichnen wir wieder den orthogonalen Projektor auf Bild(X) bezüglich des von Σ und damit auch bezüglich des von Σ_0 induzierten Skalarprodukts.

Lemma 2.35
Sei $\mathrm{rg}(X) = r \le p$ *und für (iii)-(v) gelte ferner* $r < N$.

(i) Für den Residuenvektor $\hat{\mathsf{e}}(\boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\eta(\vartheta)}(\boldsymbol{Y})$ *gilt*

$$\mathrm{E}_{\vartheta,\sigma}(\hat{\mathsf{e}}(\boldsymbol{Y})) = 0 \quad \textit{und} \quad \mathrm{D}_{\vartheta,\sigma}(\hat{\mathsf{e}}(\boldsymbol{Y})) = \sigma^2(I_N - \mathcal{P}_{X,\Sigma})\Sigma_0.$$

(ii) Im Fall $S^{\eta,y}(\hat{\vartheta}(y)) \neq 0$ *(vgl. Fußnote 1, S. 5) ist die Maximum-Likelihood-Schätzung* $\widehat{\sigma^2_{ML}}(y)$ *für* σ^2 *zur Beobachtung* y *gegeben durch*

$$\widehat{\sigma^2_{ML}}(y) = \frac{1}{N}\,\|\hat{\mathsf{e}}(y)\|^2_{\Sigma_0} = \frac{1}{N}\,\|(I_N - \mathcal{P}_{X,\Sigma})(y - y_0)\|^2_{\Sigma_0} = \frac{1}{N}(y - y_0)^{\mathrm{T}}R(y - y_0)$$

mit $R = \Sigma_0^{-1}(I_N - \mathcal{P}_{X,\Sigma}) = \Sigma_0^{-1} - \Sigma_0^{-1}X(X^{\mathrm{T}}\Sigma_0^{-1}X)^+X^{\mathrm{T}}\Sigma_0^{-1}$.

(iii) Der Schätzer

$$\widehat{\sigma^2} = \widehat{\sigma^2}(\boldsymbol{Y}) = \frac{1}{N - r}\,\|\hat{\mathsf{e}}(y)\|^2_{\Sigma_0}$$

ist ein erwartungstreuer Schätzer für σ^2 *und optimal in der Klasse der erwartungstreuen Schätzer bezüglich jeder konvexen Verlustfunktion; insbesondere gilt*

$$\mathrm{D}_{\vartheta,\sigma}(\widehat{\sigma^2}(\boldsymbol{Y})) \le \mathrm{D}_{\vartheta,\sigma}(\delta(\boldsymbol{Y}))$$

für alle $\delta : (\mathbb{R}^N, \mathcal{B}^N) \to (\mathbb{R}, \mathcal{B})$ *mit* $\mathrm{E}_{\vartheta,\sigma}(\delta(\boldsymbol{Y})) = \sigma^2$ *für alle* $(\vartheta, \sigma) \in \mathbb{R}^p \times (0, \infty)$. *Ferner gilt*

$$\frac{N - r}{\sigma^2}\,\widehat{\sigma^2}(\boldsymbol{Y}) \sim \chi^2_{N-r}\,.$$

(iv) Der Schätzer

$$\hat{\sigma}_e(\boldsymbol{Y}) = \frac{\sqrt{N - r}}{\sqrt{2}}\,\frac{\Gamma((N - r)/2)}{\Gamma((N + 1 - r)/2)}\,\sqrt{\widehat{\sigma^2}(\boldsymbol{Y})}$$

ist erwartungstreu für σ *und optimal in der Klasse der erwartungstreuen Schätzer bezüglich jeder konvexen Verlustfunktion; insbesondere gilt*

$$\mathrm{D}_{\vartheta,\sigma}(\hat{\sigma}_e(\boldsymbol{Y})) \le \mathrm{D}_{\vartheta,\sigma}(\delta(\boldsymbol{Y}))$$

für alle $\delta : (\mathbb{R}^N, \mathcal{B}^N) \to (\mathbb{R}, \mathcal{B})$ *mit* $\mathrm{E}_{\vartheta,\sigma}(\delta(\boldsymbol{Y})) = \sigma$ *für alle* $(\vartheta, \sigma) \in \mathbb{R}^p \times (0, \infty)$.

(v) *Sei* $L \in \mathbb{R}^{q\times p}$ *mit* $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$. *Der Gauß-Markov-Schätzer* $\widehat{L\vartheta}(\boldsymbol{Y})$ *für* $L\vartheta$ *und* $\widehat{\sigma^2}(\boldsymbol{Y})$ *sind stochastisch unabhängig. Sei ferner* $\mathrm{rg}(L) = q$. *Für*

$$Q(\boldsymbol{Y}) = (\widehat{L\vartheta}(\boldsymbol{Y}) - L\vartheta)^{\mathrm{T}} V_L^{-1} (\widehat{L\vartheta}(\boldsymbol{Y}) - L\vartheta)$$

mit

$$V_L = L(X^{\mathrm{T}}\Sigma_0^{-1}X)^{+}L^{\mathrm{T}}$$

gilt dann

$$Q(\boldsymbol{Y})/\sigma^2 \sim \chi^2_q$$

und

$$F(\boldsymbol{Y}) = \frac{Q(\boldsymbol{Y})}{q\,\widehat{\sigma^2}(\boldsymbol{Y})} \sim F_{q,N-r}.$$

Für $c \in \mathrm{Bild}(X^{\mathrm{T}}) \setminus \{\mathbf{0}_p\}$ *gilt*

$$\frac{1}{\sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^{+}c \cdot \widehat{\sigma^2}(\boldsymbol{Y})}} (\widehat{c^{\mathrm{T}}\vartheta}(\boldsymbol{Y}) - c^{\mathrm{T}}\vartheta) \sim t_{N-r}.$$

Insbesondere folgt aus (v), dass im Fall $\mathrm{rg}(X) = p$ der Kleinste-Quadrate-Schätzer $\hat{\vartheta}(\boldsymbol{Y})$ und $\widehat{\sigma^2}(\boldsymbol{Y})$ stochastisch unabhängig sind und es gilt

$$F(\boldsymbol{Y}) = \frac{1}{p\,\widehat{\sigma^2}(\boldsymbol{Y})} (\hat{\vartheta} - \vartheta)^{\mathrm{T}} X^{\mathrm{T}}\Sigma_0^{-1}X(\hat{\vartheta} - \vartheta) \sim F_{p,N-p}.$$

Für eine Approximation der Vorfaktoren in (iv) für großes $N - r$ siehe Gurland u. Tripathi (1971). Oft verwendet man jedoch $\hat{\sigma}(\boldsymbol{Y}) = \sqrt{\widehat{\sigma^2}(\boldsymbol{Y})}$ zur Schätzung von σ.
Aussagen zur Optimalität von $\widehat{\sigma^2}(\boldsymbol{Y})$ und zur Unkorreliertheit von $\widehat{\sigma^2}(\boldsymbol{Y})$ und $\widehat{L\vartheta}(\boldsymbol{Y})$ unter Bedingungen an die ersten vier Momente von $\boldsymbol{Y}$ statt der Normalverteilungsannahme finden sich in Witting (1985, S. 456 ff.).

Nun kurz zum Fall $\boldsymbol{A}$, dass die Kovarianzmatrix Σ vollständig bekannt ist.

Bemerkung 2.36

(i) Für den Residuenvektor $\hat{\mathsf{e}}(\boldsymbol{Y}) = \boldsymbol{Y} - \widehat{\eta(\vartheta)}(\boldsymbol{Y})$ gilt

$$\mathrm{E}_{\vartheta,\sigma}(\hat{\mathsf{e}}(\boldsymbol{Y})) = 0 \quad \text{und} \quad \mathrm{D}_{\vartheta,\sigma}(\hat{\mathsf{e}}(\boldsymbol{Y})) = (I_N - \mathcal{P}_{X,\Sigma})\Sigma.$$

(ii) Sei $L \in \mathbb{R}^{q\times p}$ mit $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$ und $\mathrm{rg}(L) = q$. Für

$$Q(\boldsymbol{Y}) = (\widehat{L\vartheta}(\boldsymbol{Y}) - L\vartheta)^{\mathrm{T}} V_L^{-1} (\widehat{L\vartheta}(\boldsymbol{Y}) - L\vartheta)$$

mit

$$V_L = L(X^{\mathrm{T}}\Sigma^{-1}X)^{+}L^{\mathrm{T}}$$

gilt dann

$$Q(\boldsymbol{Y}) \sim \chi^2_q.$$

Für $c \in \mathrm{Bild}(X^{\mathrm{T}}) \setminus \{\mathbf{0}_p\}$ gilt

$$\frac{1}{\sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}\Sigma^{-1}X)^+c}}\,(\widehat{c^{\mathrm{T}}\vartheta}(\boldsymbol{Y}) - c^{\mathrm{T}}\vartheta) \sim \mathcal{N}(0,1).$$

Testen linearer Hypothesen

Wir verzichten bei den folgenden Testprozeduren auf Aussagen zur Optimalität und verweisen hierfür auf Witting (1985). Soweit nicht extra angegeben, verwenden wir in den folgenden beiden Korollaren die Bezeichnungen von Lemma 2.35 und betrachten wieder zuerst den Fall $\boldsymbol{B}$. Mit $F_{q,s;\alpha}$ bezeichnen wir im Folgenden das α-Quantil der $F_{q,s}$-Verteilung, mit $t_{s;\alpha}$ das α-Quantil der t_s-Verteilung, mit $\chi^2_{s;\alpha}$ das α-Quantil der χ^2_s-Verteilung und schließlich mit Φ die Verteilungsfunktion der Standardnormalverteilung.

Korollar 2.37
Sei $\alpha \in (0,1)$ und $\mathrm{rg}(X) = r < N$.

(i) Sei $c \in \mathbb{R}^q$, $L \in \mathbb{R}^{q\times p}$ mit $\mathrm{rg}(L) = q$ und $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$. Für das Testproblem

$$H_0 : L\vartheta = c \quad \text{gegen} \quad H_1 : L\vartheta \neq c$$

ist

$$\phi(\boldsymbol{Y}) = \mathbf{1}_{(F_{q,N-r;1-\alpha},\,\infty)}\,(T_1(\boldsymbol{Y}))$$

mit der Teststatistik

$$T_1(\boldsymbol{Y}) = \frac{Q(\boldsymbol{Y})}{q \cdot \widehat{\sigma^2}(\boldsymbol{Y})}$$

ein α-Niveau-Test, wobei

$$Q(\boldsymbol{Y}) = (\widehat{L\vartheta}(\boldsymbol{Y}) - c)^{\mathrm{T}} V_L^{-1} (\widehat{L\vartheta}(\boldsymbol{Y}) - c).$$

(ii) Sei $b \in \mathbb{R}$, $c \in \mathrm{Bild}(X^{\mathrm{T}}) \setminus \{\mathbf{0}_p\}$. Für das Testproblem

$$H_0 : c^{\mathrm{T}}\vartheta \leq b \quad \text{gegen} \quad H_1 : c^{\mathrm{T}}\vartheta > b$$

ist

$$\phi(\boldsymbol{Y}) = \mathbf{1}_{(t_{N-r;1-\alpha},\,\infty)}\,(T_2(\boldsymbol{Y}))$$

mit der Teststatistik

$$T_2(\boldsymbol{Y}) = \frac{\widehat{c^{\mathrm{T}}\vartheta}(\boldsymbol{Y}) - b}{\sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^+c \cdot \widehat{\sigma^2}(\boldsymbol{Y})}}$$

ein α-Niveau-Test.

(iii) Sei $\tau > 0$. Für das Testproblem

$$H_0 : \sigma^2 \leq \tau^2 \quad \textit{gegen} \quad H_1 : \sigma^2 > \tau^2$$

ist

$$\phi(\boldsymbol{Y}) = \mathbf{1}_{(\chi^2_{N-r;1-\alpha},\, \infty)}\,(T_3(\boldsymbol{Y}))$$

mit der Teststatistik

$$T_3(\boldsymbol{Y}) = (N-r)\,\frac{\widehat{\sigma^2}(\boldsymbol{Y})}{\tau^2}$$

ein α-Niveau-Test.

(iv) Sei $\tau > 0$. Für das Testproblem

$$H_0 : \sigma^2 \geq \tau^2 \quad \textit{gegen} \quad H_1 : \sigma^2 < \tau^2$$

ist

$$\phi(\boldsymbol{Y}) = \mathbf{1}_{(0,\, \chi^2_{N-r;\alpha})}\,(T_3(\boldsymbol{Y}))$$

ein α-Niveau-Test.

(v) Seien $\tau > 0$ und $\pi_1, \pi_2 \in (0,1)$ mit $\pi_2 - \pi_1 = 1 - \alpha$. Für das Testproblem

$$H_0 : \sigma^2 = \tau^2 \quad \textit{gegen} \quad H_1 : \sigma^2 \neq \tau^2$$

ist

$$\phi(\boldsymbol{Y}) = 1 - \mathbf{1}_{\left[\chi^2_{N-r;\pi_1},\, \chi^2_{N-r;\pi_2}\right]}\,(T_3(\boldsymbol{Y}))$$

ein α-Niveau-Test.

Den Test unter (i) kann man auch als Likelihood-Ratio-Test auffassen. Zur optimalen Wahl von π_1 und π_2 in (v) vergleiche Witting (1985, S. 392); man verwendet meist $\pi_1 = \alpha/2$ und $\pi_2 = 1 - \alpha/2$.

Beispiel 2.38

a) Will man im Fall $\operatorname{rg}(X) = p$ für $t \in \mathbb{R}^p$ das Testproblem

$$H_0 : \vartheta = t \quad \text{gegen} \quad H_1 : \vartheta \neq t$$

untersuchen, so kommt (i) zur Anwendung mit $L = I_p$, $c = t$ und

$$Q(\boldsymbol{Y}) = (\hat{\vartheta}(\boldsymbol{Y}) - t)^{\mathrm{T}} X^{\mathrm{T}} \Sigma_0^{-1} X (\hat{\vartheta}(\boldsymbol{Y}) - t).$$

b) Testet man auf einen Teilvektor von ϑ, betrachtet also z.B. für $\vartheta^{\mathrm{T}} = (\vartheta_1^{\mathrm{T}}, \vartheta_2^{\mathrm{T}})$ mit $\vartheta_1 \in \mathbb{R}^u$, $\vartheta_2 \in \mathbb{R}^v$, $0 < u < p$, $v = p - u$ und $t \in \mathbb{R}^u$ das Testproblem

$$H_0 : \vartheta_1 = t \quad \text{gegen} \quad H_1 : \vartheta_1 \neq t,$$

so kann man (i) mit

$$L = \left[\; I_u, \; \mathbf{0}_{u \times v} \; \right]$$

anwenden, wobei $L \in \mathbb{R}^{u \times p}$ die Blockmatrix gebildet aus I_u und $\mathbf{0}_{u \times v}$ bezeichnet. Unterteilt man entsprechend $X = [X_1 \;\; X_2]$ mit einer Matrix $X_1 \in \mathbb{R}^{N \times u}$ und einer Matrix $X_2 \in \mathbb{R}^{N \times v}$, so gilt

$$\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}}) \iff \mathrm{rg}((I_N - \mathcal{P}_{X_2})X_1) = u.$$

Hierbei bezeichnet $\mathcal{P}_{X_2} = \mathcal{P}_{X_2, I_N}$ den orthogonalen Projektor auf $\mathrm{Bild}(X_2)$ bezüglich des Standardskalarprodukts; Details zum Beweis finden sich im Anhang A.3. Ferner gilt

$$\mathrm{rg}(X) = p \Rightarrow \mathrm{rg}((I_N - \mathcal{P}_{X_2})X_1) = u \Rightarrow \mathrm{rg}(X_1) = u,$$

d.h. $\mathrm{rg}(X_1) = u$ ist notwendig und $\mathrm{rg}(X) = p$ ist hinreichend für $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$ und man erhält unter der Bildinklusion mit Hilfe von Searle (1982, S. 263)

$$V_L = (X_1^{\mathrm{T}} \Sigma_0^{-1} (I_N - \mathcal{P}_{X_2}) X_1)^{-1}$$

und hiermit

$$\begin{aligned} Q(\boldsymbol{Y}) &= (\widehat{L\vartheta}(\boldsymbol{Y}) - t)^{\mathrm{T}} V_L^{-1} (\widehat{L\vartheta}(\boldsymbol{Y}) - t) \\ &= (L\hat{\vartheta}(\boldsymbol{Y}) - t)^{\mathrm{T}} X_1^{\mathrm{T}} \Sigma_0^{-1} (I_N - \mathcal{P}_{X_2}) X_1 (L\hat{\vartheta}(\boldsymbol{Y}) - t). \end{aligned}$$

Im Fall ***A*** der vollständig bekannten Kovarianzmatrix Σ werden oft die folgenden Tests verwendet, die man mittels Bemerkung 2.36 erhält.

Bemerkung 2.39

(i) Sei $c \in \mathbb{R}^q$, $L \in \mathbb{R}^{q \times p}$ mit $\mathrm{rg}(L) = q$ und $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$. Für das Testproblem

$$H_0 : L\vartheta = c \quad \text{gegen} \quad H_1 : L\vartheta \neq c$$

ist

$$\phi(\boldsymbol{Y}) = \mathbf{1}_{(\chi^2_{q;1-\alpha}, \, \infty)} \left(Q(\boldsymbol{Y}) \right)$$

mit der Teststatistik

$$Q(\boldsymbol{Y}) = (\widehat{L\vartheta}(\boldsymbol{Y}) - c)^{\mathrm{T}} V_L^{-1} (\widehat{L\vartheta}(\boldsymbol{Y}) - c)$$

ein α-Niveau-Test.

(ii) Sei $b \in \mathbb{R}$, $c \in \mathrm{Bild}(X^{\mathrm{T}}) \setminus \{\mathbf{0}_p\}$. Für das Testproblem

$$H_0 : c^{\mathrm{T}}\vartheta \leq b \quad \text{gegen} \quad H_1 : c^{\mathrm{T}}\vartheta > b$$

ist

$$\phi(\boldsymbol{Y}) = \mathbf{1}_{(\Phi^{-1}(1-\alpha),\,\infty)}\left(T_2(\boldsymbol{Y})\right)$$

mit der Teststatistik

$$T_2(\boldsymbol{Y}) = \frac{\widehat{c^{\mathrm{T}}\vartheta}(\boldsymbol{Y}) - b}{\sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}\Sigma^{-1}X)^{+}c}}$$

ein α-Niveau-Test.

Konfidenzbereiche für lineare Aspekte

Wir verzichten bei den folgenden exemplarischen Konfidenzbereichen, die man aus obigen Tests durch den üblichen Zusammenhang zwischen Konfidenzbereichen und Tests erhält, wieder auf Aussagen zur Optimalität und verweisen auf Witting (1985). Ferner betrachten wir nur Fall ***B***.

Korollar 2.40
Sei $\alpha \in (0,1)$.

(i) Sei $a \in \mathbb{R}^q$ und $L \in \mathbb{R}^{q\times p}$, wobei $\mathrm{Bild}(L^{\mathrm{T}}) \subseteq \mathrm{Bild}(X^{\mathrm{T}})$ und $\mathrm{rg}(L) = q$ gelte. Mit

$$V_L = L(X^{\mathrm{T}}\Sigma_0^{-1}X)^{+}L^{\mathrm{T}}$$

ist ein $(1-\alpha)$-Konfidenzellipsoid für $L\vartheta + a$ gegeben durch

$$\left\{z \in \mathbb{R}^q : (z - \widehat{L\vartheta}(\boldsymbol{Y}) - a)^{\mathrm{T}}V_L^{-1}(z - \widehat{L\vartheta}(\boldsymbol{Y}) - a) \leq q \cdot \widehat{\sigma^2}(\boldsymbol{Y})F_{q,N-r;1-\alpha}\right\}.$$

(ii) Sei $c \in \mathrm{Bild}(X^{\mathrm{T}}) \setminus \{\mathbf{0}_p\}$. Dann ist

$$\left[\widehat{c^{\mathrm{T}}\vartheta}(\boldsymbol{Y}) - \sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^{+}c \cdot \widehat{\sigma^2}(\boldsymbol{Y})}\, t_{N-r;1-\alpha/2},\right.$$
$$\left.\widehat{c^{\mathrm{T}}\vartheta}(\boldsymbol{Y}) + \sqrt{c^{\mathrm{T}}(X^{\mathrm{T}}\Sigma_0^{-1}X)^{+}c \cdot \widehat{\sigma^2}(\boldsymbol{Y})}\, t_{N-r;1-\alpha/2}\right]$$

ein $(1-\alpha)$-Konfidenzintervall für $c^{\mathrm{T}}\vartheta$.

(iii) Im Fall $\mathrm{rg}(X) = p$ ist ein $(1-\alpha)$-Konfidenzellipsoid für ϑ gegeben durch

$$\{\vartheta \in \mathbb{R}^p : (\vartheta - \hat{\vartheta}(\boldsymbol{Y}))^{\mathrm{T}}X^{\mathrm{T}}\Sigma_0^{-1}X(\vartheta - \hat{\vartheta}(\boldsymbol{Y})) \leq p \cdot \widehat{\sigma^2}(\boldsymbol{Y})F_{p,N-p;1-\alpha}\}.$$

(iv) Seien $\pi_1, \pi_2 \in (0,1)$ mit $\pi_2 - \pi_1 = 1 - \alpha$. Dann ist

$$\left[(N-r)\, \frac{\widehat{\sigma^2}(\boldsymbol{Y})}{\chi^2_{N-r;\pi_2}},\ (N-r)\, \frac{\widehat{\sigma^2}(\boldsymbol{Y})}{\chi^2_{N-r;\pi_1}} \right]$$

ein $(1-\alpha)$-Konfidenzintervall für σ^2.
(Üblich sind die Werte $\pi_1 = \alpha/2$ und $\pi_2 = 1 - \alpha/2$, vgl. die Bemerkung im Anschluss an Korollar 2.37.)

Beispiel 2.41

a) Ein $(1-\alpha)$-Konfidenzintervall für σ ist gegeben durch

$$\left[\sqrt{(N-r)\, \frac{\widehat{\sigma^2}(\boldsymbol{Y})}{\chi^2_{N-r;1-\alpha/2}}}\, ,\ \sqrt{(N-r)\, \frac{\widehat{\sigma^2}(\boldsymbol{Y})}{\chi^2_{N-r;\alpha/2}}} \right].$$

b) Sei $e_i \in \mathrm{Bild}(X^{\mathrm{T}})$. Mit

$$\mathrm{se}(\widehat{\vartheta_i}) = \mathrm{se}(\widehat{\vartheta_i})(\boldsymbol{Y}) = \sqrt{e_i^{\mathrm{T}} (X^{\mathrm{T}} \Sigma_0^{-1} X)^{+} e_i \cdot \widehat{\sigma^2}(\boldsymbol{Y})}$$

ist ein $(1-\alpha)$-Konfidenzintervall für ϑ_i gegeben durch

$$\left[\widehat{\vartheta_i}(\boldsymbol{Y}) - \mathrm{se}(\widehat{\vartheta_i}) t_{N-r;1-\alpha/2},\ \widehat{\vartheta_i}(\boldsymbol{Y}) + \mathrm{se}(\widehat{\vartheta_i}) t_{N-r;1-\alpha/2} \right].$$

c) Für simultane Konfidenzbereiche z.B. nach Bonferroni oder nach Scheffé verweisen wir auf Seber (1977, Chapter 5) oder Christensen (2002, Chapter 5).

2.2.2. Nichtlineares Modell

Wir werden uns im Folgenden auf Konfidenzbereiche beschränken. Mit dem bekannten Zusammenhang zwischen Konfidenzbereichen und Tests erhält man die zugehörigen Tests.

Die einfachste Idee, um in nichtlinearen Modellen Konfidenzbereiche zu erhalten, besteht darin, die Modellfunktion zu linearisieren. Hierfür gibt es zwei Methoden: Zum einen kann man in einem vorher gewählten Punkt ϑ_0 oder in der Kleinste-Quadrate-Schätzung $\widehat{\vartheta}(y)$ des nichtlinearen Modells linearisieren. Wir werden dieses Vorgehen, das auch durch die Resultate aus der Asymptotik motiviert ist, vgl. Bemerkung 2.34, beispielhaft erläutern. Eine entsprechende Übertragung auf die anderen Konfidenzbereiche (und Tests) aus dem vorhergehenden Abschnitt 2.2.1 über das lineare Modell geschieht nach dem gleichen Muster.

Approximiert man in einem Punkt ϑ_0 mit $\mathrm{rg}(J_\eta(\vartheta_0)) = p$ durch das entsprechende lineare Modell mit der Modellfunktion

$$\mathfrak{l} : \Theta \to \mathbb{R}^N,\ \mathfrak{l}(\vartheta) = J_\eta(\vartheta_0) \cdot (\vartheta - \vartheta_0) + \eta(\vartheta_0),$$

dann erhält man im Fall $\boldsymbol{B}$ (σ unbekannt) für $N > p$ mit

$$\widehat{\sigma^2}(\boldsymbol{Y}) = \frac{1}{N-p} \|\boldsymbol{Y} - \eta(\hat{\vartheta}(\boldsymbol{Y}))\|^2_{\Sigma_0}$$

mit Korollar 2.40(iii)

$$\{\vartheta \in \Theta : (\vartheta - \hat{\vartheta}(\boldsymbol{Y}))^{\mathrm{T}} J_\eta^{\mathrm{T}}(\vartheta_0)\Sigma_0^{-1} J_\eta(\vartheta_0)(\vartheta - \hat{\vartheta}(\boldsymbol{Y})) \le p\,\widehat{\sigma^2}(\boldsymbol{Y}) F_{p,N-p;1-\alpha}\}$$

als approximatives $(1-\alpha)$-Konfidenzellipsoid für ϑ bzw. linearisiert man stattdessen zur Beobachtung y im Punkt $\vartheta_0 = \hat{\vartheta}(y)$, so erhält man für $N > p$ und für ein reguläres Modell, d.h. mit $\mathrm{rg}(J_\eta(\vartheta)) = p$ für alle $\vartheta \in \Theta$,

$$\{\vartheta \in \Theta : (\vartheta - \hat{\vartheta}(\boldsymbol{Y}))^{\mathrm{T}} J_\eta^{\mathrm{T}}(\hat{\vartheta}(\boldsymbol{Y}))\Sigma_0^{-1} J_\eta(\hat{\vartheta}(\boldsymbol{Y}))(\vartheta - \hat{\vartheta}(\boldsymbol{Y})) \le p\,\widehat{\sigma^2}(\boldsymbol{Y}) F_{p,N-p;1-\alpha}\}$$

als approximatives $(1-\alpha)$-Konfidenzellipsoid für ϑ.[9] Im Fall $\boldsymbol{A}$ (Σ bekannt) erhält man durch Übertragung von Bemerkung 2.39 entsprechend die approximativen $(1-\alpha)$-Konfidenzellipsoide

$$\{\vartheta \in \Theta : (\vartheta - \hat{\vartheta}(\boldsymbol{Y}))^{\mathrm{T}} J_\eta^{\mathrm{T}}(\vartheta_0)\Sigma^{-1} J_\eta(\vartheta_0)(\vartheta - \hat{\vartheta}(\boldsymbol{Y})) \le \chi^2_{p;1-\alpha}\}$$

bzw.

$$\{\vartheta \in \Theta : (\vartheta - \hat{\vartheta}(\boldsymbol{Y}))^{\mathrm{T}} J_\eta^{\mathrm{T}}(\hat{\vartheta}(\boldsymbol{Y}))\Sigma^{-1} J_\eta(\hat{\vartheta}(\boldsymbol{Y}))(\vartheta - \hat{\vartheta}(\boldsymbol{Y})) \le \chi^2_{p;1-\alpha}\}$$

für ϑ.

Die Güte der Approximation hängt davon ab, wie gut das linearisierte Modell das nichtlineare Modell beschreibt. Ausgehend von Beale (1960) wurden daher Nichtlinearitätsmaße eingeführt, die auf einer quadratischen Approximation, d.h. auf den zweiten Ableitungen von η basieren. Diesbezüglich verweisen wir auf Bates u. Watts (1988, Chapter 7), Seber u. Wild (1989, Chapter 4) und Pázman (1993, Section 5.5 und Section 6.2).

Eine zweite Methode liefert sogar exakte Konfidenzbereiche für reguläre Modellfunktionen, d.h. $\mathrm{rg}(J_\eta(\vartheta)) = p$ für alle $\vartheta \in \Theta$. Wie im linearen Fall (vgl. auch Lemma 2.35) zeigt man, dass die Zufallsvariablen

$$P_{J_\eta(\vartheta),\Sigma}(\boldsymbol{Y} - \eta(\vartheta)) \sim \mathcal{N}\left(0, P_{J_\eta(\vartheta),\Sigma}\right)$$

und

$$(I_N - P_{J_\eta(\vartheta),\Sigma})(\boldsymbol{Y} - \eta(\vartheta)) \sim \mathcal{N}\left(0, (I_N - P_{J_\eta(\vartheta),\Sigma})\right)$$

[9] Ist $\Sigma_0^{-1} = U^{\mathrm{T}}U$ und ist für $X = J_\eta(\vartheta_0)$ bzw. $X = J_\eta(\hat{\vartheta}(y))$ eine QR-Zerlegung $QR = UX$ von UX mit $R = \left(\begin{smallmatrix} R_1 \\ 0 \end{smallmatrix}\right)$ gegeben, so gilt $X^{\mathrm{T}}\Sigma_0^{-1}X = R_1^{\mathrm{T}}R_1$ und man erhält

$$\{\vartheta \in \Theta : (\vartheta - \hat{\vartheta}(y))^{\mathrm{T}} X^{\mathrm{T}}\Sigma_0^{-1} X(\vartheta - \hat{\vartheta}(y)) \le p\,\widehat{\sigma^2}(y) F_{p,N-p;1-\alpha}\}$$
$$= \left\{\hat{\vartheta}(y) + \sqrt{p\,\widehat{\sigma^2}(y) F_{p,N-p;1-\alpha}}\, R_1^{-1} t : t \in \mathbb{R}^p, \|t\| \le 1\right\} \cap \Theta.$$

stochastisch unabhängig sind und daher auch die Zufallsvariablen

$$\left\|P_{J_\eta(\vartheta),\Sigma}(\boldsymbol{Y}-\eta(\vartheta))\right\|_\Sigma^2 \sim \chi_p^2$$

und

$$\left\|(I_N - P_{J_\eta(\vartheta),\Sigma})(\boldsymbol{Y}-\eta(\vartheta))\right\|_\Sigma^2 \sim \chi_{N-p}^2$$

stochastisch unabhängig sind. Dies führt im Fall $\boldsymbol{A}$ zum exakten $(1-\alpha)$-Konfidenzbereich

$$\left\{\vartheta\in\Theta : \left\|P_{J_\eta(\vartheta),\Sigma}(\boldsymbol{Y}-\eta(\vartheta))\right\|_\Sigma^2 \le \chi_{p;1-\alpha}^2\right\}$$

für ϑ und im Fall $\boldsymbol{B}$ für $N>p$ zum exakten $(1-\alpha)$-Konfidenzbereich

$$\left\{\vartheta\in\Theta : \frac{\left\|P_{J_\eta(\vartheta),\Sigma_0}(\boldsymbol{Y}-\eta(\vartheta))\right\|_{\Sigma_0}^2}{\left\|(I_N - P_{J_\eta(\vartheta),\Sigma_0})(\boldsymbol{Y}-\eta(\vartheta))\right\|_{\Sigma_0}^2} \le F_{p,N-p;1-\alpha}\right\} \tag{2.22}$$

für ϑ.

Beispiel 2.42
Wie in Beispiel 2.32 sei die Modellfunktion

$$\eta:\mathbb{R}\to\mathbb{R}^2,\ \eta(\vartheta)=\begin{pmatrix}e^{\vartheta}\\ e^{-\vartheta}\end{pmatrix}.$$

und $\Sigma = I_2$ gegeben. Dann ist

$$P_{J_\eta(\vartheta),I_2} = \frac{1}{e^{2\vartheta}+e^{-2\vartheta}}\begin{pmatrix}e^{2\vartheta} & -1\\ -1 & e^{-2\vartheta}\end{pmatrix}$$

und

$$g_y(\vartheta) = \left\|P_{J_\eta(\vartheta),I_2}(y-\eta(\vartheta))\right\|^2 = \frac{\left(e^{4\vartheta}-y_1e^{3\vartheta}+y_2e^{\vartheta}-1\right)^2}{e^{2\vartheta}+e^{6\vartheta}}. \tag{2.23}$$

Die Bestimmung von Konfidenzbereichen erfordert numerische Verfahren zur Bestimmung der Subniveaumengen der nichtlinearen Funktion g_y, also insbesondere die Bestimmung aller Lösungen der nichtlinearen Gleichung $g_y(\vartheta)=c$ für den kritischen Wert c. Das Verhalten der Funktion $g_y:\mathbb{R}\to\mathbb{R}$ ist jedoch im Allgemeinen nicht sofort zu überblicken.
Sei $y=(7,2)^{\mathrm{T}}$. Dann ergibt sich der Kleinste-Quadrate-Schätzer als der Logarithmus der einzigen reellen Nullstelle des Polynoms

$$p_y(t) = t^4 - y_1t^3 + y_2t - 1 = t^4 - 7t^3 + 2t - 1,$$

vgl. (2.21), also $\widehat{\vartheta}(y)\approx\ln(6.916971)\approx 1.94042$ und damit $\eta(\widehat{\vartheta}(y))\approx(6.91697, 0.1436)^{\mathrm{T}}$. Mit $\chi^2_{1;0.95}=3.84145882$ ergibt sich aber etwas überraschend zu $y=(7,2)^{\mathrm{T}}$ der 95%-Konfidenzbereich

$$\{\vartheta\in\mathbb{R} : \left\|P_{J_\eta(\vartheta),I_2}(y-\eta(\vartheta))\right\|^2 \le \chi^2_{1;0.95}\} \approx [-1.21815,\ -0.464639]\cup[1.60251, 2.19014],$$

vgl. Abbildung 2.7, obwohl das Intervall mit den negativen ϑ-Werten unplausibel erscheint.

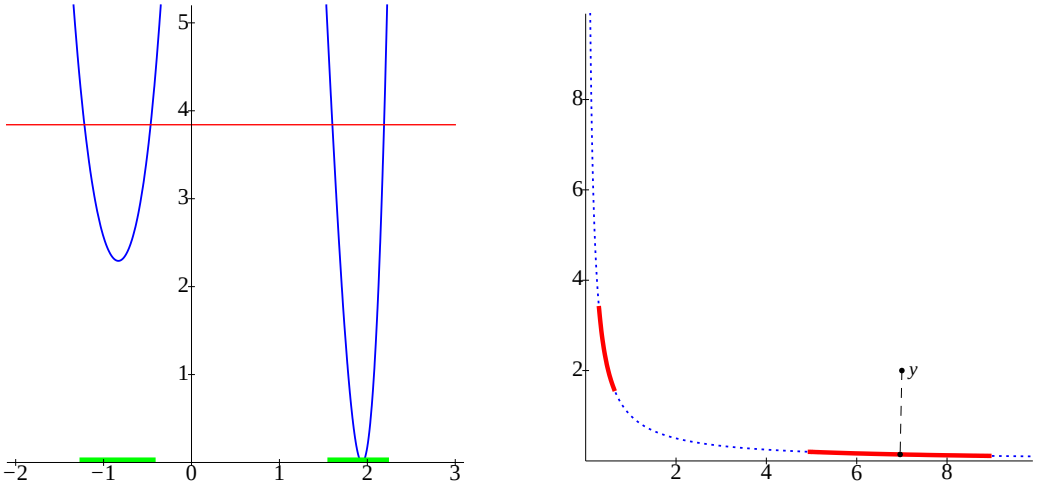


Abbildung 2.7.: Die Funktion g_y aus (2.23), der kritische Wert $\chi^2_{1;0.95}$ und das Konfidenzintervall zu $y = (7,2)^{\mathrm{T}}$ aus Beispiel 2.42 (links) und der entsprechende Bereich auf der Erwartungswertmenge $\mathcal{E}$ (rechts)

Zum einen sind diese exakten Konfidenzbereiche also numerisch schwierig zu bestimmen, zum anderen liefern sie mitunter unplausible unzusammenhängende Konfidenzbereiche. Das erklärt, warum die Konfidenzbereiche aus der Linearisierung in der Praxis so populär sind. Das Phänomen solcher unplausibler Konfidenzbereiche wird z.T. vermieden durch approximative $(1-\alpha)$-Konfidenzbereiche der Form

$$\left\{\vartheta \in \Theta : \left\|P_{J_\eta(\vartheta),\Sigma}(\boldsymbol{Y} - \eta(\vartheta))\right\|^2_\Sigma \leq \chi^2_{p;1-\alpha} \text{ und } \|\eta(\widehat{\vartheta}(\boldsymbol{Y})) - \eta(\vartheta)\|^2_\Sigma \leq r_0\right\}$$

für ϑ im Fall $\boldsymbol{A}$ und im Fall $\boldsymbol{B}$ durch approximative $(1-\alpha)$-Konfidenzbereiche

$$\left\{\vartheta \in \Theta : \frac{\left\|P_{J_\eta(\vartheta),\Sigma_0}(\boldsymbol{Y} - \eta(\vartheta))\right\|^2_{\Sigma_0}}{\left\|(I_N - P_{J_\eta(\vartheta),\Sigma_0})(\boldsymbol{Y} - \eta(\vartheta))\right\|^2_{\Sigma_0}} \leq F_{p,N-p;1-\alpha} \text{ und } \|\eta(\widehat{\vartheta}(\boldsymbol{Y})) - \eta(\vartheta)\|^2_{\Sigma_0} \leq r_0\right\}$$

für ϑ, die Pázman (1993, Chapter 8) untersucht.

Als weitere Möglichkeit machen wir abschließend noch auf Konfidenzbereiche basierend auf dem Likelihood-Ratio-Test und mögliche Verfeinerungen aufmerksam, siehe z.B. Pázman (1993, Section 8.3).

3. Kleinste-Quadrate-Schätzung: numerische Methoden

3.1. Lineares Modell

Wir beschreiben in diesem Abschnitt vier der gebräuchlichsten numerischen Verfahren zur Lösung der Normalgleichungen (2.5), also von

$$X^{\mathrm{T}}\Sigma^{-1}X(\vartheta^* - \vartheta_0) = X^{\mathrm{T}}\Sigma^{-1}(y - y_0),$$

wobei wir bei Mehrdeutigkeit ($\mathrm{rg}(X) < p$) an der Minimum-Norm-Lösung interessiert sind, d.h. wir wollen (2.6) bzw. (2.9) bestimmen.

Wir erinnern zuerst an den Begriff der Singulärwertzerlegung und der Konditionszahl einer Matrix.

3.1.1. Singulärwertzerlegung, Konditionszahl und Sensitivitätsanalysen

Bemerkung 3.1
Sei $C \in \mathbb{R}^{m\times n}$ mit $r = \mathrm{rg}(C) \leq \min\{m, n\}$. Mit $\lambda_1 \geq \ldots \geq \lambda_r$ bezeichnen wir die von 0 verschiedenen Eigenwerte von $C^{\mathrm{T}}C$. Sei $v_1, \ldots, v_n$ ein Orthonormalsystem des $\mathbb{R}^n$ von Eigenvektoren von $C^{\mathrm{T}}C$, wobei v_i Eigenvektor zum Eigenwert λ_i für $i = 1, \ldots, r$ und $v_{r+1}, \ldots, v_n$ Eigenvektoren zum Eigenwert 0 seien. Ferner ergänzen wir $u_i = \frac{1}{\sqrt{\lambda_i}} Cv_i$, $i = 1, \ldots, r$, zu einem Orthonormalsystem $u_1, \ldots, u_m$ von $\mathbb{R}^m$.[1]
Setzt man $V = [v_1, \ldots, v_n] \in \mathbb{R}^{n\times n}$, $U = [u_1, \ldots, u_m] \in \mathbb{R}^{m\times m}$, $d_i = \sqrt{\lambda_i}$ für $i = 1, \ldots, r$ und $d_i = 0$ für $i = r+1, \ldots, \min\{m, n\}$ und

$$D = (s_{ij})_{\substack{i=1,\ldots,m\\ j=1,\ldots,n}} \quad \text{mit} \quad s_{ij} = \begin{cases} d_i & \text{für } i = j, \\ 0 & \text{für } i \neq j \end{cases}$$

[1] Man beachte, dass $u_1, \ldots, u_r$ dann normiert und orthogonal sind, weil

$$u_i^{\mathrm{T}}u_i = \frac{1}{\sqrt{\lambda_i}}\frac{1}{\sqrt{\lambda_i}} v_i^{\mathrm{T}}C^{\mathrm{T}}Cv_i = \frac{1}{\lambda_i} v_i^{\mathrm{T}}\lambda_i v_i = 1 \text{ und } u_i^{\mathrm{T}}u_j = \frac{\sqrt{\lambda_j}}{\sqrt{\lambda_i}} v_i^{\mathrm{T}}v_j = 0 \text{ für } 1 \leq i, j \leq r \text{ mit } i \neq j$$

und für $i = 1, \ldots, r$, ferner u_i Eigenvektor von CC^{T} zum Eigenwert λ_i ist, da

$$CC^{\mathrm{T}}u_i = \frac{1}{\sqrt{\lambda_i}} CC^{\mathrm{T}}Cv_i = \frac{1}{\sqrt{\lambda_i}} C\lambda_i v_i = \lambda_i u_i$$

gilt. Dann lassen sich $u_1, \ldots, u_r$ mit Eigenvektoren $u_{r+1}, \ldots, u_m$ von CC^{T} zum Eigenwert 0 zu einem Orthogonalsystem ergänzen.

für $i = 1, \ldots, m$, $j = 1, \ldots, n$, dann ist

$$C = UDV^{\mathrm{T}}$$

eine **Singulärwertzerlegung**[2] von C. Die positiven Zahlen $d_1 \geq \ldots \geq d_r$ heißen auch die **singulären Werte** von C.
Insbesondere ist dann[3]

$$C^+ = VD^+U^{\mathrm{T}}$$

mit

$$D^+ = (t_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}} \quad \text{mit} \quad t_{ij} = \begin{cases} 1/d_i & \text{für } i = j \leq r, \\ 0 & \text{sonst,} \end{cases}$$

d.h. $1/d_r \geq \ldots \geq 1/d_1$ sind die singulären Werte von C^+.

Wir definieren im Folgenden zwei verschiedene Konditionszahlen, die man je nachdem verwendet, ob man an der allgemeinen Lösung oder an der Minimum-Norm-Lösung eines linearen Gleichungssystems interessiert ist. Mit $\|\cdot\|$ bezeichnen wir die Spektralnorm einer Matrix, also die induzierte Matrixnorm zur Euklidischen Vektornorm, d.h. es gilt

$$\|C\| = \max_{x \in \mathbb{R}^n,\, x \neq \mathbf{0}_n} \frac{\|Cx\|}{\|x\|}.$$

Für eine Matrix $C \in \mathbb{R}^{m \times n}$ ist daher $\|C\|^2$ der maximale Eigenwert von $C^{\mathrm{T}}C$.

Definition 3.2
Sei $C \in \mathbb{R}^{n \times n}$ mit $C \neq \mathbf{0}_{n \times n}$. Ist C regulär, so setzt man

$$\mathrm{cond}(C) = \|C\| \left\|C^{-1}\right\|,$$

ist C nichtregulär, so setzt man $\mathrm{cond}(C) = \infty$.
Man nennt $\mathrm{cond}(C)$ die **Konditionszahl** von C.

Bemerkung 3.3
Sind $d_1 \geq \ldots \geq d_n > 0$ die singulären Werte der regulären Matrix $C \in \mathbb{R}^{n \times n}$, so ist

$$\mathrm{cond}(C) = \frac{d_1}{d_n} = \frac{\max\limits_{x \in \mathbb{R}^n,\, x \neq \mathbf{0}_n} \|Cx\| \,/\, \|x\|}{\min\limits_{x \in \mathbb{R}^n,\, x \neq \mathbf{0}_n} \|Cx\| \,/\, \|x\|}\,.$$

[2] Eine Singulärwertzerlegung $C = UDV^{\mathrm{T}}$ liegt genau dann vor, wenn U und V orthogonale Matrizen sind und D nur auf der Hauptdiagonalen Einträge ungleich 0 hat, die positiv und absteigend sortiert sind.

[3] Die Darstellung $C^+ = VD^+U^{\mathrm{T}}$ ist jedoch keine Singulärwertzerlegung von C^+, hierzu wäre noch die Reihenfolge der ersten r Spalten von U und V und der ersten r Diagonaleinträge von D^+ umzudrehen.

Die Motivation für den Begriff der Konditionszahl liegt in folgendem Satz begründet, der eine Abschätzung für den relativen Fehler der (allgemeinen) Lösung eines linearen Gleichungssystems mit Koeffizientenmatrix C bei gestörter rechter Seite liefert.

Satz 3.4
Seien $C \in \mathbb{R}^{n\times n}$ und $b, \tilde{b} \in \mathbb{R}^n$. Sei $x^* \in \mathbb{R}^n$ eine Lösung des linearen Gleichungssystems $Cx = b$ und $\tilde{x}^* \in \mathbb{R}^n$ eine Lösung des linearen Gleichungssystems $Cx = \tilde{b}$. Dann gilt (für $b \neq \mathbf{0}_n$)

$$\frac{\|\tilde{x}^* - x^*\|}{\|x^*\|} \leq \operatorname{cond}(C)\, \frac{\|\tilde{b} - b\|}{\|b\|}.$$

Ist C regulär und ist $b \neq \tilde{b}$, wird Gleichheit genau dann angenommen, falls $\tilde{b} - b$ im Eigenraum zum größten Eigenwert von $(C^{-1})^{\mathrm{T}} C^{-1}$ liegt und b im Eigenraum zum größten Eigenwert von CC^{T} liegt. Ist C singulär, dann kann die linke Seite der Ungleichung beliebig groß werden.

Beweis: Ist C singulär, dann ist die Abschätzung wegen $\operatorname{cond}(C) = \infty$ trivial. Da C singulär ist, gibt es ein $z \in \mathbb{R}^n$, so dass $Cz = \mathbf{0}_n$. Mit $\tilde{x}^*$ ist auch $\tilde{x}^* + \alpha z$ für $\alpha \in \mathbb{R}$ eine Lösung des Gleichungssystems $Cx = \tilde{b}$, so dass der zweite Teil der Behauptung folgt.
Wir betrachten nun den Fall, dass C regulär ist. Dann ist

$$\tilde{x}^* - x^* = C^{-1}(\tilde{b} - b)$$

und daher

$$\|\tilde{x}^* - x^*\| \leq \left\|C^{-1}\right\| \|\tilde{b} - b\|. \tag{3.1}$$

Wegen

$$\|b\| \leq \|C\|\, \|x^*\| \tag{3.2}$$

folgt

$$\frac{\|\tilde{x}^* - x^*\|}{\|x^*\|} \leq \left\|C^{-1}\right\| \|C\|\, \frac{\|\tilde{b} - b\|}{\|b\|}. \tag{3.3}$$

Gleichheit wird in (3.3) im Fall von $\tilde{b} \neq b$ und $b \neq \mathbf{0}_n$ genau dann angenommen, wenn in beiden Ungleichungen (3.1) und (3.2) Gleichheit angenommen wird. In der Ungleichung (3.1) wird genau dann Gleichheit angenommen, wenn $\tilde{b} - b$ ein Eigenvektor zum größten Eigenwert von $(C^{-1})^{\mathrm{T}} C^{-1}$ ist. In der Ungleichung (3.2) wird genau dann Gleichheit angenommen, wenn x^* Eigenvektor zum größten Eigenwert von $C^{\mathrm{T}}C$ ist, d.h. wenn

$$C^{\mathrm{T}} C x^* = \|C\|^2\, x^*.$$

Wegen $Cx^* = b$ und $\|C\| = \left\|C^{\mathrm{T}}\right\|$ ist dies durch Multiplikation mit C äquivalent zu

$$CC^{\mathrm{T}} b = \left\|C^{\mathrm{T}}\right\|^2 b,$$

d.h. dass b Eigenvektor zum größten Eigenwert von CC^{T} ist. □

Eine weitere Interpretation für die Konditionszahl liefert der Satz von Gastinel und Kahan, siehe Kahan (1966). Hierzu definieren wir den **relativen Abstand** einer Matrix $C \in \mathbb{R}^{n\times n}$ mit $C \neq \mathbf{0}_{n\times n}$ **zur Singularität** durch

$$\operatorname{dist}(C) = \inf\left\{\frac{\|\tilde{C} - C\|}{\|C\|} : \tilde{C} \in \mathbb{R}^{n\times n} \text{ singulär}\right\}. \tag{3.4}$$

Satz 3.5 (Gastinel und Kahan)
Sei $C \in \mathbb{R}^{n\times n}$ regulär. Dann gilt

$$\operatorname{dist}(C) = 1/\operatorname{cond}(C).$$

Beweis: siehe Higham (2002, Theorem 6.5) oder als Spezialfall von Satz 3.12 □

Die Konditionszahl ist also der Kehrwert des Abstandes zur Singularität. Ist die Konditionszahl hoch, so ist der relative Abstand zu einer singulären Matrix gering.
Störungen in der Koeffizientenmatrix werden durch die folgende Abschätzung erfasst.

Satz 3.6
Seien $C \in \mathbb{R}^{n\times n}$ regulär, $\tilde{C} \in \mathbb{R}^{n\times n}$ und $b \in \mathbb{R}^n$ mit $b \neq \mathbf{0}_n$ und es gelte

$$\operatorname{cond}(C)\,\frac{\|\tilde{C} - C\|}{\|C\|} < 1. \tag{3.5}$$

Dann ist auch $\tilde{C}$ regulär und für $x^* = C^{-1}b$ und $\tilde{x}^* = \tilde{C}^{-1}b$ gilt

$$\frac{\|\tilde{x}^* - x^*\|}{\|x^*\|} \leq \frac{\operatorname{cond}(C)}{1 - \operatorname{cond}(C)\frac{\|\tilde{C}-C\|}{\|C\|}}\,\frac{\|\tilde{C} - C\|}{\|C\|}.$$

Beweis: Mittels Satz 3.5 folgt aus (3.5) sofort, dass $\tilde{C}$ regulär ist. Für einen Beweis der weiteren Aussagen siehe z.B. Hämmerlin u. Hoffmann (1991, Abschnitt 5.2). □

Bemerkung 3.7

a) Da für $z \in \mathbb{R}$, $|z| < 1$,

$$\frac{1}{1-z} = \sum_{i=0}^{\infty} z^i$$

ist, erhält man mit $z = \operatorname{cond}(C)\frac{\|\tilde{C}-C\|}{\|C\|}$, dass für festes C und für $\|\tilde{C} - C\|$ hinreichend klein

$$\frac{\|\tilde{x}^* - x^*\|}{\|x^*\|} \leq \operatorname{cond}(C)\frac{\|\tilde{C} - C\|}{\|C\|} + O(\|\tilde{C} - C\|^2)$$

gilt.

b) Kombiniert man Satz 3.4 mit Satz 3.6, so ergibt sich unter den Voraussetzungen von Satz 3.6, dass für $\tilde{x}^* = \tilde{C}^{-1}\tilde{b}$ die Abschätzung

$$\frac{\|\tilde{x}^* - x^*\|}{\|x^*\|} \leq \frac{\mathrm{cond}(C)}{1 - \mathrm{cond}(C)\frac{\|\tilde{C}-C\|}{\|C\|}} \left(\frac{\|\tilde{C} - C\|}{\|C\|} + \frac{\|\tilde{b} - b\|}{\|b\|} \right)$$

gilt.

Wir kommen jetzt zur zweiten Definition einer Konditionszahl, die im Fall $b \in \mathrm{Bild}(C)$ auf das Problem der Minimum-Norm-Lösung eines linearen Gleichungssystems $Cx = b$ bzw. im Fall von $b \notin \mathrm{Bild}(C)$ auf das Minimierungsproblem $\min_{x\in\mathbb{R}^n} \|Cx - b\|$ zugeschnitten ist.

Definition 3.8
Sei $C \in \mathbb{R}^{m\times n}$ mit $C \neq \mathbf{0}_{m\times n}$, $r = \mathrm{rg}(C)$ und singulären Werten $d_1 \geq \ldots \geq d_r > 0$. Dann heißt

$$\mathrm{cond}_+(C) = \frac{d_1}{d_r}$$

die Konditionszahl von C (bezüglich des Minimum-Norm-Problems) oder auch die **Spektralkonditionszahl**.

Bemerkung 3.9
Es gilt also $\mathrm{cond}_+(C) = \|C\|\,\|C^+\|$ und im Falle einer regulären Matrix $C \in \mathbb{R}^{n\times n}$ gilt daher $\mathrm{cond}_+(C) = \mathrm{cond}(C)$. Trivialerweise gilt $\mathrm{cond}_+(C) \geq 1$.

Wir erinnern daran, dass wir mit $\mathcal{P}_C = C(C^{\mathrm{T}}C)^+C^{\mathrm{T}} = CC^+$ den orthogonalen Projektor auf $\mathrm{Bild}(C)$ bezüglich des Euklidischen Skalarprodukts bezeichnen.

Satz 3.10 (Verallgemeinerung von Satz 3.4)
Sei $C \in \mathbb{R}^{m\times n}$ mit $C \neq \mathbf{0}_{m\times n}$ und $b, \tilde{b} \in \mathbb{R}^n$ mit $\mathcal{P}_C b \neq \mathbf{0}_m$. Für $x^* = C^+b$ und $\tilde{x}^* = C^+\tilde{b}$ gilt dann

$$\frac{\|\tilde{x}^* - x^*\|}{\|x^*\|} \leq \mathrm{cond}_+(C)\,\frac{\|\tilde{b} - b\|}{\|\mathcal{P}_C b\|}\,.$$

Hierbei wird Gleichheit angenommen, falls $\tilde{b} - b$ im Eigenraum zum größten Eigenwert von $C^{+\mathrm{T}}C^+$ und b im Eigenraum zum größten Eigenwert von CC^{T} liegt.

Beweis: Seien $r = \mathrm{rg}(C)$, $d_1 \geq \ldots \geq d_r > 0$ die singulären Werte von C und $C = UDV^{\mathrm{T}}$ eine Singulärwertzerlegung mit den entsprechenden Bezeichnungen von Bemerkung 3.1. Es gilt

$$\tilde{x}^* - x^* = C^+(\tilde{b} - b)$$

und daher

$$\|\tilde{x}^* - x^*\| \leq \|C^+\|\,\|\tilde{b} - b\|\,. \tag{3.6}$$

Ferner[4] gilt mit $c = U^{\mathrm{T}}b$

$$x^* = C^+b = VD^+U^{\mathrm{T}}b = \sum_{i=1}^{r} (c_i/d_i)\ v_i$$

und wegen

$$\mathcal{P}_C b = CC^+b = UDD^+U^{\mathrm{T}}b = U \begin{pmatrix} I_r & \mathbf{0}_{r\times m-r} \\ \mathbf{0}_{m-r\times r} & \mathbf{0}_{m-r\times m-r} \end{pmatrix} c = \sum_{i=1}^{r} c_i u_i$$

weiter

$$\|x^*\|^2 = \sum_{i=1}^{r} (c_i/d_i)^2 \geq \frac{1}{d_1^2} \sum_{i=1}^{r} c_i^2 = \frac{1}{d_1^2} \left\| \sum_{i=1}^{r} c_i u_i \right\|^2 = \frac{1}{\|C\|^2} \|\mathcal{P}_C b\|^2 . \tag{3.7}$$

Es folgt mit (3.6)

$$\frac{\|\tilde{x}^* - x^*\|}{\|x^*\|} \leq \left\|C^+\right\| \|C\| \frac{\|\tilde{b} - b\|}{\|\mathcal{P}_C b\|}$$

und wegen $\mathrm{cond}_+(C) = \|C^+\|\,\|C\|$ die behauptete Ungleichung.
Wir kommen zur Gleichheitsdiskussion: In (3.6) gilt Gleichheit genau dann, wenn $\tilde{b} - b$ Eigenvektor zum größten Eigenwert von $(C^+)^{\mathrm{T}} C^+$ ist.
Sei s die Dimension des Eigenraums zum größten Eigenwert von CC^{T}, d.h. für die singulären Werte von C gilt $d_1 = \ldots = d_s > d_{s+1} \geq \ldots \geq d_r > 0$. Dann gilt Gleichheit in (3.7) genau dann, wenn $c_i = (U^{\mathrm{T}}b)_i = 0$ für $i = s+1, \ldots, r$. Dies ist äquivalent dazu, dass b im vom Eigenraum zum größten Eigenwert von CC^{T} und vom Nullraum $\mathrm{Null}(CC^{\mathrm{T}}) = \mathrm{Null}(C^{\mathrm{T}})$ aufgespannten Raum liegt. □

Bemerkung 3.11
Man beachte, dass $x^* = C^+b$ die Minimum-Norm-Lösung des Minimierungsproblems $\min_{x\in\mathbb{R}^n} \|Cx - b\|$ ist. Falls zudem $b \in \mathrm{Bild}(C)$ gilt, ist $x^* = C^+b$ die Minimum-Norm-Lösung des Gleichungssystems $Cx = b$. Ist C regulär, so gilt natürlich, dass $x^* = C^{-1}b$.
Ferner ist der größte Eigenwert von $C^{+\mathrm{T}}C^+ = U(D^+)^{\mathrm{T}}D^+U^{\mathrm{T}}$ gleich dem Kehrwert des kleinsten Eigenwertes $\lambda_r \neq 0$ von $CC^{\mathrm{T}} = UDD^{\mathrm{T}}U^{\mathrm{T}}$ und die zugehörigen Eigenräume sind identisch.
Die Konditionszahl $\mathrm{cond}_+(C)$ liefert also eine Abschätzung zum relativen Fehler der Minimum-Norm-Lösung des in der rechten Seite gestörten Gleichungssystems, die für gewisse Werte der rechten Seite sogar scharf ist.
Die Abschätzungen in Satz 3.10 und in Satz 3.4 wie auch die Abschätzungen in Satz 3.6 und Satz 3.17 sind zwar nicht das Gleiche wie eine Abschätzung für den relativen Fehler bei der numerischen Lösung durch einen Algorithmus, bei dem nicht nur Inputgrößen gestört sind, sondern in jedem Berechnungsschritt durch Rundung auf die vom Programm verwendeten Gleitkommazahlen gestört wird. Im letzteren Fall hat man dann im

[4] Alternativ kann man analog zum Beweis von Satz 3.4 die Formel (3.7) wieder aus $\|C\|\,\|x^*\| \geq \|Cx^*\| = \left\|CC^+b\right\| = \|\mathcal{P}_C b\|$ folgern.

Allgemeinen den zugehörigen Term $\tilde{C}$ der Störung nicht unter Kontrolle. Trotzdem werden aber beide Konditionszahlen oft zur praktischen Einschätzung für die numerischen Schwierigkeiten des Problems verwendet.

Eine weitere Interpretation für die Spektralkonditionszahl liefert eine Verallgemeinerung des Satzes 3.5 von Gastinel und Kahan. Hierzu definieren wir den **relativen Abstand** einer Matrix $C \in \mathbb{R}^{m\times n}$ mit $C \neq \mathbf{0}_{m\times n}$ **zu Rangverlust** durch

$$\operatorname{dist}_+(C) = \inf\left\{ \frac{\|\tilde{C} - C\|}{\|C\|} : \tilde{C} \in \mathbb{R}^{m\times n},\ \operatorname{rg}(\tilde{C}) < \operatorname{rg}(C) \right\}. \tag{3.8}$$

Satz 3.12
Sei $C \in \mathbb{R}^{m\times n}$ mit $C \neq \mathbf{0}_{m\times n}$. Dann gilt

$$\operatorname{dist}_+(C) = 1/\operatorname{cond}_+(C).$$

Bevor wir diesen Satz beweisen, führen wir folgende Bezeichnung ein:
Sei $A = (a_{ij})_{\substack{i=1,\dots,m\\ j=1,\dots,n}} \in \mathbb{R}^{m\times n}$. Für $k \in \{1,\dots,n\}$ sei $A_{(k)} = (\check{a}_{ij})_{\substack{i=1,\dots,m\\ j=1,\dots,n}} \in \mathbb{R}^{m\times n}$ die Matrix mit den Einträgen

$$\check{a}_{ij} = \begin{cases} a_{ij}, & \text{falls } j \leq k, \\ 0 & \text{falls } j > k. \end{cases}$$

Beweis von Satz 3.12: Sei $\tilde{C} \in \mathbb{R}^{m\times n}$ mit $\operatorname{rg}(\tilde{C}) < \operatorname{rg}(C)$. Dann gibt es $x \in \operatorname{Bild}(C^{\mathrm{T}}) = \operatorname{Bild}(C^+)$ mit $x \neq \mathbf{0}_n$, so dass $\tilde{C}x = 0$. Sei $z \in \mathbb{R}^m$ mit $x = C^+z$. Dann folgt aufgrund der Eigenschaften der Moore-Penrose-Inverse, dass $C^+Cx = C^+CC^+z = C^+z = x$ und damit

$$\|x\| = \|C^+(\tilde{C} - C)x\| \leq \|C^+\|\,\|\tilde{C} - C\|\,\|x\|\,.$$

Somit ergibt sich

$$\frac{\|\tilde{C} - C\|}{\|C\|} \geq \frac{1}{\|C\|\,\|C^+\|} = 1/\operatorname{cond}_+(C).$$

Es genügt also zu zeigen, dass diese untere Schranke auch angenommen wird.
Seien $d_1 \geq \ldots \geq d_r > 0$ die singulären Werte von C und $C = UDV^{\mathrm{T}}$ eine Singulärwertzerlegung. Setze $\check{C} = UD_{(r-1)}V^{\mathrm{T}}$. Dann gilt $\operatorname{rg}(\check{C}) = r - 1$ und

$$\left\|C - \check{C}\right\| = \left\|D - D_{(r-1)}\right\| = d_r = 1/\left\|C^+\right\|,$$

weil die Spektralnorm gegenüber orthogonalen Transformationen invariant ist. Somit gilt

$$\frac{\|\check{C} - C\|}{\|C\|} = \frac{1}{\|C\|\,\|C^+\|} = 1/\operatorname{cond}_+(C).$$

□

Wir zeigen kurz einen interessanten Zusammenhang des Satzes 3.12 mit dem Approximationstheorem von E. Schmidt.[5] Hierzu bezeichne

$$\|A\|_{\text{Frob}} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2 \right)^{1/2}$$

die Frobenius-Norm von A.

Satz 3.13 (Approximationstheorem von E. Schmidt)
Sei $C \in \mathbb{R}^{m\times n}$ mit $\operatorname{rg}(C) = r$ und den singulären Werten $d_1 \geq \ldots \geq d_r > 0$ und sei $C = UDV^{\mathrm{T}}$ eine Singulärwertzerlegung.
Dann gilt für $1 \leq k < r$:

$$\inf\left\{ \|C - \tilde{C}\|_{\text{Frob}} : \tilde{C} \in \mathbb{R}^{m\times n}, \operatorname{rg}(\tilde{C}) \leq k \right\} = \left\| C - UD_{(k)}V^{\mathrm{T}} \right\|_{\text{Frob}} = \left(\sum_{i=k+1}^{r} d_i^2 \right)^{1/2}.$$

Ferner ist $\tilde{C} = UD_{(k)}V^{\mathrm{T}}$ die eindeutige Minimalstelle genau dann, wenn $d_{k+1} \neq d_k$.

Beweis: siehe Ben-Israel u. Greville (2003, Chapter 6, Theorem 3) □

Bemerkung 3.14
Insbesondere gilt nach dem Beweis von Satz 3.12 und nach Satz 3.13 für $C \in \mathbb{R}^{m\times n}$ der Zusammenhang

$$\begin{aligned}\inf\left\{ \|\tilde{C} - C\| : \tilde{C} \in \mathbb{R}^{m\times n},\ \operatorname{rg}(\tilde{C}) < \operatorname{rg}(C) \right\} \\ = \inf\left\{ \|\tilde{C} - C\|_{\text{Frob}} : \tilde{C} \in \mathbb{R}^{m\times n},\ \operatorname{rg}(\tilde{C}) < \operatorname{rg}(C) \right\}.\end{aligned}$$

Der Beweis zeigt ferner, dass in (3.8) und damit auch in (3.4) das Infimum angenommen wird und wie eine zugehörige Matrix aussieht.[6]

Abschließend betrachten wir auch hier wieder Störungen in der Koeffizientenmatrix. Will man ein Analogon zu Satz 3.6 zeigen, dann liegt ein Problem allerdings darin, dass die Operation der Bildung der Moore-Penrose-Inversen unstetig bei Rangveränderungen ist. So gilt z.B.

$$C_k = \begin{pmatrix} 1 & 0 \\ 0 & 1/k \end{pmatrix} \to C = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{für } k \to \infty,$$

[5] Das Approximationstheorem von E. Schmidt wird auch Satz von Eckart-Young genannt, vgl. Eckart u. Young (1936), wurde aber erstmals von Schmidt (1907) bewiesen.

[6] Dass für jede orthogonal invariante Norm $\|\cdot\|$ und $k < \operatorname{rg}(C)$

$$\inf\left\{ \|\tilde{C} - C\| : \tilde{C} \in \mathbb{R}^{m\times n},\ \operatorname{rg}(\tilde{C}) \leq k \right\} = \|D - D_{(k)}\|$$

gilt, wurde erstmals von Mirsky (1960) gezeigt.

aber

$$C_k^+ = \begin{pmatrix} 1 & 0 \\ 0 & k \end{pmatrix} \not\to C^+ = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{für } k \to \infty.$$

Allgemeiner zeigt folgender Satz, dass für Matrizen C und $\tilde{C}$, die sehr kleinen Abstand, aber unterschiedlichen Rang haben, die jeweiligen Moore-Penrose-Inversen C^+ und $\tilde{C}^+$ weit voneinander entfernt sind.

Satz 3.15
Seien $\tilde{C}, C \in \mathbb{R}^{m\times n}$.

a) Falls $\operatorname{rg}(\tilde{C}) \neq \operatorname{rg}(C)$ *ist, dann gilt*

$$\left\|\tilde{C}^+ - C^+\right\| \geq 1/\left\|\tilde{C} - C\right\|.$$

b) Falls $\operatorname{rg}(\tilde{C}) = \operatorname{rg}(C)$ *ist, dann gilt*

$$\left\|\tilde{C}^+ - C^+\right\| \leq \mu \left\|\tilde{C}^+\right\| \left\|C^+\right\| \left\|\tilde{C} - C\right\|,$$

wobei

$$\mu = \begin{cases} (1+\sqrt{5})/2, & \textit{falls } \operatorname{rg}(C) < \min\{m,n\}, \\ \sqrt{2}, & \textit{falls } \operatorname{rg}(C) = \min\{m,n\} \textit{ und } m \neq n, \\ 1, & \textit{falls } \operatorname{rg}(C) = m = n. \end{cases}$$

Beweis: siehe Wedin (1973, Theorem 4.1 und Theorem 7.2) □

Bemerkung 3.16

a) Im Falle von $\lim_{k\to\infty} C_k = C$ gilt daher $\lim_{k\to\infty} C_k^+ = C^+$ genau dann, wenn $\lim_{k\to\infty} \operatorname{rg}(C_k) = \operatorname{rg}(C)$.

b) Um eine vernünftige Abschätzung für die Lösung des gestörten Problems erreichen zu können, muss man aufgrund von Satz 3.15 folglich voraussetzen, dass die Störung den Rang von C nicht ändert. Dies ist ein Hinweis, dass es für numerische Verfahren wichtig ist, dass der Rang bei Matrixumformungen erhalten bleibt.

Satz 3.17
Seien $\tilde{C}, C \in \mathbb{R}^{m\times n}$ *mit* $\operatorname{rg}(\tilde{C}) = \operatorname{rg}(C)$ *und* $C \neq \mathbf{0}_{m\times n}$ *und sei*

$$\gamma = \operatorname{cond}_+(C)\,\frac{\|\tilde{C} - C\|}{\|C\|} < 1.$$

Dann ist

$$\|\tilde{C}^+\| \leq \frac{1}{1-\gamma}\left\|C^+\right\|.$$

Seien ferner $b, \tilde{b} \in \mathbb{R}^m$. *Dann gilt für* $\tilde{x}^* = \tilde{C}^+\tilde{b}$ *und* $x^* = C^+b$ *mit*

$$M = \frac{\mathrm{cond}_+(C)}{1-\gamma}\left(\frac{\|\tilde{C}-C\|}{\|C\|}\,\|x^*\| + \frac{\|\tilde{b}-b\|}{\|C\|} + \frac{\|\tilde{C}-C\|}{\|C\|}\,\mathrm{cond}_+(C)\,\frac{\|(I_m-\mathcal{P}_C)b\|}{\|C\|}\right)$$

die Abschätzung

$$\|\tilde{x}^* - x^*\| \le M + \|\tilde{C}-C\|\,\left\|(CC^{\mathrm{T}})^+b\right\|. \tag{3.9}$$

Falls $\mathrm{rg}(C) = n$, *dann kann man die Abschätzung* (3.9) *verschärfen zu*

$$\|\tilde{x}^* - x^*\| \le M.$$

Beweis: siehe Wedin (1973, Theorem 5.1) □

Bemerkung 3.18

a) Aus der Bedingung $\gamma < 1$ folgt mit Satz 3.12, dass $\mathrm{rg}(\tilde{C}) \ge \mathrm{rg}(C)$. Falls $\mathrm{rg}(C) = \min\{m,n\}$, folgt damit bereits $\mathrm{rg}(C) = \mathrm{rg}(\tilde{C})$.

b) Falls $\mathrm{rg}(C) = n$ kann man durch geeignete Wahl von $\tilde{C}$, b und $\tilde{b}$ in der Abschätzung (3.9) die Differenz zwischen rechter und linker Seite beliebig klein machen, auch wenn man im Allgemeinen nicht 0 erreichen kann, vgl. Wedin (1973, Section 6). Für eine geometrische Interpretation des Terms mit $(\mathrm{cond}_+(C))^2$, die zeigt, dass man diesen quadratischen Terms nicht weglassen kann, und Abschätzungen basierend auf der Norm der Spalten von C bzw. von $\tilde{C}$ siehe van der Sluis (1974).

c) Wegen $(CC^{\mathrm{T}})^+ = (C^+)^{\mathrm{T}}C^+$ kann man den letzten Term in (3.9) auch abschätzen zu

$$\|\tilde{C}-C\|\,\left\|(CC^{\mathrm{T}})^+b\right\| \le \|\tilde{C}-C\|\,\left\|C^+\right\|\,\|x^*\| = \frac{\|\tilde{C}-C\|}{\|C\|}\,\mathrm{cond}_+(C)\,\|x^*\|. \tag{3.10}$$

Analog zu Bemerkung 3.7a) erhält man damit Entwicklungen erster Ordnung. Wenn man $\|b\| = \|Cx^*\| \le \|C\|\,\|x^*\|$ und (3.10) in (3.9) verwendet, so ergibt sich für $x^* \neq 0$ bzw. äquivalent für $b \notin \mathrm{Null}(C^{\mathrm{T}})$

$$\frac{\|\tilde{x}^*-x^*\|}{\|x^*\|} \le \mathrm{cond}_+(C)\left(2\,\frac{\|\tilde{C}-C\|}{\|C\|} + \frac{\|\tilde{b}-b\|}{\|P_Cb\|} + \frac{\|\tilde{b}-b\|}{\|P_Cb\|}\,\frac{\|\tilde{C}-C\|}{\|C\|}\,\mathrm{cond}_+(C)\right.$$
$$\left. + \frac{\|(I_m-P_C)b\|}{\|P_Cb\|}\,\frac{\|\tilde{C}-C\|}{\|C\|}\,\mathrm{cond}_+(C)\right) + O(\|\tilde{C}-C\|^2 + \|\tilde{C}-C\|^2\,\|\tilde{b}-b\|^2).$$

Ist etwa $\|\tilde{C}-C\| / \|C\| < \epsilon$ und $\|\tilde{b}-b\| / \|P_Cb\| < \epsilon$, dann ist

$$\frac{\|\tilde{x}^*-x^*\|}{\|x^*\|} \le \left(3\,\mathrm{cond}_+(C) + \frac{\|(I_m-P_C)b\|}{\|P_Cb\|}(\mathrm{cond}_+(C))^2\right)\epsilon + O(\epsilon^2).$$

Falls $\mathrm{rg}(C) = m$, dann ist $P_C = I_m$ und der Term mit $(\mathrm{cond}_+(C))^2$ verschwindet. Ansonsten ist je nachdem, ob die Norm $\|Cx^*-b\| \overset{(2.12)}{=} \|(I_m-P_C)b\|$ des Residuums groß ist oder nicht, der Term mit $(\mathrm{cond}_+(C))^2$ oder der Term mit $\mathrm{cond}_+(C)$ dominant.

Bemerkung 3.19
Für Resultate, wie sich Störungen in der Koeffizientenmatrix und der rechten Seite, die komponentenweise beschränkt sind, auf die Abweichung in der Lösung auswirken, verweisen wir auf Bauer (1966), Oettli u. Prager (1964) und Björck (1991) sowie auf den Übersichtsartikel Higham (1994). Für Störungen, bei denen z.B. die Matrizen $\tilde{C}, C$ in der Klasse der symmetrischen Matrizen verbleiben, siehe Rump (2003a, b).

Von den zu Beginn des Kapitels angekündigten Algorithmen zur Lösung der Normalgleichungen bestimmen das Verfahren basierend auf der Cholesky-Zerlegung und das Verfahren basierend auf der LR-Zerlegung, die wir in den Abschnitten 3.1.2 und 3.1.3 vorstellen, eine Lösung des Gleichungssystem $Ax = b$ mit $A = X^{\mathrm{T}}\Sigma^{-1}X$ und $b = X^{\mathrm{T}}\Sigma^{-1}(y - y_0)$ und benötigen, dass die Matrix A regulär ist. Die relevante Konditionszahl ist hier also $\mathrm{cond}(A)$.
Das Verfahren basierend auf der QR-Zerlegung und das Verfahren basierend auf der Singulärwertzerlegung, die wir in den Abschnitten 3.1.4 und 3.1.5 vorstellen, bestimmen die Kleinste-Quadrate-Lösung des Gleichungssystems $Bx = a$ mit $\Sigma^{-1} = U^{\mathrm{T}}U$, $B = UX$ und $a = U(y - y_0)$, d.h. sie lösen statt der Normalgleichungen das äquivalente Minimierungsproblem

$$\min_{x\in\mathbb{R}^p} \|Bx - a\| .$$

Die relevante Konditionszahl ist bei diesem Vorgehen $\mathrm{cond}_+(B)$.
Hierbei gilt $\mathrm{cond}_+(A) = (\mathrm{cond}_+(B))^2$. Da außer im Fall, dass alle singulären Werte von B gleich sind, $\mathrm{cond}_+(B) > 1$ gilt, ist die Spektralkonditionszahl von B kleiner als die Spektralkonditionszahl $\mathrm{cond}_+(A)$ von A, wobei ja $\mathrm{cond}(A) = \mathrm{cond}_+(A)$, falls A regulär ist. Da somit der relative Abstand zu Rangverlust für die Matrix A kleiner ist als für die Matrix B, ist es gemäß Satz 3.4 und Satz 3.10 plausibel anzunehmen, dass die numerischen Schwierigkeiten einer Zerlegung von A bei Matrizen mit großer Konditionszahl größer sind als die einer Zerlegung von B und dass das Ergebnis basierend auf einer Zerlegung von A ungenauer ist als das Ergebnis basierend auf einer Zerlegung von B.

Wir illustrieren die unterschiedlichen Konditionszahlen und die Auswirkung der Matrixmultiplikation bei der expliziten Bestimmung der Normalengleichungen an einem Beispiel von derselben Form wie in Läuchli (1961).

Beispiel 3.20
Sei

$$X = \begin{pmatrix} 1 & 1 & 1 \\ \delta & 0 & 0 \\ 0 & \delta & 0 \\ 0 & 0 & \delta \end{pmatrix}$$

mit $\delta \in \mathbb{R}$. Rechnet man, wie in den meisten modernen Computern verwendet, mit 64-bit-Gleitkommazahlen[7] („double precision“) nach IEEE 754-2008, so ergibt die Mul-

[7] Die verwendeten Zahlen haben eine 52-bit-Mantisse und damit ca. 15-16 Stellen Genauigkeit im De-

tiplikation $X^{\mathrm{T}}X$ für $\delta = 10^{-8}$ anstelle von

$$X^{\mathrm{T}}X = \begin{pmatrix} 1+\delta^2 & 1 & 1 \\ 1 & 1+\delta^2 & 1 \\ 1 & 1 & 1+\delta^2 \end{pmatrix}$$

das Ergebnis

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

In diesem Fall ist $\mathrm{cond}_+(X) = \sqrt{3+\delta^2}/|\delta|$ und $\mathrm{cond}(X^{\mathrm{T}}X) = (3+\delta^2)/\delta^2$, d.h. für $\delta = 10^{-8}$ ergibt sich $\mathrm{cond}_+(X) \approx 1.73205 \cdot 10^8$ und $\mathrm{cond}(X^{\mathrm{T}}X) \approx 3 \cdot 10^{16}$. Während die Matrix X für 64-bit-Gleitkommazahlen $|\delta| > 2^{-52} \approx 2.22 \cdot 10^{-16}$ z.B. auch nach Addition von Zeile 1 auf Zeile 2 und Zeile 3 in der Darstellung durch 64-bit-Gleitkommazahlen noch Rang 3 hat, geht dieser Rang für $X^{\mathrm{T}}X$ in dieser Darstellung für 64-bit-Gleitkommazahlen $|\delta| < 2^{-26} \approx 1.49 \cdot 10^{-8}$ und daher in unserem Beispiel schon mit $\delta = 10^{-8}$ verloren.

Da bei allen Verfahren, die wir in diesem Abschnitt noch vorstellen werden, zu Beginn die Matrixmultiplikation $A = X^{\mathrm{T}}\Sigma^{-1}X$ bzw. $B = UX$ anfällt, eine Vorbemerkung zur Matrixmultiplikation.

Bemerkung 3.21
Gegeben seien $C \in \mathbb{R}^{m\times n}$ und $D \in \mathbb{R}^{n\times k}$. Nutzt man für die Multiplikation CD die naive Matrixmultiplikation, so fallen

$$m(n-1)k \quad \text{Additionen und} \quad mnk \quad \text{Multiplikationen}$$

an. Nutzt man den Algorithmus von Strassen (1969), so kann die Anzahl der notwendigen Rechenschritte einer Matrixmultiplikation verringert werden.
Seien m, n, k gerade Zahlen (ansonsten ergänze man die Matrizen um eine Zeile oder Spalte mit 0-Einträgen). Die Idee von Strassen besteht in der Unterteilung der Matrizen C und D in jeweils 4 Blöcke gleicher Größe, d.h.

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

mit Matrizen $C_{11}, C_{12}, C_{21}, C_{22} \in \mathbb{R}^{m/2\times n/2}$ und

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$$

zimalsystem, wenn man vom Problem des Exponentenüberlaufs oder Exponentenunterlaufs absieht; genauer ist die Maschinengenauigkeit $\epsilon_{\mathrm{Mach}} = 2^{-52}$. Die Maschinengenauigkeit ist dabei definiert als die kleinste verwendete Gleitkommazahl $\tau > 0$, so dass $1+\tau$ bei einer Berechnung in den verwendeten Gleitkommazahlen noch von 1 unterschieden werden kann.

mit Matrizen $D_{11}, D_{12}, D_{21}, D_{22} \in \mathbb{R}^{n/2 \times k/2}$. Setzt man

$$\begin{aligned}
M_1 =&(C_{11} + C_{22})(D_{11} + D_{22}),\\
M_2 =&(C_{21} + C_{22})D_{11},\\
M_3 =&C_{11}(D_{12} - D_{22}),\\
M_4 =&C_{22}(D_{21} - D_{11}),\\
M_5 =&(C_{11} + C_{12})D_{22},\\
M_6 =&(C_{21} - C_{11})(D_{11} + D_{12}),\\
M_7 =&(C_{12} - C_{22})(D_{21} + D_{22}),\\
F_{11} =&M_1 + M_4 - M_5 + M_7,\\
F_{12} =&M_3 + M_5,\\
F_{21} =&M_2 + M_4,\\
F_{22} =&M_1 + M_3 - M_2 + M_6,
\end{aligned}$$

so ergibt sich

$$CD = \begin{pmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{pmatrix}.$$

Zur Berechnung der Matrizen $F_{11}, F_{12}, F_{21}, F_{22}$ werden also 7 Multiplikationen von $m/2 \times n/2$ mit $n/2 \times k/2$ Matrizen und 18 Additionen von Matrizen unterschiedlicher Bauart, davon 5 Additionen zweier $m/2 \times n/2$ Matrizen, 5 Additionen zweier $n/2 \times k/2$ Matrizen und 8 Additionen zweier $m/2 \times k/2$ Matrizen benötigt. Benutzt man für die einzelnen Blöcke die naive Matrixmultiplikation, so fallen folglich insgesamt

$$\frac{7}{8}mnk \quad \text{Multiplikationen}$$

und

$$\frac{7}{8}m(n-2)k + \frac{5}{4}mn + \frac{5}{4}nk + \frac{8}{4}mk \quad \text{Additionen}$$

an. Einmalige Anwendung der Idee von Strassen verkleinert also die Zahl der Rechenoperationen für große m, n, k um den Faktor $\frac{7}{8}$. Bei rekursiver Anwendung der Unterteilung in Blockmatrizen kann man für $m = n = k$ zeigen, dass die Zahl der Rechenoperationen (Additionen und Multiplikationen zusammen) kleiner als $4m^{\log_2 7}$ ist, wobei $\log_2 7 \approx 2.807$, siehe Higham (2002, Section 23.1).

Dabei zeigen numerische Experimente, dass sich (im Fall $m = n = k$) ca. ab $m \geq 100$ die Rechenzeit durch die Verwendung des Algorithmus von Strassen gegenüber der naiven Matrixmultiplikation tatsächlich verkürzt, während für kleinere Matrizen der zusätzliche Speicherverwaltungsaufwand die Zeitersparnis durch die geringere Anzahl an Rechenoperationen zunichte macht. Bei praktischen Implementierungen etwa im Fall $m = n = k$ wendet man daher die Unterteilung in Teilmatrizen an, bis die Spalten- bzw. Zeilenzahl kleiner als 100 ist und verwendet dann die naive Matrixmultiplikation. Hierzu und für weitergehende Ausführungen zur schnellen Matrixmultiplikation vergleiche Higham (2002, Section 23.1) und Skiena (1998, Chapter 8.2).

Verwendet man hingegen bei der Matrixmultiplikation auf Parallelrechner zugeschnittene Algorithmen, so sind andere Überlegungen zur Komplexität nötig, für die wir z.B. auf Golub u. Van Loan (1996, Chapter 6) verweisen.

Nun zu vier Verfahren zur Lösung der Normalgleichungen

$$X^{\mathrm{T}}\Sigma^{-1}X(\vartheta^* - \vartheta_0) = X^{\mathrm{T}}\Sigma^{-1}(y - y_0). \tag{3.11}$$

Wir werden bei den Verfahren annehmen, dass $\mathrm{rg}(X) = p$ gilt, außer bei dem Verfahren basierend auf der Singulärwertzerlegung, siehe Abschnitt 3.1.5. Nur dort werden wir also die Mehrdeutigkeit der Lösung von (3.11) zulassen und bestimmen dann die Lösung ϑ^* mit der minimaler Norm $\|\vartheta^* - \vartheta_0\|$.

Wir werden der Einfachheit halber bei den folgenden Komplexitätsbetrachtungen außer beim Gesamtvergleich am Ende des Abschnitts jeweils mit der naiven Matrixmultiplikation rechnen.

Ferner werden wir annehmen, dass die Multiplikation mit Σ^{-1} einfach ist in dem Sinne, dass entweder die Bestimmung von Σ^{-1} bereits geschehen ist oder man eine Cholesky-Zerlegung von Σ kennt, und werden daher die Kosten einer Bestimmung von Σ^{-1} ignorieren. Bei den Verfahren basierend auf einer QR-Zerlegung bzw. auf einer Singulärwertzerlegung werden wir sogar annehmen, dass man schon eine Zerlegung

$$\Sigma^{-1} = U^{\mathrm{T}}U$$

mit $U \in \mathbb{R}^{N\times N}$ kennt, etwa durch eine Cholesky-Zerlegung, wobei allerdings U nicht notwendigerweise eine Dreiecksmatrix sein muss.

Für den Fall, dass Σ schlecht konditioniert ist und daher die Bestimmung von U numerisch instabil ist, verweisen wir auf Paige (1979a, b) für entsprechende Verfahren, die ohne Verwendung von U auskommen.

3.1.2. Verfahren basierend auf der Cholesky-Zerlegung

1. Berechne $A = X^{\mathrm{T}}\Sigma^{-1}X$ und $b = X^{\mathrm{T}}\Sigma^{-1}(y - y_0)$.
2. Bestimme eine Cholesky-Zerlegung $A = LL^{\mathrm{T}}$ (L ist linke Dreiecksmatrix).
3. Löse zwei Dreiecksgleichungssysteme
   a) $Lz^* = b$ nach z^* und
   b) $L^{\mathrm{T}}x^* = z^*$ nach x^*.

   Lösung: $\vartheta^* = x^* + \vartheta_0$.

Komplexität: (für Schritt 2 siehe Hämmerlin u. Hoffmann (1991, Abschnitt 2.2.3))

Rechenschritt	Additionen/ Subtraktionen	Multiplikationen/ Divisionen	Quadratwurzeln
Schritt 1			
$X^{\mathrm{T}}\Sigma^{-1}$	$p(N-1)N$	pN^2	
$(X^{\mathrm{T}}\Sigma^{-1})X$	$p(N-1)p$	p^2N	
$(X^{\mathrm{T}}\Sigma^{-1})(y-y_0)$	$p(N-1)+N$	pN	
Schritt 2	$\frac{1}{6}(p^3-p)$	$\frac{1}{6}(p^3+3p^2-4p)$	p
Schritt 3			
$Lz^*=b$	$\frac{1}{2}p(p-1)$	$\frac{1}{2}p(p+1)$	
$L^{\mathrm{T}}x^*=z^*$	$\frac{1}{2}p(p-1)$	$\frac{1}{2}p(p+1)$	
insgesamt	$pN^2+p^2N+\frac{1}{6}p^3$ $+O(p^2+pN)$	$pN^2+p^2N+\frac{1}{6}p^3$ $+O(p^2+pN)$	p

Nutzt man die Symmetrie der Matrix $A=(X^{\mathrm{T}}\Sigma^{-1})X$, so kann man in der zweiten Zeile $p(N-1)p$ Additionen auf $\frac{1}{2}p(N-1)(p+1)$ Additionen und p^2N Multiplikationen auf $\frac{1}{2}p(p+1)N$ Multiplikationen reduzieren. Ist Σ^{-1} eine Diagonalmatrix, so sinkt die Komplexität in der ersten Zeile auf 0 Additionen und pN Multiplikationen.

Wenn die Matrix A große Konditionszahl hat und der Abstand zu einer singulären Matrix klein ist, und damit auch der Abstand zu den Matrizen, die nicht positiv semidefinit sind, klein ist, dann steigt das Risiko, dass die Bestimmung der Cholesky-Zerlegung numerisch nicht möglich ist, typischerweise weil bei einigen Rechenschritten, bei denen Wurzeln gezogen werden müssen, durch Rechenungenauigkeiten keine nichtnegativen Zahlen mehr vorliegen.[8]
Bemerkenswert ist in diesem Zusammenhang ein Satz von Wilkinson (1968, Theorem 2) zur numerischen Cholesky-Zerlegung bei kleiner Konditionszahl:

Sei A eine $p\times p$ Matrix, deren Einträge Gleitkommazahlen in Binärdarstellung mit einer Mantisse der Länge t und beliebig großem oder kleinem Exponenten sind, und bei Additionen, Subtraktionen, Multiplikationen und Divisionen werde fortlaufend in dieser Gleitkommadarstellung gerechnet, d.h. die Maschinengenauigkeit beträgt $\epsilon_{\mathrm{Mach}}=2^{-t}$. Ferner sei $t_1=t-\log_2(1.06)$ und für Gleitkommazahlen $a>0$ gelte für das Ergebnis $x(a)$ der Quadratwurzeloperation, dass $(x(a))^2/a\in[1-2.2^{-t_1},1+2.2^{-t_1}]$. Dann kann die Cholesky-Zerlegung von A numerisch bestimmt werden, falls

$$2p^{3/2}2^{-t_1}\operatorname{cond}(A)<0.1$$

und für den berechneten Cholesky-Faktor $\tilde{L}$ gilt

$$\left\|A-\tilde{L}\tilde{L}^{\mathrm{T}}\right\|\le 2.2^{-t_1}\left(1+(\sqrt{p}+2.2)2^{-t_1}\right)^p\left(\frac{2}{3}(p+1)^{3/2}+1.1p\right)\|A\|\,. \tag{3.12}$$

Im Fall $p>10$ kann man (3.12) weiter abschätzen zu

$$\|A-\tilde{L}\tilde{L}^{\mathrm{T}}\|\le 2.5p^{3/2}2^{-t_1}\|A\|\,.$$

[8] Zur Bestimmung einer (verallgemeinerten) Cholesky-Zerlegung im singulären Fall vergleiche Higham (1990).

3.1.3. Verfahren basierend auf der LR-Zerlegung mittels Gaußscher Elimination

1. Berechne $A = X^{\mathrm{T}}\Sigma^{-1}X$ und $b = X^{\mathrm{T}}\Sigma^{-1}(y - y_0)$.
2. Bestimme eine LR-Zerlegung $PA = LR$ (P Permutationsmatrix, L ist linke Dreiecksmatrix, R rechte Dreiecksmatrix) mittels Gauß-Algorithmus mit (Zeilen-/Spalten-/Turm- oder vollständiger) Pivotisierung. Berechne $c = Pb$.
3. Löse zwei Dreiecksgleichungssysteme
   a) $Lz^* = c$ nach z^* und
   b) $Rx^* = z^*$ nach x^*.

   Lösung: $\vartheta^* = x^* + \vartheta_0$.

Komplexität: (für Schritt 2 siehe Hämmerlin u. Hoffmann (1991, Abschnitt 2.1.6))

Rechenschritt	Additionen/ Subtraktionen	Multiplikationen/ Divisionen
Schritt 1		
$X^{\mathrm{T}}\Sigma^{-1}$	$p(N-1)N$	pN^2
$(X^{\mathrm{T}}\Sigma^{-1})X$	$p(N-1)p$	p^2N
$(X^{\mathrm{T}}\Sigma^{-1})(y-y_0)$	$p(N-1)+N$	pN
Schritt 2	$\frac{1}{3}p^3 - \frac{1}{2}p^2 + \frac{1}{6}p$	$\frac{1}{3}p^3 - \frac{1}{2}p$
Schritt 3		
$Lz^* = b$	$\frac{1}{2}p(p-1)$	$\frac{1}{2}p(p-1)$
$Rx^* = z^*$	$\frac{1}{2}p(p-1)$	$\frac{1}{2}p(p+1)$
insgesamt	$pN^2 + p^2N + \frac{1}{3}p^3 + O(p^2 + pN)$	$pN^2 + p^2N + \frac{1}{3}p^3 + O(p^2 + pN)$

Der wesentliche Unterschied in der Komplexität zum Verfahren basierend auf der Cholesky-Zerlegung liegt darin, dass Schritt 2 der Matrixzerlegung von A ca. doppelt so aufwendig ist. Die entsprechenden Bemerkungen aus Abschnitt 3.1.2 zur Komplexität bei Ausnutzung der Symmetrie von A oder im Fall einer Diagonalmatrix Σ^{-1} gelten analog.

Wenn die Matrix A schlecht konditioniert ist, d.h. cond(A) groß ist, dann kann die LR-Zerlegung numerisch versagen.
Sie ist aber bei der Anwendung von Pivotisierungsstrategien in der Regel numerisch stabiler als die Cholesky-Zerlegung, ohne Pivotisierung besteht in numerischer Hinsicht kaum ein Unterschied.[9] Einen Kompromiss zwischen Rechenaufwand und numerischer

[9] Für entsprechende Modifizierungen des Cholesky-Verfahrens durch Pivotsuche vergleiche Fang u. O'Leary (2008).

Stabilität stellt dabei die Rook-Pivotisierung[10] dar, vgl. Higham (2002, Chapter 9). Zwar fallen im Worst Case wie bei kompletter Pivotsuche auch insgesamt $O(p^3)$ Vergleiche an, unter einem sehr speziellen stochastischen Modell für die Matrizen $A \in \mathbb{R}^{p\times p}$ kann man aber zeigen, dass der Erwartungswert für die Anzahl der Vergleiche kleiner als $e/2(p-1)p \approx 1.36(p-1)p$ ist, also von der gleichen Ordnung wie bei Spaltenpivotsuche oder Zeilenpivotsuche, bei denen jeweils $p(p-1)/2$ Vergleiche anfallen. Diese Ordnungsbeziehung bestätigt sich auch unter einem anderen stochastischen Modell und in Simulationen unter weiteren stochastischen Modellen, vgl. Foster (1997, 1998) und Poole u. Neal (2000). Mit der Begründung, dass wegen der positiven Definitheit von A die Absolutbeträge der Einträge der Matrizen A_k nach den Zeilenumformungen im k-ten Schritt ($k = 1, \ldots, p-1$, d.h. $A_{p-1} = R$) durch das Maximum der Absolutbeträge der Einträge von A beschränkt sind (vgl. Higham (2002, Problem 10.6)), wird auch ein Verzicht auf eine Pivotisierungsstrategie befürwortet, vgl. Higham (2002, Section 9.3) oder Reid (1971).

Für die LR-Zerlegung ist kein Resultat bekannt, das die Norm von $A-\tilde{L}\tilde{R}$ als Ausdruck in p, der Norm von A und der Mantisse t der verwendeten Gleitkommazahlen beschränkt, also kein Analogon zum Satz von Wilkinson bei der Cholesky-Zerlegung (siehe S. 53). Für Aussagen, in denen die Einträge von $A - \tilde{L}\tilde{R}$ durch die Einträge von $\tilde{L}$ und $\tilde{R}$ beschränkt werden, siehe z.B. Higham (2002, Section 9.3, 9.8 und 9.13).

3.1.4. Verfahren basierend auf der QR-Zerlegung

Wir betrachten hier nur den Fall $N \geq p$ und $\mathrm{rg}(X) = p$ und formulieren das Vorgehen für die QR-Zerlegung mit Pivotisierung.[11] Falls keine Pivotisierung stattfindet, also im Folgenden $P = I_p$ ist, entfällt Schritt 4b).[12]

1. Berechne $B = UX$.
2. Bestimme eine QR-Zerlegung $BP = Q\binom{R}{\mathbf{0}}$ mit der Nullmatrix $\mathbf{0}_{(N-p)\times p}$, einer rechten Dreiecksmatrix $R \in \mathbb{R}^{p\times p}$, einer orthogonalen Matrix $Q \in \mathbb{R}^{N\times N}$ und einer Permutationsmatrix $P \in \mathbb{R}^{p\times p}$.
   (Insbesondere ist R invertierbar. Schreibt man $Q = (Q_1\, Q_2)$, wobei Q_1 die ersten p Spalten von Q sind, so gilt $BP = Q_1 R$.)

[10] So genannt, weil das Verfahren zum Aufsuchen des Pivotelements den Zugmöglichkeiten eines Turms auf einem Schachbrett entspricht und „Rook“ die englische Bezeichnung für einen Turm im Schach ist. Auf diese Weise erreicht man ein Pivotelement, das sowohl in seiner Spalte als auch in seiner Zeile maximalen Betrag hat.

[11] Im Fall $N < p$ bestimmt man eine QR-Zerlegung $AP = Q\,(R\ S)$ mit einer Permutationsmatrix $P \in \mathbb{R}^{p\times p}$, einer orthogonalen Matrix $Q \in \mathbb{R}^{N\times N}$, einer rechten Dreiecksmatrix $R \in \mathbb{R}^{N\times N}$ und einer Matrix $S \in \mathbb{R}^{N\times(p-N)}$. Die Normalgleichungen sind aber nicht eindeutig lösbar und man könnte die allgemeine Lösung zwar daraus bestimmen, aber dies ist einfacher mit Hilfe einer Singulärwertzerlegung - vor allem, wenn man an der Minimum-Norm-Lösung $\hat{\vartheta}_{MP}(y)$ aus (2.9) interessiert ist. Wir werden diesen Weg daher nicht weiter verfolgen und verweisen auf Björck (1990, Remark 7.5 und Section 11), wie man mit einer QR-Zerlegung die Minimum-Norm-Lösung ermittelt.

[12] Dieses Verfahren ohne Pivotisierung zur Lösung des linearen Kleinste-Quadrate-Problems wurde im Detail zuerst von Golub (1965) vorgeschlagen und heißt daher auch Golub-Methode.

3. Berechne $c = Q_1^{\mathrm{T}} U(y - y_0)$.
4. Löse das Gleichungssystem
   a) $Rz^* = c$ nach z^* und permutiere
   b) $x^* = Pz^*$.

   Lösung: $\vartheta^* = x^* + \vartheta_0$.

Die QR-Zerlegung verschlechtert die Kondition der zu zerlegenden Matrix nicht, wie dies bei den Verfahren basierend auf der Cholesky- bzw. auf der LR-Zerlegung schon durch Bildung von $A = X^{\mathrm{T}}\Sigma^{-1}X$ geschieht.

Wir geben nun die Anzahl der benötigten Rechenoperationen des obigen Kleinste-Quadrate-Verfahrens an, falls $N > p$ gilt und das QR-Verfahren ohne Pivotstrategie mit Hilfe von Householder-Matrizen erfolgt. Im Fall $N = p$ ändern sich nur im Schritt 2 und im Schritt 3 Glieder kleinerer Ordnung. Für Detailberechnungen verweisen wir auf Anhang A.4.

Komplexität:

Rechenschritt	Additionen/ Subtraktionen	Multiplikationen/ Divisionen	Quadratwurzeln
Schritt 1:	$p(N-1)N$	pN^2	
Schritt 2	$Np^2 - \frac{1}{3}p^3 + \frac{1}{3}p$	$Np^2 - \frac{1}{3}p^3 + p^2 + \frac{4}{3}p$	p
Schritt 3	$3Np + N - p^2 - p$	$3Np - p^2 + 2p$	
Schritt 4 $Rx^* = c$	$\frac{1}{2}p(p-1)$	$\frac{1}{2}p(p+1)$	
insgesamt	$pN^2 + p^2N - \frac{1}{3}p^3 + O(p^2 + pN)$	$pN^2 + p^2N - \frac{1}{3}p^3 + O(p^2 + pN)$	p

Zwar benötigt die QR-Zerlegung in Schritt 2 selbst im Fall von $N = p$ ca. doppelt so viele Rechenoperationen wie die LR-Zerlegung, die anderen Schritte wiegen diesen Nachteil allerdings wieder auf, so dass von den führenden Gliedern dritter Ordnung in p und N die Koeffizienten zu pN^2 und p^2N gleich sind und der Koeffizient zu p^3 sogar kleiner ist als beim Verfahren basierend auf der LR-Zerlegung.[13] Mit der Pivotisierungsstrategie fallen allerdings insgesamt $\frac{3}{2}N^2p + Np^2 - \frac{1}{2}p^3 + O(p^2 + pN)$ Additionen/Subtraktionen und $\frac{3}{2}N^2p + Np^2 - \frac{1}{2}p^3 + O(p^2 + pN)$ Multiplikationen an.

In Higham (2002, Section 19.3 und 19.4) finden sich Resultate zur Berechnung der QR-Zerlegung bei Verwendung von Gleitkommazahlen mit endlicher Mantisse, die dem Satz von Wilkinson bei der Cholesky-Zerlegung (siehe S. 53) entsprechen.

Falls noch größere numerische Stabilität gewünscht ist oder falls $\mathrm{rg}(X) < p$, dann empfiehlt sich ein Algorithmus basierend auf der Singulärwertzerlegung:

[13] Kommt allerdings die schnelle Matrixmultiplikation gewinnbringend zum Einsatz, so profitieren das Verfahren basierend auf der Cholesky-Zerlegung und das Verfahren basierend auf der LR-Zerlegung mehr als das Verfahren basierend auf der QR-Zerlegung.

3.1.5. Verfahren basierend auf der Singulärwertzerlegung

1. Berechne $B = UX$.
2. Im Fall $N \geq p$:
   Bestimme eine Singulärwertzerlegung $B = W\binom{D}{\mathbf{0}}V^{\mathrm{T}}$ mit einer orthogonalen Matrix $W \in \mathbb{R}^{N\times N}$, einer orthogonalen Matrix $V \in \mathbb{R}^{p\times p}$, einer Diagonalmatrix $D = \operatorname{diag}(d_1, \ldots, d_p) \in \mathbb{R}^{p\times p}$ mit Einträgen $d_1 \geq d_2 \geq d_p \geq 0$ und einer Nullmatrix $\mathbf{0}_{(N-p)\times p}$. Ferner sei $r = \#\{i \in \{1, \ldots, p\} : d_i > 0\}$.
   Im Fall $N < p$:
   Bestimme eine Singulärwertzerlegung $B = W\,(D\ \mathbf{0})\,V^{\mathrm{T}}$ mit einer orthogonalen Matrix $W \in \mathbb{R}^{N\times N}$, einer orthogonalen Matrix $V \in \mathbb{R}^{p\times p}$, einer Diagonalmatrix $D = \operatorname{diag}(d_1, \ldots, d_N) \in \mathbb{R}^{N\times N}$ mit Einträgen $d_1 \geq d_2 \geq d_N \geq 0$ und einer Nullmatrix $\mathbf{0}_{N\times(p-N)}$. Ferner sei $r = \#\{i \in \{1, \ldots, N\} : d_i > 0\}$.
3. Berechne $c = U(y - y_0)$.
4. Berechne:
$$x^* = \sum_{i=1}^{r} \frac{w_i^{\mathrm{T}} c}{d_i}\, v_i. \tag{3.13}$$
   Hierbei bezeichnen v_i, $i = 1, \ldots, p$, die Spalten von V und w_i, $i = 1, \ldots, N$, die Spalten von W.
   Lösung: $\vartheta^* = x^* + \vartheta_0$.

Wir geben eine kurze Begründung für die Korrektheit des Verfahrens, d.h. für Formel (3.13). Sei $r = \operatorname{rg}(X)$ und sei $B = W\tilde{D}V^{\mathrm{T}}$ eine Singulärwertzerlegung mit singulären Werten $d_1 \geq d_2 \geq d_r > 0$, d.h. im Fall $N \geq p$ gilt $\tilde{D} = \binom{D}{\mathbf{0}}$ und im Fall $N < p$ gilt $\tilde{D} = (D\ \mathbf{0})$. Da für $\eta(\vartheta) = X \cdot (\vartheta - \vartheta_0) + y_0$ die Beziehung

$$\begin{aligned}\operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^p} \|y - \eta(\vartheta)\|_\Sigma^2 &= \operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^p} \|UX \cdot (\vartheta - \vartheta_0) - U(y - y_0)\|^2 \\ &= \operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^p} \|B \cdot (\vartheta - \vartheta_0) - c\|^2 \end{aligned} \tag{3.14}$$

gilt, ergibt sich für die Lösung ϑ^* mit der minimalen Norm $\|\vartheta^* - \vartheta_0\|$, siehe (2.9), dass

$$\vartheta^* = \vartheta_0 + B^+ c = \vartheta_0 + V\tilde{D}^+ W^{\mathrm{T}} c = \vartheta_0 + \sum_{i=1}^{r} \frac{w_i^{\mathrm{T}} c}{d_i}\, v_i.$$

Da für das System $v_1, \ldots, v_p$ der orthogonalen Eigenvektoren von $B^{\mathrm{T}}B = X^{\mathrm{T}}\Sigma^{-1}X$ gebildet durch die Spalten von V gilt, dass $Bv_i = \mathbf{0}_N$ für $i = r+1, \ldots, p$ gilt, folgt

$$\mathcal{L}(y) = \left\{ \vartheta_0 + \sum_{i=1}^{r} \frac{w_i^{\mathrm{T}} c}{d_i}\, v_i + \sum_{i=r+1}^{p} t_i v_i : t_i \in \mathbb{R} \text{ für } i = r+1, \ldots, p \right\}$$

für die allgemeine Lösung von (3.14) aus (2.8).

Falls $r = \mathrm{rg}(X)$ nicht bekannt ist, so ist für ein geeignetes $\delta > 0$ der sogenannte numerische δ-Rang

$$\hat{r} = \#\{i \in \{1, \ldots, \min(p, N)\} : d_i > \delta\} \tag{3.15}$$

der Matrix X eine zweckmäßige Wahl für r, vgl. Björck (1990, Section 10), d.h. man setzt die Einträge $d_{r+1}, \ldots, d_p$ bzw. $d_{r+1}, \ldots, d_N$ in der im Schritt 2 numerisch berechneten Diagonalmatrix $D = \mathrm{diag}(d_1, \ldots, d_p)$ bzw. $D = \mathrm{diag}(d_1, \ldots, d_N)$ auf 0. In Gleichung (3.15) empfehlen Golub u. Van Loan (1996, Section 5.5.8) $\delta = 10^{-t} \|A\|_\infty$ zu verwenden, wenn die Matrixeinträge auf t Dezimalstellen genau sind, bzw. $\delta = \epsilon_{\mathrm{Mach}} \|A\|_\infty$ zu verwenden, wenn die Ungenauigkeit in den Einträgen nur von der Verwendung von Gleitkommazahlen herrührt.

Abschließend betrachten wir die Zahl der Rechenschritte, die das Verfahren im Fall $N \geq p$ bei einer Matrix X mit Rang $r = \mathrm{rg}(X)$ benötigt; im Fall $N < p$ sind bei der Singulärwertzerlegung in den Formeln N und p entsprechend zu vertauschen. Für die zweite Phase der Singulärwertzerlegung ist allerdings keine obere Schranke für die Anzahl der Rechenoperationen bekannt. Daher finden sich in der folgenden Tabelle bei Phase 2 von Schritt 2, bei der D explizit und W und V implizit als Produkt von Householder-Matrizen und Givens-Matrizen bestimmt werden, kein Eintrag. Da in Schritt 4 die entsprechenden Informationen aus den Matrizen W und V bestimmt werden müssen, findet sich ebenfalls kein Eintrag. Rechnet man mit Annahmen aus numerischen Experimenten, dann ergeben sich im Erwartungswert insgesamt $2pN^2 + 4Np^2 + 8p^3 + O(p^2 + pN)$ Rechenoperationen bzw. $2pN^2 + 2Np^2 + 11p^3 + O(p^2 + pN)$ Rechenoperationen bei einem Alternativ-Verfahren. Genauere Ausführen finden sich hierzu im Anhang A.5.

Für die Anzahl der Rechenoperation, wie sie in der Tabelle aufgelistet sind, verweisen wir für Details ebenfalls auf Anhang A.5.

Komplexität:

Rechenschritt	Additionen/ Subtraktionen	Multiplikationen/ Divisionen	Quadratwurzeln
Schritt 1:	$p(N-1)N$	pN^2	
Schritt 2:			
Phase 1	$2Np^2 - \frac{2}{3}p^3$ $-2Np + p^2 + \frac{2}{3}p - 2$	$2Np^2 - \frac{2}{3}p^3$ $+p^2 - 4N + \frac{20}{3}p - 8$	$2p - 2$
Phase 2	/	/	/
Schritt 3	N^2	N^2	
Schritt 4	/	/	
insgesamt	$\geq pN^2 + 2Np^2 - \frac{2}{3}p^3$ $+O(p^2 + pN)$	$\geq pN^2 + 2Np^2 - \frac{2}{3}p^3$ $+O(p^2 + pN)$	$2p - 2$

Für einen Vergleich der Anzahl der Rechenoperationen bei den vier vorgestellten Verfahren beschränken wir uns auf die Glieder 3. Ordnung in N und p und definieren

$$\begin{aligned}
T_{\text{Chol}}(p,N) &= 2pN^2 + 2p^2N + \frac{1}{3}p^3,\\
T_{\text{LR}}(p,N) &= 2pN^2 + 2p^2N + \frac{2}{3}p^3,\\
T_{\text{QR}}(p,N) &= 2pN^2 + 2p^2N - \frac{2}{3}p^3,\\
T_{\text{QR-Pivot}}(p,N) &= 3pN^2 + 2p^2N - p^3,\\
T_{\text{SWZ}}(p,N) &= \min(2pN^2 + 4Np^2 + 8p^3, 2pN^2 + 2Np^2 + 11p^3),
\end{aligned}$$

d.h. das Verfahren basierend auf der Cholesky-Zerlegung benötigt $T_{\text{Chol}}(p,N) + O(p^2 + pN)$ Rechenoperationen. Die Graphen der Funktionen $T_{\text{Chol}}(p,\cdot)$, $T_{\text{LR}}(p,\cdot)$, $T_{\text{QR}}(p,\cdot)$, $T_{\text{QR-Pivot}}(p,\cdot)$ und $T_{\text{SWZ}}(p,\cdot)$ finden sich in Abbildung 3.1 (oben). Kommt schnelle Matrixmultiplikation zum Einsatz, etwa durch dreimalige Unterteilung in Teilmatrizen, so reduzieren sich die Glieder 3. Ordnung zu

$$\begin{aligned}
T^{\text{sM}}_{\text{Chol}}(p,N) &= 2\left(\frac{7}{8}\right)^3 pN^2 + 2\left(\frac{7}{8}\right)^3 p^2N + \frac{1}{3}p^3,\\
T^{\text{sM}}_{\text{LR}}(p,N) &= 2\left(\frac{7}{8}\right)^3 pN^2 + 2\left(\frac{7}{8}\right)^3 p^2N + \frac{2}{3}p^3,\\
T^{\text{sM}}_{\text{QR}}(p,N) &= 2\left(\frac{7}{8}\right)^3 pN^2 + 2p^2N - \frac{2}{3}p^3,\\
T^{\text{sM}}_{\text{QR-Pivot}}(p,N) &= \left(2\left(\frac{7}{8}\right)^3 + 1\right) pN^2 + 2p^2N - p^3,\\
T^{\text{sM}}_{\text{SWZ}}(p,N) &= \min\left(2\left(\frac{7}{8}\right)^3 pN^2 + 4Np^2 + 8p^3, 2\left(\frac{7}{8}\right)^3 pN^2 + 2Np^2 + 11p^3\right)
\end{aligned}$$

und die entsprechenden Graphen finden sich in Abbildung 3.1 (unten).

Man erkennt, dass die Unterschiede zwischen den Verfahren basierend auf Cholesky-Zerlegung, der LR-Zerlegung und der QR-Zerlegung (ohne Pivotisierung) gering sind und nur bei Anwendung der schnellen Matrixmultiplikation das Cholesky-Verfahren und das LR-Verfahren weniger Rechenoperationen als das QR-Verfahren benötigen. Das QR-Verfahren mit Pivotisierung benötigt für kleines N nicht wesentlich mehr Rechenoperationen als die Verfahren basierend auf der Cholesky-Zerlegung, der LR-Zerlegung und der QR-Zerlegung (ohne Pivotisierung), bei großem N/p sind es aber fast 50 Prozent mehr Rechenoperationen. Das Verfahren basierend auf der Singulärwertzerlegung benötigt zwar deutlich mehr Rechenoperationen als die Verfahren basierend auf der Cholesky-Zerlegung, der LR-Zerlegung und der QR-Zerlegung (ohne Pivotisierung), schneidet aber bei großem N/p wesentlich besser ab als das Verfahren basierend auf der QR-Zerlegung mit Pivotisierung. Da das Verfahren basierend auf der QR-Zerlegung die besseren numerischen Eigenschaften hat als die Verfahren basierend auf der Cholesky-Zerlegung und

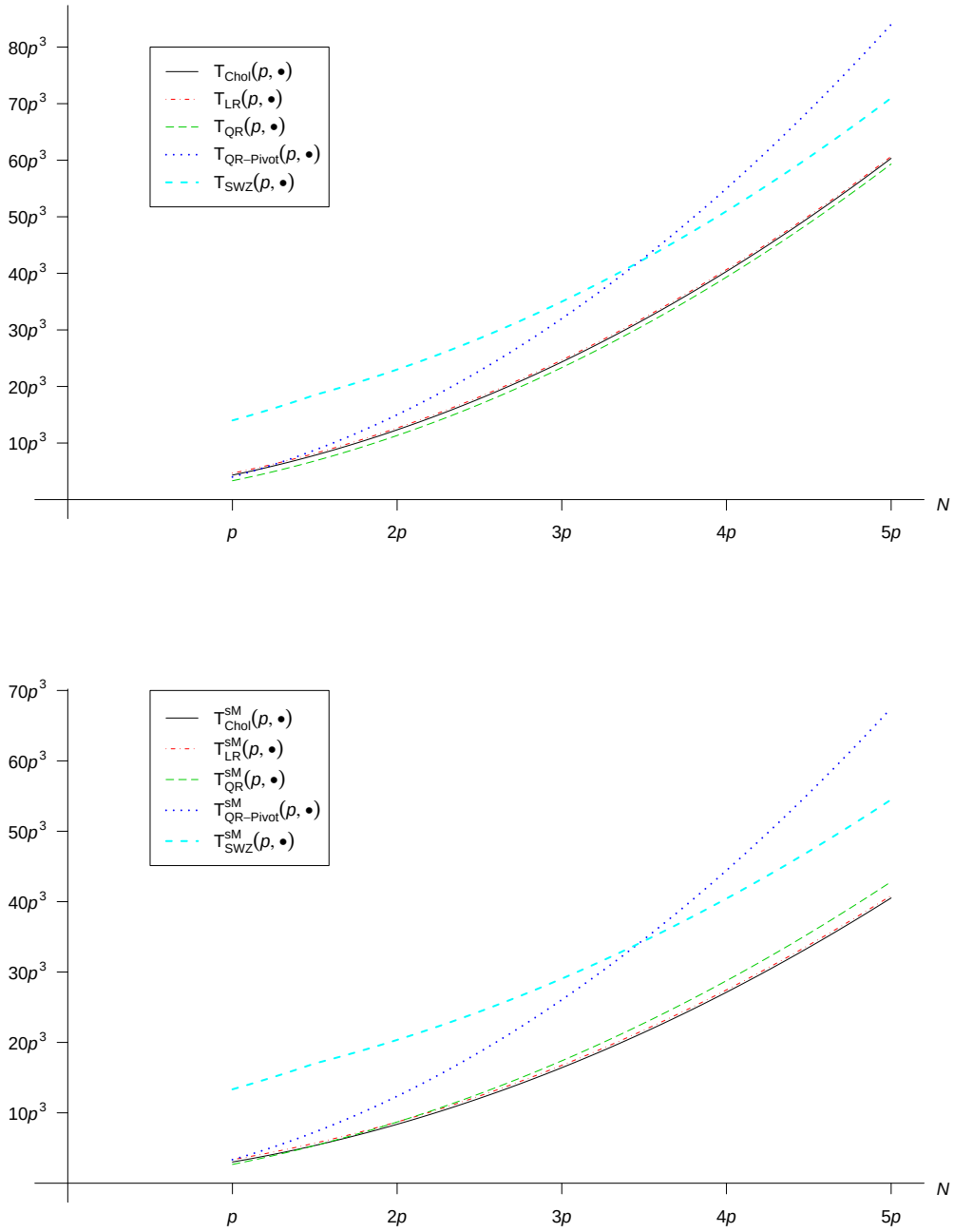


Abbildung 3.1.: Rechenoperationen der vier Verfahren zur Bestimmung der Kleinste-Quadrate-Lösung im linearen Fall in Abhängigkeit von N.
In der unteren Abbildung dreimalige Unterteilung in Teilmatrizen und Einsatz der schnellen Matrixmultiplikation, oben ohne Verwendung der schnellen Matrixmultiplikation.

der LR-Zerlegung, ist es diesen daher in der Regel vorzuziehen, für moderates N/p kann man sogar die QR-Zerlegung mit Pivotisierung erwägen. Für Probleme mit schlechter Konditionszahl oder Rangdefekt ist das Verfahren basierend auf der Singulärwertzerlegung das Verfahren der Wahl.

Für weitergehende Ausführungen zum Kleinste-Quadrate-Problem im linearen Modell vergleiche Björck (1996, 2004) und Higham (2002, Chapter 20).

3.2. Nichtlineares Modell

Wir betrachten in diesem Abschnitt numerische Verfahren zur Lösung des Kleinste-Quadrate-Problems (2.15), wobei wir voraussetzen, dass $\eta : \mathbb{R}^p \to \mathbb{R}^N$ stetig differenzierbar ist. Für den Fall, dass η nicht-differenzierbar ist, verweisen wir auf Fletcher (1987), Bertsekas (1999, Chapter 6) und Osborne (1985). Für Methoden, die zwar die Differenzierbarkeit voraussetzen, aber ohne Ableitungen arbeiten, siehe z.B. Bazaraa u. a. (2006, Chapter 8) und Bertsekas (1999, Chapter 1), und für Methoden zur numerischen Berechnung von Ableitungen siehe Nocedal u. Wright (1999, Chapter 7).

Für die Verfahren, die wir betrachten, benötigen wir den Gradienten und teilweise die Hesse-Matrix von

$$S^{\eta,y} : \Theta \to \mathbb{R},\ S^{\eta,y}(\vartheta) = \|y - \eta(\vartheta)\|_\Sigma^2$$

und damit auch zweimalige Differenzierbarkeit von η. Die erste Ableitung haben wir bereits bei der Herleitung der Normalgleichungen (2.19) bestimmt; es gilt

$$\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta) = 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - y\right). \tag{3.16}$$

Für die Hesse-Matrix ergibt sich

$$\operatorname{Hess} S^{\eta,y}(\vartheta) = 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + 2\sum_{j=1}^{N}\left(\Sigma^{-1}\left(\eta(\vartheta) - y\right)\right)_j \operatorname{Hess}\eta_j(\vartheta). \tag{3.17}$$

Wir behandeln im folgenden die Situation $\Theta = \mathbb{R}^p$. In diesem Abschnitt betrachten wir Verfahren, die Folgen $(\vartheta_k)_{k\in\mathbb{N}_0}$ liefern, die ausgehend von einem geeigneten Startwert ϑ_0 durch

$$\vartheta_{k+1} = \vartheta_k + \lambda_k d_k,\ \text{für } k = 0, 1, \ldots$$

rekursiv definiert sind. Hierbei wird im k-ten Schritt zuerst, je nach Verfahren auf unterschiedliche Weise, eine sogenannte Suchrichtung $d_k \in \mathbb{R}^p$ bestimmt, so dass es gibt ein $\delta > 0$ gibt mit

$$S^{\eta,y}(\vartheta_k + \lambda d_k) < S^{\eta,y}(\vartheta_k)$$

für alle $\lambda \in (0, \delta)$. Notwendig hierzu ist $\operatorname{grad} S^{\eta,y}(\vartheta_k) \cdot d_k \leq 0$, hinreichend und, falls die Hesse-Matrix positiv definit ist, sogar äquivalent ist

$$\operatorname{grad} S^{\eta,y}(\vartheta_k) \cdot d_k < 0. \tag{3.18}$$

Wir nennen eine Richtung $d_k \in \mathbb{R}^p$ mit (3.18) eine **Abstiegsrichtung** in ϑ_k. Zu einer Abstiegsrichtung d_k wird anschließend das Liniensuchproblem

$$\min_{\lambda \in (0,1]} S^{\eta,y}(\vartheta_k + \lambda d_k)$$

zur Bestimmmung von λ_k betrachtet, also ein eindimensionales Minimierungsproblem.[14]

Eine Methode besteht darin, diese Liniensuche mit einem sogenannten exakten Verfahren zu lösen, etwa dem Verfahren der Fibonacci-Suche, dem Verfahren des goldenen Schnitts oder Verfahren basierend auf quadratischer oder kubischer Interpolation, vgl. z.B. Bazaraa u. a. (2006, Chapter 8) oder Gill u. a. (1989, Abschnitt 4.1). Da bei den von uns angestrebten Anwendungen Funktionsauswertungen von η teuer sind, aber diese Verfahren viele Funktionsauswertungen benötigen, ist ein Ausweg, nicht das Liniensuchproblem zu lösen, sondern stattdessen sich mit gewissen Verbesserungen durch den neuen Punkt $\vartheta_k + \lambda d_k$, die in der Regel mit sehr wenigen Funktionsauswertungen erreicht werden, zufrieden zu geben und daher nur dementsprechende Bedingungen an λ_k zu stellen, so dass z.B.

$$\lim_{k \to \infty} \operatorname{grad} S^{\eta,y}(\vartheta_k) = \mathbf{0}_p^{\mathrm{T}}$$

gilt, vgl. Korollar 3.25.

Wir stellen hier die **Wolfe-Bedingungen**[15] für zulässige Werte λ_k vor:[16]

Die Folgen $(\vartheta_k)_{k \in \mathbb{N}_0}$, $(d_k)_{k \in \mathbb{N}_0}$ in $\mathbb{R}^p$ und $(\lambda_k)_{k \in \mathbb{N}_0}$ in $\mathbb{R}_0^+$ mit $\vartheta_{k+1} = \vartheta_k + \lambda_k d_k$ für $k \in \mathbb{N}_0$ erfüllen die Wolfe-Bedingungen, falls es $\alpha \in (0, 1)$ und $\beta \in (\alpha, 1)$ gibt, so dass

$$S^{\eta,y}(\vartheta_k + \lambda_k d_k) \leq S^{\eta,y}(\vartheta_k) + \alpha \lambda_k \operatorname{grad} S^{\eta,y}(\vartheta_k) \cdot d_k \tag{3.19}$$

und

$$\operatorname{grad} S^{\eta,y}(\vartheta_k + \lambda_k d_k) \cdot d_k \geq \beta \operatorname{grad} S^{\eta,y}(\vartheta_k) \cdot d_k \tag{3.20}$$

für alle $k \in \mathbb{N}_0$.

Wir geben eine kurze Motivation der Bedingungen: Es gilt die Taylor-Formel

$$S^{\eta,y}(\vartheta_k + \lambda d_k) = S^{\eta,y}(\vartheta_k) + \lambda \operatorname{grad} S^{\eta,y}(\vartheta_k) \cdot d_k + O(\lambda^2).$$

Für kleines $\lambda > 0$ ist

$$\frac{S^{\eta,y}(\vartheta_k + \lambda d_k) - S^{\eta,y}(\vartheta_k)}{\lambda} \approx \operatorname{grad} S^{\eta,y}(\vartheta_k) \cdot d_k$$

die Verbesserung (negativ, wenn d_k Abstiegsrichtung) relativ zur Schrittlänge λ. Daher sichert die Bedingung (3.19), dass die Verbesserungsrate noch mit dem Faktor α oder besser erhalten bleibt. Es sollen auf diese Weise zu kleine Verbesserungen relativ zur Schrittlänge verhindert werden.

[14] Alternativ betrachtet man auch $\min_{\lambda > 0} S^{\eta,y}(\vartheta_k + \lambda d_k)$.

[15] auch Wolfe-Powell-Bedingungen genannt

[16] Für andere Bedingungen wie Armijos Regel oder den Goldstein-Test verweisen wir auf Nocedal u. Wright (1999, Chapter 3), Bazaraa u. a. (2006, Chapter 8) oder Moré u. Thuente (1994).

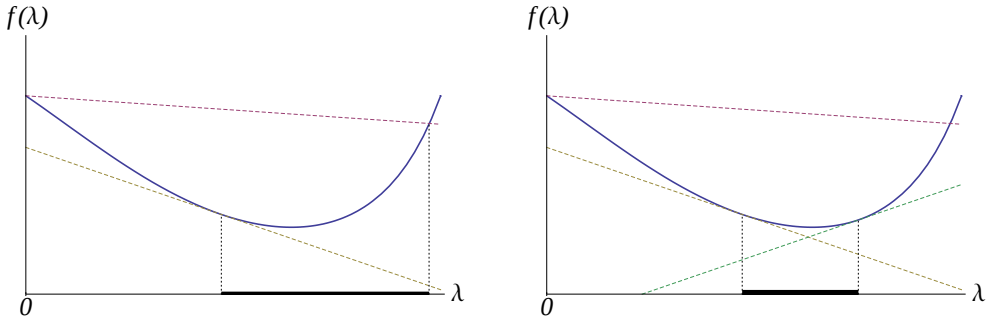


Abbildung 3.2.: Veranschaulichung der Wolfe-Bedingungen (links) bzw. der strengen Wolfe-Bedingungen (rechts): Die beiden markierten Teilbereiche kennzeichnen jeweils die zulässigen λ-Werte.

Die Bedingung (3.20) sorgt dagegen dafür, dass die Schrittlänge nicht zu kurz wird. Diese Bedingung wird auch Krümmungsbedingung genannt, weil die Steigung $f'(\lambda_k)$ von $f : [0, \infty) \to \mathbb{R}$, $f(\lambda) = S^{\eta,y}(\vartheta_k + \lambda d_k)$ im Punkt λ_k größer sein soll als $\beta f'(0)$. Will man zusätzlich sichern, dass die Steigung nicht beliebig groß wird, so kann man die Bedingung (3.20) zu

$$|\text{grad}\, S^{\eta,y}(\vartheta_k + \lambda_k d_k) \cdot d_k| \leq \beta\, |\text{grad}\, S^{\eta,y}(\vartheta_k) \cdot d_k| \tag{3.21}$$

verschärfen. Dann erfüllt $\lambda_k \geq 0$ die **strengen Wolfe-Bedingungen**, falls die beiden Bedingungen (3.19) und (3.21) gelten. Für eine graphische Veranschaulichung zulässiger λ-Werte vergleiche Abbildung 3.2.

Der folgende Satz nach Wolfe (1969, 1971) zeigt, dass die Wolfe-Bedingungen bei gegebenem $\vartheta_k \in \mathbb{R}^p$ und $d_k \in \mathbb{R}^p$ immer durch ein $\lambda_k > 0$ erfüllbar sind, wenn d_k eine Abstiegsrichtung in ϑ_k ist.

Satz 3.22
Sei d_k eine Abstiegsrichtung in ϑ_k für $S^{\eta,y}$ und $\alpha \in (0, 1)$, $\beta \in (\alpha, 1)$. Dann gibt es ein Intervall von Punkten $\lambda > 0$, die die strengen Wolfe-Bedingungen (3.19) *und* (3.21) *erfüllen.*

Beweis: Betrachte die Funktion

$$f : [0, \infty) \to \mathbb{R},\ f(\lambda) = S^{\eta,y}(\vartheta_k + \lambda d_k).$$

Da d_k Abstiegsrichtung, gilt $f'(0) < 0$ und die strengen Wolfe-Bedingungen werden zu

$$f(\lambda) \leq f(0) + \alpha\lambda f'(0) \tag{3.22}$$

und

$$|f'(\lambda)| \leq \beta |f'(0)|. \tag{3.23}$$

Angenommen es gelte

$$f(\lambda) \geq f(0) + \alpha\lambda f'(0) \quad \text{für alle } \lambda > 0.$$

Dann wäre

$$f'(0) = \lim_{\lambda \to 0} \frac{1}{\lambda}(f(\lambda) - f(0)) \geq \alpha f'(0)$$

im Widerspruch zu $f'(0) < 0$ und $\alpha < 1$. Daher gibt es ein $\lambda_L > 0$, so dass

$$f(\lambda_L) < f(0) + \alpha\lambda_L f'(0).$$

Weil damit für die Funktion $g : (0, \infty) \to \mathbb{R}$, $g(\lambda) = f(\lambda) - f(0) - \alpha\lambda f'(0)$ gilt, dass $g(\lambda_L) < 0$ ist, und weil wegen $\alpha > 0$, $f \geq 0$ und $f'(0) < 0$ gilt, dass $\lim_{\lambda \to \infty} g(\lambda) = \infty$, gibt es nach dem Zwischenwertsatz eine Nullstelle von g, d.h. die Funktion f schneidet sich mit der Halbgeraden $H = \{(\lambda, f(0) + \alpha\lambda f'(0)) : \lambda > 0\}$ in einem Punkt $(\lambda, f(\lambda))$.

Angenommen, es gäbe eine Folge $(\lambda_i)_{i \in \mathbb{N}}$ von Nullstellen von g mit $\lambda_i \to 0$ für $i \to \infty$. Dann wäre

$$f'(0) = \lim_{i \to \infty} \frac{f(\lambda_i) - f(0)}{\lambda_i} = \alpha f'(0)$$

im Widerspruch zu $\alpha < 1$ und $f'(0) < 0$. Daher ist $\inf\{\lambda > 0 : g(\lambda) = 0\} > 0$. Da g aber stetig ist, wird das Infimum angenommen, es gibt also eine kleinste Nullstelle λ_H von g. Nach Definition von g ist das der kleinste Schnittpunkt von f mit der Halbgeraden H, d.h. es gilt

$$f(\lambda_H) = f(0) + \alpha\lambda_H f'(0). \tag{3.24}$$

Wegen der Stetigkeit von f und mit der gleichen Argumentation wie oben für λ_L folgt

$$f(\lambda) < f(0) + \alpha\lambda f'(0) \quad \text{für } \lambda \in (0, \lambda_H). \tag{3.25}$$

Aus Stetigkeitsgründen gibt es ein $\lambda_S \in (0, \lambda_H)$ mit

$$f(0) > f(\lambda_S) > f(\lambda_H). \tag{3.26}$$

Da $\lambda_S \in (0, \lambda_H)$ gilt wegen (3.25)

$$f(\lambda_S) < f(0) + \alpha\lambda_S f'(0). \tag{3.27}$$

Nach dem Mittelwertsatz gibt es ein $\bar{\lambda} \in (\lambda_S, \lambda_H)$, so dass

$$f'(\bar{\lambda}) = \frac{f(\lambda_H) - f(\lambda_S)}{\lambda_H - \lambda_S}.$$

Daher folgt wegen $\beta > \alpha$, $f'(0) < 0$, (3.26), (3.27) und (3.24)

$$0 > f'(\bar{\lambda}) > \alpha f'(0) > \beta f'(0). \tag{3.28}$$

Wegen (3.25) und (3.28) erfüllt $\bar{\lambda}$ die strengen Wolfe-Bedingungen (3.22) und (3.23) und es gibt aufgrund der Stetigkeit von f und f' ein Intervall um $\bar{\lambda}$, in dem die Punkte die strengen Wolfe-Bedingungen einhalten. □

Bemerkung 3.23

a) Unter den Bedingungen des Satzes 3.22 gibt es trivialerweise auch ein Intervall von Punkten λ, die die Wolfe-Bedingungen (3.19) und (3.20) erfüllen, da diese Bedingungen schwächer sind.

b) Betrachtet man statt der nichtnegativen Funktion $S^{\eta,y} : \mathbb{R}^p \to \mathbb{R}$ eine nach unten beschränkte, stetig differenzierbare Funktion $G : \mathbb{R}^p \to \mathbb{R}$, so gilt Satz 3.22 offensichtlich immer noch. Falls aber G nicht nach unten beschränkt ist, so gibt es nicht notwendig einen Schnittpunkt wie im Beweis. Dass dann das Lemma nicht mehr gilt, zeigt das Beispiel

$$G : \mathbb{R} \to \mathbb{R},\ G(\vartheta) = -\exp(\vartheta).$$

Für alle $\vartheta \in \mathbb{R}$ ist hier $d = +1$ eine Abstiegsrichtung, aber $|G'(\vartheta + \lambda d)| > |G'(\vartheta)|$, so dass die strengen Wolfe-Bedingungen nicht erfüllbar sind, wie auch immer $\alpha \in (0, 1)$ und $\beta \in (\alpha, 1)$ gewählt werden.

Lemma 3.24
Sei $\gamma \geq 0$ und $S^{\eta,y}$ erfülle die Lipschitz-Bedingung

$$\left\| \operatorname{grad} S^{\eta,y}(\vartheta) - \operatorname{grad} S^{\eta,y}(\tilde{\vartheta}) \right\| \leq \gamma \left\| \vartheta - \tilde{\vartheta} \right\| \tag{3.29}$$

für alle $\vartheta, \tilde{\vartheta} \in \mathbb{R}^p$, d.h. $\operatorname{grad} S^{\eta,y}$ ist Lipschitz-stetig.
Gegeben seien Folgen $(\vartheta_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$, $(d_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$ und $(\lambda_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_0^+$ mit folgenden Eigenschaften: Für jedes $k \in \mathbb{N}_0$ sei entweder

a) $\operatorname{grad} S^{\eta,y}(\vartheta_k) = \mathbf{0}_p^{\mathrm{T}}$, $\vartheta_{k+1} = \vartheta_k$ und $d_k \neq \mathbf{0}_p$ (es gibt keine Abstiegsrichtung in ϑ_k)

oder

b) d_k ist Abstiegsrichtung in ϑ_k, die Wolfe-Bedingungen (3.19) und (3.20) sind erfüllt und $\vartheta_{k+1} = \vartheta_k + \lambda_k d_k$.

Dann gilt

$$\lim_{k\to\infty} \frac{\operatorname{grad} S^{\eta,y}(\vartheta_k) \cdot d_k}{\|d_k\|} = 0.$$

Beweis: Dennis u. Schnabel (1983, Theorem 6.3.3) □

Mit

$$\phi_k = \arccos\left(-\frac{\operatorname{grad} S^{\eta,y}(\vartheta_k) d_k}{\|\operatorname{grad} S^{\eta,y}(\vartheta_k)\| \, \|d_k\|} \right) \tag{3.30}$$

bezeichnen wir den Winkel zwischen d_k und $-\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta_k)$. Ist d_k Abstiegsrichtung in ϑ_k, so ist $\phi_k \leq \pi/2$. Setzt man zudem noch voraus, dass der Winkel nach oben von $\pi/2$ weg beschränkt ist, so erhält man Konvergenz der Gradienten gegen 0.

Korollar 3.25
Gilt unter den Voraussetzungen von Lemma 3.24 zusätzlich, dass

$$\limsup_{k\in\mathbb{N}_0} \phi_k < \pi/2, \tag{3.31}$$

so gilt

$$\lim_{k\to\infty} \operatorname{grad} S^{\eta,y}(\vartheta_k) = 0.$$

Bemerkung 3.26

a) Man kann bei den meisten Algorithmen, die in der Praxis verwendet werden, nicht garantieren, dass die Bedingung (3.31) erfüllt ist. Eine Ausnahme bildet z.B. die Methode des steilsten Abstiegs ($d_k = -\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta_k)$), bei der sogar $\phi_k = 0$ gilt. Die Methode des steilsten Abstiegs zeigt aber nur lineare Konvergenz, die mitunter sehr langsam ist, vgl. Dennis u. Schnabel (1983, S. 115 f.). Dennis u. Schnabel (1983) stellen fest, dass in den meisten Fällen die Methode des steilsten Abstiegs nicht verwendet werden soll, da ein Quasi-Newton-Verfahren viel effizienter ist.

b) Eine Abschwächung von (3.31) ist die Bedingung

$$\sum_{k=0}^{\infty} \cos^2(\phi_k) = \infty. \tag{3.32}$$

Diese liefert unter den Voraussetzungen von Lemma 3.24

$$\liminf_{k\to\infty} \|\operatorname{grad} S^{\eta,y}(\vartheta_k)\| = \mathbf{0}_p^{\mathrm{T}},$$

vgl. Fletcher (1987, S. 31 f.). Die Bedingung (3.32) wird zum Beispiel unter Zusatzbedingungen vom BFGS-Verfahren erfüllt, vgl. Powell (1976). Auch andere Verfahren erfüllen diese Bedingung, vgl. z.B. Al-Baali (1985).

c) Betrachtet man statt der nichtnegativen Funktion $S^{\eta,y} : \mathbb{R}^p \to \mathbb{R}$ eine nach unten beschränkte, stetig differenzierbare Funktion $G : \mathbb{R}^p \to \mathbb{R}$, so gelten Lemma 3.24 und Korollar 3.25 immer noch.

d) Die Lipschitz-Bedingung (3.29) muss nicht global erfüllt sein, damit die Aussagen von Lemma 3.24 und Korollar 3.25 gelten. Genauer genügt, damit die Aussagen für die Folge $(\vartheta_k)_{k\in\mathbb{N}_0}$ gelten, dass die Lipschitz-Bedingung auf der Subniveaumenge $U_0 = \{\vartheta \in \mathbb{R}^p : S^{\eta,y}(\vartheta) \leq S^{\eta,y}(\vartheta_0)\}$ gilt.
Ferner gilt: Falls η stetig differenzierbar auf einer offenen Umgebung von U_0 und J_η Lipschitz-stetig auf U_0 und $\|J_\eta\|$ beschränkt auf U_0 bzw. äquivalent, falls η stetig differenzierbar auf einer offenen Umgebung von U_0 und für $i = 1, \ldots, N$ $\operatorname{grad} \eta_i$ Lipschitz-stetig auf U_0 und $\|\operatorname{grad} \eta_i\|$ beschränkt auf U_0 sind, dann ist auch $S^{\eta,y}$ stetig differenzierbar auf einer offenen Umgebung von U_0 und $\operatorname{grad} S^{\eta,y}$ ist Lipschitz-stetig auf U_0.

Beweis: Es gelte $\|J_\eta\| \leq \beta$ auf U_0 und J_η sei Lipschitz-stetig auf U_0 mit Lipschitz-Konstante $L \geq 0$. Damit ist η Lipschitz-stetig auf U_0 mit Lipschitz-Konstante β und es gilt

$$\begin{aligned}
&\left\|\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta) - \operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\tilde{\vartheta})\right\| \\
&= \left\|2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - y\right) - 2J_\eta^{\mathrm{T}}(\tilde{\vartheta})\Sigma^{-1}\left(\eta(\tilde{\vartheta}) - y\right)\right\| \\
&= \left\|2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - \eta(\tilde{\vartheta})\right) + 2\left(J_\eta^{\mathrm{T}}(\vartheta) - J_\eta^{\mathrm{T}}(\tilde{\vartheta})\right)\Sigma^{-1}\left(\eta(\tilde{\vartheta}) - y\right)\right\| \\
&\leq 2\left\|J_\eta(\vartheta)\right\|\left\|\Sigma^{-1}\right\|\left\|\eta(\vartheta) - \eta(\tilde{\vartheta})\right\| + 2\left\|J_\eta^{\mathrm{T}}(\vartheta) - J_\eta^{\mathrm{T}}(\tilde{\vartheta})\right\|\left\|\Sigma^{-1}\right\|\left\|\eta(\tilde{\vartheta}) - y\right\| \\
&\leq 2\beta\left\|\Sigma^{-1}\right\|\beta\left\|\vartheta - \tilde{\vartheta}\right\| + 2L\left\|\vartheta - \tilde{\vartheta}\right\|\left\|\Sigma^{-1}\right\|\sqrt{S^{\eta,y}(\vartheta_0)} \\
&= 2\left\|\Sigma^{-1}\right\|\left(\beta^2 + L\sqrt{S^{\eta,y}(\vartheta_0)}\right)\left\|\vartheta - \tilde{\vartheta}\right\|
\end{aligned}$$

für alle $\vartheta, \tilde{\vartheta} \in U_0$. □

Eine hinreichende Bedingung hierfür ist, dass U_0 beschränkt ist und η zweimal stetig differenzierbar auf einer offenen Umgebung von U_0 ist.

e) Das Korollar 3.25 folgert die Konvergenz von $(\operatorname{grad} S^{\eta,y}(\vartheta_k))_{k\in\mathbb{N}}$ gegen 0, aber damit ist noch nicht gezeigt, dass $(\vartheta_k)_{k\in\mathbb{N}}$ konvergiert und Konvergenz gegen ein (lokales) Minimum erfolgt. Die Voraussetzungen an $S^{\eta,y}$ sind allerdings so schwach, dass gar kein lokales Minimum zu existieren braucht. Man vergleiche dazu Beispiel 2.17b) mit $y = (-1, 2)^{\mathrm{T}}$. Dann ist

$$S^{\eta,y}(\vartheta) = 6 + \frac{2 - 4\vartheta}{\sqrt{1 + \vartheta^2}},$$

vergleiche Abbildung 3.3. Das globale Infimum ergibt sich aus $\vartheta \to \infty$ und es gibt nur ein lokales Maximum und kein lokales Minimum, da

$$\frac{\partial}{\partial\vartheta} S^{\eta,y}(\vartheta) = \frac{-2(2 + \vartheta)}{(1 + \vartheta^2)^{3/2}}.$$

Ferner gilt aber offensichtlich die Lipschitz-Bedingung (3.29), da die 2. Ableitung

$$\frac{\partial^2}{\partial\vartheta^2} S^{\eta,y}(\vartheta) = \frac{2(-1 + 6\vartheta + 2\vartheta^2)}{(1 + \vartheta^2)^{5/2}}$$

beschränkt ist, vergleiche auch Abbildung 3.3.

Erweitert man die Voraussetzungen noch um die zusätzliche Bedingung, dass die Subniveaumenge $\{\vartheta \in \mathbb{R}^p : S^{\eta,y}(\vartheta) \leq S^{\eta,y}(\vartheta_0)\}$ beschränkt (und damit kompakt) ist, ergibt sich zumindest, dass die Häufungspunkte stationäre Punkte sind, wie das folgende Korollar zeigt.

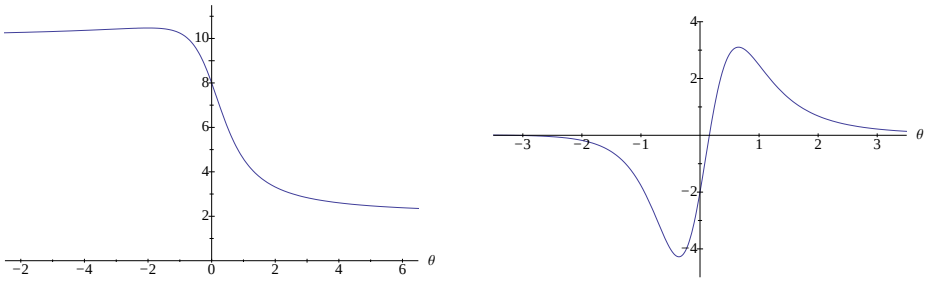


Abbildung 3.3.: Die Funktion $S^{\eta,y}$ (links) und ihre 2.Ableitung $\frac{\partial^2}{\partial\vartheta^2}S^{\eta,y}$ (rechts) aus Bemerkung 3.26e)

Korollar 3.27
Gilt unter den Voraussetzungen von Lemma 3.24 zusätzlich (3.31) *oder* (3.32) *und ist ferner noch* $\{\vartheta \in \mathbb{R}^p : S^{\eta,y}(\vartheta) \leq S^{\eta,y}(\vartheta_0)\}$ *beschränkt, dann ist die Folge* $(\vartheta_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$ *beschränkt und jeder Häufungspunkt der Folge ist ein stationärer Punkt von* $S^{\eta,y}$*.*

Beweis: analog Bertsekas (1999, Proposition 1.2.4) □

Wir wenden uns nun konkreten Algorithmen zu, um bei gegebenem ϑ_k, Abstiegsrichtung d_k und $\alpha \in (0,1)$, $\beta \in (\alpha,1)$ zu einer nach unten beschränkten Funktion G ein durch Satz 3.22 garantiertes $\lambda_k > 0$ zu finden, so dass die Wolfe-Bedingungen

$$G(\vartheta_k + \lambda_k d_k) \leq G(\vartheta_k) + \alpha\lambda_k \operatorname{grad}^{\mathrm{T}} G(\vartheta_k) d_k \qquad \text{und} \tag{3.33}$$

$$\operatorname{grad} G(\vartheta_k + \lambda_k d_k) \cdot d_k \geq \beta \operatorname{grad} G(\vartheta_k) \cdot d_k \qquad \text{oder} \tag{3.34}$$

$$|\operatorname{grad} G(\vartheta_k + \lambda_k d_k) \cdot d_k| \leq \beta\, |\operatorname{grad} G(\vartheta_k) \cdot d_k| \tag{3.35}$$

erfüllt werden. In dieser Klasse von Algorithmen startet man mit dem Versuch $\lambda_k = 1$ und versucht dann in der Regel die Schrittweite solange zu verkürzen („Backtracking“), bis die Wolfe-Bedingungen eingehalten werden.

Algorithmus 3.28 (Generischer Backtracking-Algorithmus)

Input: $G : \mathbb{R}^p \to \mathbb{R}$ stetig differenzierbar,
$\vartheta \in \mathbb{R}^p$,
Abstiegsrichtung $d \in \mathbb{R}^p$ bezüglich G in ϑ,
$\alpha \in (0,1)$,
$0 < l < u < 1$.

Output: $\lambda > 0$, so dass λ, ϑ, d die Bedingung

$$G(\vartheta + \lambda d) \leq G(\vartheta) + \alpha\lambda \operatorname{grad} G(\vartheta) d \tag{3.36}$$

erfüllen.

1. Start mit $\lambda = 1$

2. Falls (3.36) nicht erfüllt, setze $\lambda := \rho\lambda$ für ein $\rho \in [l, u]$ (in jedem Schritt verschiedenes ρ möglich) und wiederhole (2).
   Ansonsten STOP mit Output λ.

Bemerkung 3.29

a) Bei vielen Algorithmen kann man die Konstanten $\alpha \in (0, 1)$ und $\beta \in (\alpha, 1)$ frei wählen. Dennis u. Schnabel (1983, Section 6.3.2) und Nocedal u. Wright (1999, Chapter 3) empfehlen bei der Verwendung des Backtracking-Algorithmus zur Liniensuche bei Newton- oder Quasi-Newton-Verfahren die Wahl von $\alpha = 10^{-4}$ und $\beta = 0.9$.

b) In der Praxis verzichtet man meist auf eine Implementation der 2. Ungleichung aus den Wolfe-Bedingungen, weil man hofft, dass durch die Backtracking-Strategie gesichert ist, dass die Schrittlänge nicht zu klein ist, vgl. die Motivation der 2. Ungleichung auf S. 63. Will man sicher gehen, dass die 2.Ungleichung auch erfüllt ist, dann muss man event. von $\lambda_0 = 1$ ausgehend die λ-Werte erst noch erhöhen. Für eine algorithmische Umsetzung vergleiche Dennis u. Schnabel (1983, A 6.3.1mod).

c) Die konkreten Backtracking-Algorithmen unterscheiden sich in der Strategie der Wahl von ρ. Wir werden einen Backtracking-Algorithmus vorstellen, der auf einer quadratischen Anpassung der aktuellen Funktionsinformationen basiert. Ein ähnlicher Algorithmus wird von Dennis u. Schnabel (1983, A 6.3.1) vorgeschlagen; dort erfolgt im ersten Schritt ebenfalls eine quadratische Anpassung, in den weiteren Schritten allerdings eine kubische Anpassung. Für weitere Backtracking-Algorithmen vergleiche Nocedal u. Wright (1999, Chapter 3) oder Bazaraa u. a. (2006, Chapter 8).

Algorithmus 3.30 (Backtracking-Algorithmus mit iterierter quadratischer Anpassung)

Input: $G : \mathbb{R}^p \to \mathbb{R}$ stetig differenzierbar,
$\vartheta \in \mathbb{R}^p$,
Abstiegsrichtung $d \in \mathbb{R}^p$ bezüglich G in ϑ,
$\alpha = 10^{-4}$,
$l = 0.1$.

Output: $\lambda > 0$, so dass λ, ϑ, d

$$G(\vartheta + \lambda d) \leq G(\vartheta) + \alpha\lambda \operatorname{grad} G(\vartheta) d$$

erfüllen.

1. Setze $\lambda^{(0)} = 1$, $k = 0$.

2. Falls
$$G(\vartheta + \lambda^{(k)} d) \leq G(\vartheta) + \alpha\lambda^{(k)} \operatorname{grad} G(\vartheta) d$$
STOP mit Output $\lambda = \lambda^{(k)}$, ansonsten setze
$$\lambda^{(k+1)} := \lambda^{(k)} \max\left(l,\ -\frac{\operatorname{grad} G(\vartheta) d}{2(G(\vartheta + \lambda^{(k)} d) - G(\vartheta) - \operatorname{grad} G(\vartheta) d)}\right) \tag{3.37}$$

und $k := k + 1$ und wiederhole den Schritt.

Bemerkung 3.31
Wir geben eine Interpretation der Updateformel (3.37). Es gelte also

$$G(\vartheta + \lambda^{(k)} d) > G(\vartheta) + \alpha \lambda^{(k)} \operatorname{grad} G(\vartheta) d. \tag{3.38}$$

Wir bestimmen dann eine quadratische Funktion $q : \mathbb{R} \to \mathbb{R}$ mit $q(0) = G(\vartheta)$, $q'(0) = \operatorname{grad} G(\vartheta) d < 0$ und $q(1) = G(\vartheta + \lambda^{(k)} d)$, d.h.

$$q(\lambda) = (q(1) - q(0) - q'(0))\lambda^2 + q'(0)\lambda + q(0).$$

Wegen (3.38) gilt $q(1) > q(0) + \alpha q'(0)$ und damit wegen $q'(0) < 0$ und $\alpha < 1$ insbesondere $q(1) - q(0) - q'(0) > 0$. Die quadratische Funktion q hat folglich ein eindeutiges Minimum in

$$\tilde{\lambda} = -\frac{q'(0)}{2(q(1) - q(0) - q'(0))} > 0.$$

Wegen (3.38) und $\alpha > 0$ gilt $\tilde{\lambda} < \frac{1}{2(1-\alpha)} < 0.5$. Man kann also Algorithmus 3.30 als spezielle Form des generischen Algorithmus 3.28 mit Obergrenze $u = \frac{1}{2(1-\alpha)}$ (oder $u = 1/2$) ansehen.

Wir wenden uns nun verschiedenen Verfahren zur Ermittlung von Suchrichtungen d_k zu.

3.2.1. Newton-Verfahren

Das Newton-Verfahren basiert auf der Idee, ausgehend von einem Startpunkt ϑ_0 im $(k+1)$-ten Schritt die zu minimierende Funktion G im Punkt ϑ_k (lokal) durch eine quadratische Funktion zu approximieren und deren Minimumstelle als ϑ_{k+1} zu wählen, bis ein geeignetes Konvergenzkriterium erfüllt ist.

Ist $G : \mathbb{R}^p \to \mathbb{R}$ zweimal stetig differenzierbar, so verwenden wir als quadratische Approximation in ϑ_k die Funktion

$$q_k : \mathbb{R}^p \to \mathbb{R},\ q_k(\vartheta) = G(\vartheta_k) + \operatorname{grad} G(\vartheta_k) \cdot (\vartheta - \vartheta_k) + \frac{1}{2}(\vartheta - \vartheta_k)^{\mathrm{T}} \cdot \operatorname{Hess} G(\vartheta_k) \cdot (\vartheta - \vartheta_k).$$

Nach dem Satz von Taylor gilt

$$G(\vartheta) = q_k(\vartheta) + o(\|\vartheta - \vartheta_k\|^2).$$

Für die quadratische Approximation q_k gilt

$$\begin{aligned} \operatorname{grad} q_k(\vartheta) &= \operatorname{grad} G(\vartheta_k) + (\vartheta - \vartheta_k)^{\mathrm{T}} \operatorname{Hess} G(\vartheta_k), \\ \operatorname{Hess} q_k(\vartheta) &= \operatorname{Hess} G(\vartheta_k). \end{aligned} \tag{3.39}$$

Notwendige Bedingung für den Minimumspunkt ϑ_{k+1} von q_k ist $\operatorname{grad} q_k(\vartheta_{k+1}) = \mathbf{0}_p^{\mathrm{T}}$; wenn $\operatorname{Hess} G(\vartheta_k)$ positiv definit ist, dann ist die Bedingung auch hinreichend. Auflösen von $\operatorname{grad} q_k(\vartheta_{k+1}) = \mathbf{0}_p^{\mathrm{T}}$ liefert aus (3.39) unter der Annahme, dass $\operatorname{Hess} G(\vartheta_k)$ regulär ist,

$$\vartheta_{k+1} = \vartheta_k - (\operatorname{Hess} G(\vartheta_k))^{-1} \operatorname{grad}^{\mathrm{T}} G(\vartheta_k). \tag{3.40}$$

Alternativ formuliert heißt dies, dass $\vartheta_{k+1} = \vartheta_k + d_k$, wobei d_k das lineare Gleichungssystem $\operatorname{Hess} G(\vartheta_k) d_k = -\operatorname{grad}^{\mathrm{T}} G(\vartheta_k)$ erfüllt.

Algorithmus 3.32 (Newton-Verfahren)
Input: $G : \mathbb{R}^p \to \mathbb{R}$ zweimal stetig differenzierbar,
Startpunkt $\vartheta_0 \in \mathbb{R}^p$,
Abbruchkriterium
Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation einer Minimumstelle von G)

1. Setze $k = 0$.
2. Löse das lineare Gleichungssystem
$$\operatorname{Hess} G(\vartheta_k) d_k = -\operatorname{grad}^{\mathrm{T}} G(\vartheta_k) \tag{3.41}$$
für d_k und setze $\vartheta_{k+1} = \vartheta_k + d_k$.
3. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten setze $k := k + 1$ und gehe zu Schritt 2.

Satz 3.33
Sei $G : \mathbb{R}^p \to \mathbb{R}$ zweimal stetig differenzierbar, $\vartheta^* \in \mathbb{R}^p$ mit $\operatorname{grad} G(\vartheta^*) = \mathbf{0}_p^{\mathrm{T}}$ und $\operatorname{Hess} G(\vartheta^*)$ regulär. Dann gibt es eine Umgebung U von ϑ^*, so dass gilt: Ist $\vartheta_0 \in U$, dann ist die nach (3.40) erzeugte Folge $(\vartheta_k)_{k\in\mathbb{N}}$ wohldefiniert und es gilt $\lim_{k\to\infty} \vartheta_k = \vartheta^*$.
Ist G sogar dreimal stetig differenzierbar,[17] *dann ist die Konvergenz lokal quadratisch, d.h. es gibt ein $c \in [0, 1)$ und eine Umgebung V von ϑ^*, so dass für alle $\vartheta_0 \in V$ gilt*[18]

$$\limsup_{k\to\infty} \frac{\|\vartheta_{k+1} - \vartheta^*\|}{\|\vartheta_k - \vartheta^*\|^2} = c.$$

Beweis: Hämmerlin u. Hoffmann (1991) oder Nocedal u. Wright (1999) □

Wir untersuchen an zwei Beispielen einige Probleme des Newton-Verfahrens.

[17] Es genügt, dass $\operatorname{Hess} G$ in einer Umgebung von ϑ^* Lipschitz-stetig ist.
[18] Hierbei ist wie im Folgenden der Fall, dass schon nach endlich vielen Schritten Konvergenz eintritt, wie in Definition 3.54 gesondert zu behandeln, indem man $0/0 := 0$ setzt.

Beispiel 3.34
Wir betrachten die Modellfunktion

$$\eta : \mathbb{R} \to \mathbb{R}^2,\ \eta(\vartheta) = (\vartheta, \vartheta^2)^{\mathrm{T}},$$

mit $\Sigma = I_2$, vgl. S. 27. Dann ist

$$S^{\eta,y} : \mathbb{R} \to \mathbb{R},\ S^{\eta,y}(\vartheta) = \vartheta^4 + (1 - 2y_2)\vartheta^2 - 2y_1\vartheta + y_1^2 + y_2^2.$$

a) Wir betrachten zum einen den Fall $y^{\mathrm{T}} = (-1, 2)$. Man erhält

$$S^{\eta,y} : \mathbb{R} \to \mathbb{R},\ S^{\eta,y}(\vartheta) = \vartheta^4 - 3\vartheta^2 + 2\vartheta + 5,$$

vgl. Abbildung 3.4. Die Funktion $S^{\eta,y}$ besitzt eine globale Minimumstelle in $\vartheta = (-1 - \sqrt{3})/2 \approx -1.366$, eine lokale Maximumsstelle in $\vartheta = (-1 + \sqrt{3})/2 \approx 0.366$ und eine lokale Minimumstelle in $\vartheta = 1$. Betrachtet man nun in Abhängigkeit vom Startwert ϑ_0 die entsprechenden Limespunkte der Newton-Iteration (3.40), so ergibt eine Berechnung bei einer Genauigkeit von 100 Stellen das Ergebnis in Abbildung 3.4. Nur für $\vartheta_0 < -\sqrt{2}/2$ konvergiert das Newton-Verfahren gegen die globale Minimumstelle, für Startwerte $\vartheta_0 \in (-\sqrt{2}/2, \sqrt{2}/2)$ erfolgt meist Konvergenz gegen die lokale Maximumsstelle und für Startwerte $\vartheta_0 > \sqrt{2}/2$ erfolgt Konvergenz gegen die lokale Minimumstelle.[19]

b) Wir betrachten nun den Fall $y^{\mathrm{T}} = (-4, 2.5)$, d.h.

$$S^{\eta,y} : \mathbb{R} \to \mathbb{R},\ S^{\eta,y}(\vartheta) = \vartheta^4 - 4\vartheta^2 + 8\vartheta + 89/4,$$

vgl. Abbildung 3.5. Hier liegt die globale Minimumstelle in

$$\vartheta^* = -\frac{(9 - \sqrt{57})^{1/3}}{3^{2/3}} - \frac{2}{(3(9 - \sqrt{57}))^{1/3}} \approx -1.769,$$

es gibt keine weiteren lokalen Minimumstellen oder Maximumsstellen. Zwar konvergiert das Newton-Verfahren für alle Startwerte $\vartheta_0 < -\sqrt{2/3}$ gegen die globale Minimumstelle, aber es gibt Startwerte $\vartheta_0 > -\sqrt{2/3}$, für die Divergenz auftritt; bei solchen Startwerten konvergieren die ungeraden Folgenglieder gegen 0 und die geraden Folgenglieder gegen 1 oder umgekehrt. Es gibt sogar eine offene Umgebung um 0 mit solchen Startwerten bei denen zugehörigen Folgen oszillieren. Als „Limespunkt“ wurde in Abbildung 3.5 für solche Punkte zur besseren Darstellung stets die 1 gewählt. Weitergehende Ausführungen zum chaotischen Verhalten in diesem Beispiel finden sich in Hubbard u. a. (2001).

Beispiel 3.35
Als zweites Beispiel betrachten wir die Minimierung der Funktion

$$G : \mathbb{R} \to \mathbb{R},\ G(\vartheta) = \ln(1 + \vartheta^2),$$

[19] Für $\vartheta_0 = -\sqrt{2}/2$ und $\vartheta_0 = +\sqrt{2}/2$ ist die Folge nicht definiert, da hierfür $\frac{\partial^2}{\partial\vartheta^2} S^{\eta,y}(\vartheta_0) = 0$.

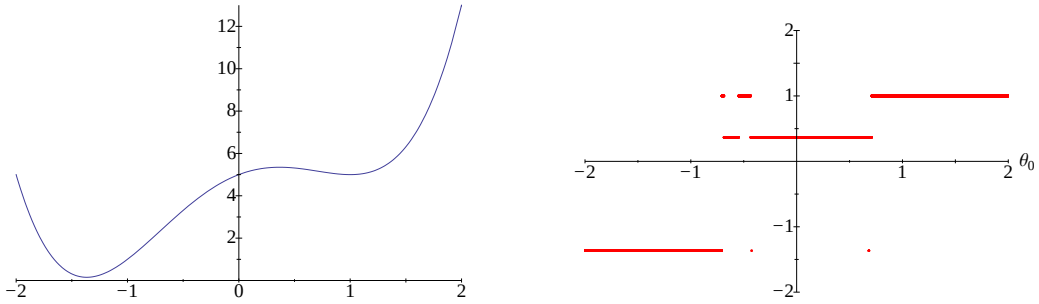


Abbildung 3.4.: Die Funktion $S^{\eta,y}$ für $y^{\mathrm{T}} = (-1, 2)$ (links) und die zugehörigen Limespunkte der Newton-Iteration in Abhängigkeit von den Startpunkten (rechts) aus Beispiel 3.34a)

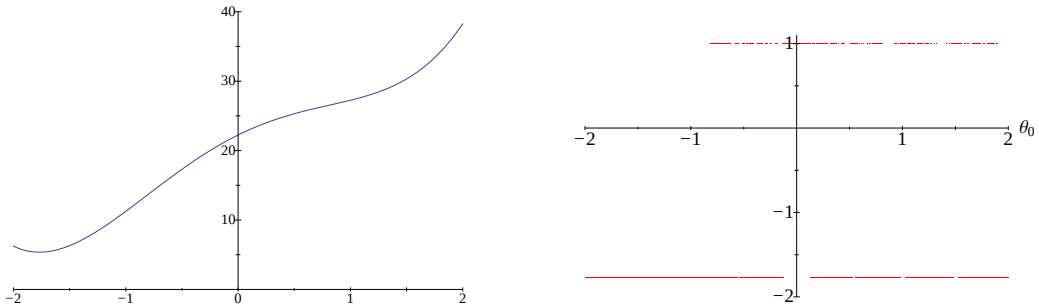


Abbildung 3.5.: Die Funktion $S^{\eta,y}$ für $y^{\mathrm{T}} = (-4, 2.5)$ (links) und die zugehörigen „Limespunkte“ der Newton-Iteration in Abhängigkeit von den Startpunkten (rechts) aus Beispiel 3.34b)

vgl. Abbildung 3.6. Betrachtet man die zur Newton-Iteration (3.40) gehörige Abbildung

$$\mathcal{G} : \mathbb{R} \to \mathbb{R},\ \mathcal{G}(\vartheta) = \vartheta - G'(\vartheta)/G''(\vartheta) = \frac{-2\vartheta^3}{1-\vartheta^2},$$

vgl. Abbildung 3.6, so erkennt man, dass für $\vartheta_0 < -1$ und $\vartheta_0 > 1$ die Absolutbeträge der Folgenglieder der Newton-Iteration immer größer werden. Man bewegt sich zwar in die richtige Richtung, schießt aber weit über das Ziel hinaus („Overshooting“). Genau für $|\vartheta_0| < 1/\sqrt{3}$ tritt dagegen Konvergenz gegen die globale Minimumstelle in $\vartheta = 0$ ein (und für $\vartheta = -1$ und $\vartheta = 1$ ist die Newton-Iteration nicht wohldefiniert).

Globale Konvergenz ist also nicht gesichert, die vom Newton-Verfahren erzeugte Folge kann gegen ∞ divergieren, zwischen zwei Häufungspunkten alternieren oder gegen eine lokale Minimumstelle oder gar eine lokale Maximumsstelle konvergieren.

Das „reine“ Newton-Verfahren kann man dadurch modifizieren, dass man es mit einer Liniensuche kombiniert. Zu beachten ist hierbei, dass d_k Abstiegsrichtung in ϑ_k ist, wenn $\operatorname{Hess} G(\vartheta_k)$ positiv definit ist. Falls keine Abstiegsrichtung berechnet worden ist, dann wählt man stattdessen die Richtung des steilsten Abstiegs.[20] Man hofft so auch bessere globale Konvergenzeigenschaften zu erreichen.

Algorithmus 3.36 (Modifiziertes Newton-Verfahren mit Liniensuche)

Input: $G : \mathbb{R}^p \to \mathbb{R}$ zweimal stetig differenzierbar,
Startpunkt $\vartheta_0 \in \mathbb{R}^p$,
Abbruchkriterium

Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation einer Minimumstelle von G)

1. Setze $k = 0$.
2. Löse das lineare Gleichungssystem
$$\operatorname{Hess} G(\vartheta_k) d_k = -\operatorname{grad}^{\mathrm{T}} G(\vartheta_k)$$
für d_k.
3. Falls das Gleichungssystem keine Lösung hat (weil $\operatorname{Hess} G(\vartheta_k)$ singulär ist) oder $\operatorname{grad} G(\vartheta_k) d_k \geq 0$ gilt (d.h. d_k ist keine Abstiegsrichtung), so setze stattdessen $d_k = -\operatorname{grad}^{\mathrm{T}} G(\vartheta_k)$ (Richtung des steilsten Abstiegs).
4. Bestimme $\lambda_k > 0$ durch eine (eventuell inexakte) Liniensuche in ϑ_k mit Suchrichtung d_k und setze $\vartheta_{k+1} = \vartheta_k + \lambda_k d_k$.
5. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten setze $k := k + 1$ und gehe zu Schritt 2.

[20] Statt dieser zweiten Modifikation gibt es viele andere Strategien, z.B. die Verwendung von $E_k + \operatorname{Hess} G(\vartheta_k)$ an Stelle von $\operatorname{Hess} G(\vartheta_k)$ mit einer geeigneten Matrix E_k (z.B. $E_k = \tau_k I_p$ für ein geeignetes $\tau_k > 0$), so dass die entstehende Matrix positiv definit ist. Vergleiche hierzu Nocedal u. Wright (1999, Chapter 6).

Wir betrachten die Eigenschaften dieses modifizierten Newton-Verfahrens in den obigen Beispielen und interessieren uns besonders dafür, ob sich die globalen Konvergenzeigenschaften verbessert haben.

Beispiel 3.37
Im Beispiel 3.34a) konvergiert ein modifiziertes Newton-Verfahren mit Liniensuche für alle Startwerte $\vartheta_0 < (-1+\sqrt{3})/2 \approx 0.366$ gegen die globale Minimumstelle, für alle Startwerte $\vartheta_0 > (-1+\sqrt{3})/2 \approx 0.366$ hingegen gegen die lokale Minimumstelle $\vartheta = 1$, vergleiche Abbildung 3.7 links. In den Beispielen 3.34b) konvergieren alle Startwerte gegen die Minimumstelle, vgl. Abbildung 3.7 rechts. Ebenso konvergieren dann alle Startwerte im Beispiel 3.35 gegen die Minimumstelle.

In den Beispielen zeigen sich also deutlich bessere globale Konvergenzeigenschaften als beim klassischen Newton-Verfahren, die Konvergenz gegen eine lokale Minimumstelle statt der globalen Minimumstelle kann so aber nicht verhindert werden. Ein übliches Verfahren, um sich hiergegen besser abzusichern, ist, das Verfahren mit mehreren Startpunkten durchzuführen.

Satz 3.38
Sei G zweimal stetig differenzierbar auf einer offenen konvexen Menge D und $\operatorname{Hess} G$ *Lipschitz-stetig auf D, d.h. es exist. $\gamma > 0$, so dass für alle $\vartheta, \tilde{\vartheta} \in D$ gilt*[21]

$$\left\| \operatorname{Hess} G(\vartheta) - \operatorname{Hess} G(\tilde{\vartheta}) \right\| \leq \gamma \left\| \vartheta - \tilde{\vartheta} \right\|.$$

Zudem gelten für die Folgen $(\vartheta_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$, $(d_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$ und $(\lambda_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_0^+$ die folgenden Eigenschaften:
d_k ist Abstiegsrichtung in ϑ_k, die Wolfe-Bedingungen (3.19) *und* (3.20) *(mit G anstelle von $S^{\eta,y}$) sind für ein $\alpha < 1/2$ erfüllt und $\vartheta_{k+1} = \vartheta_k + \lambda_k d_k$. Ferner konvergiere $(\vartheta_k)_{k\in\mathbb{N}_0}$ gegen ein $\vartheta^* \in D$, wobei* $\operatorname{Hess} G(\vartheta^*)$ *positiv definit sei, und es gelte*

$$\lim_{k\to\infty} \frac{\left\| \operatorname{grad}^{\mathrm{T}} G(\vartheta_k) + \operatorname{Hess} G(\vartheta_k) d_k \right\|}{\|d_k\|} = 0.$$

Dann gilt $\operatorname{grad} G(\vartheta^*) = \mathbf{0}_p$ *und es gibt ein $k_0 \in \mathbb{N}_0$, so dass für $k \geq k_0$ die Wolfe-Bedingungen auch von $\tilde{\lambda}_k = 1$ erfüllt werden.*
Falls es zudem ein $s \in \mathbb{N}_0$ gibt, so dass $\lambda_k = 1$ für $k \geq s$, dann konvergiert $(\vartheta_k)_{k\in\mathbb{N}_0}$ superlinear gegen ϑ^*, d.h.

$$\limsup_{k\to\infty} \frac{\|\vartheta_{k+1} - \vartheta^*\|}{\|\vartheta_k - \vartheta^*\|} = 0.$$

[21] Für $A = (a_{ij})_{1\leq i,j\leq p} \in \mathbb{R}^{p\times p}$ bezeichne hierbei $\|A\|$ irgendeine Matrixnorm auf $\mathbb{R}^{p\times p}$, also nicht unbedingt die Spektralnorm, sondern z.B. auch die Frobenius-Norm $\|A\|_{\mathrm{Frob}} = \left(\sum_{i=1}^p \sum_{j=1}^p a_{ij}^2\right)^{1/2}$; die Wahl der Matrixnorm wirkt sich nur auf die möglichen Werte der Konstanten γ aus, nicht jedoch auf die Existenz einer solchen Konstante.

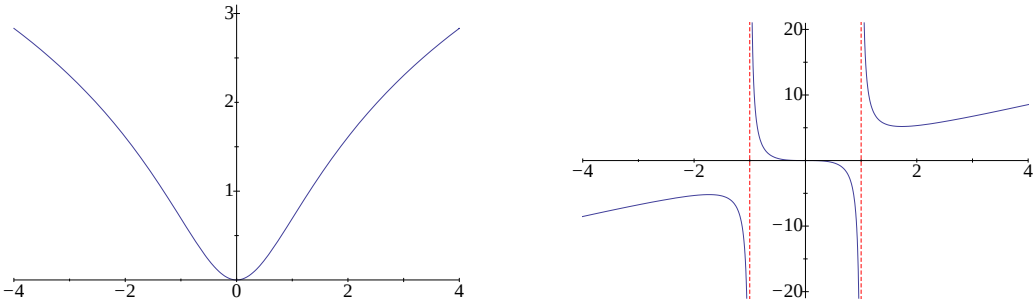


Abbildung 3.6.: Die Funktion G (links) und die Newton-Abbildung $\mathcal{G}$ (rechts) aus Beispiel 3.35

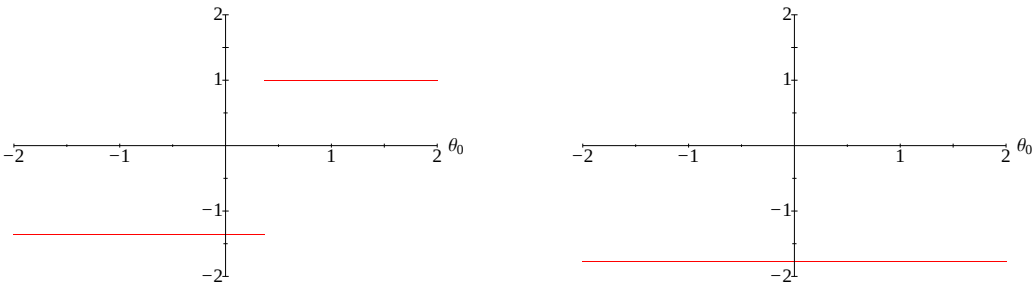


Abbildung 3.7.: Die „Limespunkte“ der modifizierten Newton-Iteration in Abhängigkeit von den Startpunkten für Beispiel 3.34a) (links) und Beispiel 3.34b) (rechts)

Beweis: Dennis u. Schnabel (1983, Theorem 6.3.4) □

Bemerkung 3.39
Wir betrachten nun ein modifizierten Newton-Verfahren mit Liniensuche, bei dem $\lambda_k = 1$ verwendet wird, falls die Wolfe-Bedingungen erfüllt sind. Dies ist z.B. der Fall, wenn ein Backtracking-Verfahren zur Liniensuche zum Einsatz kommt. Weil beim Newton-Verfahren $\operatorname{grad}^{\mathrm{T}} G(\vartheta_k) + \operatorname{Hess} G(\vartheta_k) d_k = 0$ gilt, stellt der Satz 3.38 inbesondere sicher, dass die guten lokalen Konvergenzeigenschaften des Newton-Verfahrens aus Satz 3.33 beim einem solchen modifizierten Newton-Verfahren mit Liniensuche nicht verloren gehen. Ist ferner $\operatorname{Hess} G(\vartheta^*)$ positiv definit, dann ist $\operatorname{Hess} G$ auch in einer Umgebung von ϑ^* positiv definit, also stimmen dann lokal das Newton-Verfahren und das modifizierte Newton-Verfahren mit Liniensuche überein. Bei dreimaliger stetiger Differenzierbarkeit ergibt sich lokal nicht nur superlineare Konvergenz, sondern nach Satz 3.33 sogar quadratische Konvergenz. Für eine weitere Anwendung des Satzes 3.38 vergleiche auch Satz 3.71.

Das Newton-Verfahren (oder das modifizierte Newton-Verfahren mit Liniensuche) kann man benutzen, um wie in den Beispielen 3.34a) und b) speziell $S^{\eta,y}$ zu minimieren, d.h. um den Kleinste-Quadrate-Schätzer zu bestimmen. Zu beachten ist hierbei, dass man zur Lösung des linearen Gleichungssystems nicht wie im Abschnitt 3.1 die dortige spezielle Struktur der Koeffizientenmatrix (hier hingegen $\operatorname{Hess} S^{\eta,y}$, vgl. (3.17)) beim Verfahren basierend auf der QR-Zerlegung bzw. der Singulärwertzerlegung ausnutzen kann und auf „herkömmliche“ Art lösen muss. Speziell für das Kleinste-Quadrate-Problem werden wir aber im folgenden Abschnitt ein Verfahren betrachten, bei dem die Hesse-Matrix von $S^{\eta,y}$ geeignet approximiert wird, so dass die Koeffizientenmatrix wieder eine spezielle Struktur hat – siehe (3.43) – und die Verfahren aus 3.1 ohne Einschränkungen anwendbar sind.[22]

3.2.2. Gauß-Newton-Verfahren

Die Hesse-Matrix von $S^{\eta,y}$ ergibt sich (vgl. auch Formel (3.17)) zu

$$\operatorname{Hess} S^{\eta,y}(\vartheta) = 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + 2\sum_{j=1}^{N}\left(\Sigma^{-1}\left(\eta(\vartheta)-y\right)\right)_j \operatorname{Hess}\eta_j(\vartheta). \tag{3.42}$$

Sind die Residuen $y-\eta(\vartheta)$ klein oder ist η annähernd linear, dann sind für $j = 1, \ldots, p$ die Werte $\|\operatorname{Hess}\eta_j(\vartheta)\| \,|\, \left(\Sigma^{-1}\left(\eta(\vartheta)-y\right)\right)_j \,|$ klein. Sind sie deutlich kleiner als die Eigenwerte von $J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)$, dann kann man hoffen,[23] dass der zweite Summand auf der rechten

[22] Wir schränken also die Inputklasse ein und betrachten bei der Minimierung statt allgemeiner zweimal stetig differenzierbarer Funktionen nur noch die Funktionen $S^{\eta,y}$.

[23] Insbesondere könnte die Approximation also schlecht sein, wenn das Modell nicht regulär ist in ϑ, also wenn $\operatorname{rg}(J_\eta^{\mathrm{T}}(\vartheta)) < p$ gilt.

Seite von (3.42), also $2H_\eta(\vartheta)$ mit

$$H_\eta(\vartheta) = \sum_{j=1}^{N} \left(\Sigma^{-1}\left(\eta(\vartheta) - y\right)\right)_j \operatorname{Hess} \eta_j(\vartheta),$$

vernachlässigbar ist und stattdessen die Approximation

$$\operatorname{Hess} S^{\eta,y}(\vartheta) \approx 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) \tag{3.43}$$

verwendet werden kann. Dies erspart die im Allgemeinen, und vor allem bei den von uns später betrachteten Anwendungen, aufwendigen Berechnungen der Hesse-Matrizen von η_j für $j = 1, \ldots, N$.

Algorithmus 3.40 (Gauß-Newton-Verfahren)
Input: $\eta : \mathbb{R}^p \to \mathbb{R}^N$ stetig differenzierbar,[24]
Beobachtung $y \in \mathbb{R}^N$,
Startpunkt $\vartheta_0 \in \mathbb{R}^p$,
Abbruchkriterium
Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation einer Kleinste-Quadrate-Schätzung)

1. Setze $k = 0$.
2. Löse das lineare Gleichungssystem

$$J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k)d_k = -J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(\eta(\vartheta_k) - y\right) \tag{3.44}$$

   für d_k und setze $\vartheta_{k+1} = \vartheta_k + d_k$.
3. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten setze $k := k + 1$ und gehe zu Schritt 2.

Kombiniert man das so abgewandelte Newton-Verfahren wieder mit einer Liniensuche, so erhält man

Algorithmus 3.41 (Gauß-Newton-Verfahren mit Liniensuche)
Input: $\eta : \mathbb{R}^p \to \mathbb{R}^N$ stetig differenzierbar,[25]
Beobachtung $y \in \mathbb{R}^N$,
Startpunkt $\vartheta_0 \in \mathbb{R}^p$,
Abbruchkriterium
Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation einer Kleinste-Quadrate-Schätzung)

1. Setze $k = 0$.

[24] $\operatorname{rg}(J_\eta(\vartheta)) = p$ sollte für (fast alle) $\vartheta \in \mathbb{R}^p$ gegeben sein, vgl. Bemerkung 3.43

[25] $\operatorname{rg}(J_\eta(\vartheta)) = p$ sollte für (fast alle) $\vartheta \in \mathbb{R}^p$ aus der Subniveaumenge $\{\vartheta \in \mathbb{R}^p : S^{\eta,y}(\vartheta) \leq S^{\eta,y}(\vartheta_0)\}$ gegeben sein; wie man Schritt 2 abändern kann, im Fall dass doch $\operatorname{rg}(J_\eta(\vartheta_k)) < p$ gilt, vgl. Bemerkung 3.43

2. Löse das lineare Gleichungssystem

$$J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k)d_k = -J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(\eta(\vartheta_k)-y\right) \tag{3.45}$$

   für d_k.

3. Bestimme λ_k durch eine (eventuell inexakte) Liniensuche in ϑ_k mit Suchrichtung d_k und setze $\vartheta_{k+1} = \vartheta_k + \lambda_k d_k$.

4. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten setze $k := k+1$ und gehe zu Schritt 2.

Bemerkung 3.42
Wir geben im Folgenden eine weitere aufschlussreiche Motivation für das Gauß-Newton-Verfahren.
Betrachtet man ausgehend von einem Startwert ϑ_0 die Folge $(\vartheta_k)_{k\in\mathbb{N}_0}$, die sich dadurch ergibt, dass

$$\vartheta_{k+1} = \operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^p} \|y - \mathfrak{l}_k(\vartheta)\|_\Sigma^2\,, \tag{3.46}$$

wobei

$$\mathfrak{l}_k : \mathbb{R}^p \to \mathbb{R}^N,\ \mathfrak{l}_k(\vartheta) = J_\eta(\vartheta_k)\cdot(\vartheta-\vartheta_k) + \eta(\vartheta_k)$$

die Linearisierung der Modellfunktion η im Punkt ϑ_k ist, dann erhält man wieder das Gauß-Newton-Verfahren: Denn die Kleinste-Quadrate-Probleme (3.46) werden durch die Normalgleichungen

$$J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k)(\vartheta_{k+1}-\vartheta_k) = J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(y-\eta(\vartheta_k)\right) \tag{3.47}$$

für ϑ_{k+1} gelöst, wie man durch die Ersetzungen $X = J_\eta^{\mathrm{T}}(\vartheta_k)$, $y_0 = \eta(\vartheta_k)$, $\vartheta_0 = \vartheta_k$ und $\vartheta^* = \vartheta_{k+1}$ in (2.1), (2.4) und (2.5) erkennt. Die Normalgleichungen (3.47) entsprechen dem Gleichungssystem im Schritt 2 von Algorithmus 3.40. Die Gleichungssysteme (3.44) bzw. (3.45) können daher ohne Einschränkungen mit den Verfahren aus Abschnitt 3.1 gelöst werden, d.h. insbesondere muss man nicht die Matrix $J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k)$ berechnen, wenn man das Verfahren basierend auf der QR-Zerlegung bzw. basierend auf der Singulärwertzerlegung verwendet.

Bemerkung 3.43
Es gilt

$$\begin{aligned}\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta_k)\cdot d_k &= 2\left(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(\eta(\vartheta_k)-y\right)\right)^{\mathrm{T}}\cdot d_k \\ &\overset{(3.45)}{=} -2\left(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k)d_k\right)^{\mathrm{T}} d_k = -2\,\|J_\eta(\vartheta_k)d_k\|_\Sigma^2\,.\end{aligned}$$

Wenn $\operatorname{rg}(J_\eta(\vartheta_k)) = p$, dann gilt $J_\eta(\vartheta_k)d_k = 0$ nach (3.16) und (3.45) genau dann, wenn $\operatorname{grad} S^{\eta,y}(\vartheta_k) = \mathbf{0}_p^{\mathrm{T}}$. Daher gilt im Fall von $\operatorname{rg}(J_\eta(\vartheta_k)) = p$: Wenn ϑ_k kein stationärer Punkt von $S^{\eta,y}$, dann ist $\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta_k)\cdot d_k < 0$, d.h. d_k ist Abstiegsrichtung in ϑ_k.

Falls $\operatorname{rg}(J_\eta(\vartheta)) = p$ für alle $\vartheta \in \Theta$, dann ist das Gleichungssystem eindeutig lösbar und man kann auf eine Prüfung im Algorithmus, ob jeweils eine Abstiegsrichtung vorliegt, verzichten. Falls dies nicht gegeben ist, dann sind verschiedene Strategien möglich. Man kann zuerst die Minimum-Norm-Lösung

$$d_k = -(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k))^+ J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(\eta(\vartheta_k) - y\right) \tag{3.48}$$

des Gleichungssystems (3.45) z.B. mit dem Verfahren basierend auf der Singulärwertzerlegung berechnen und hoffen, dass diese eine Abstiegsrichtung ist. Falls nicht, kann man analog zum modifizierten Newton-Verfahren mit Liniensuche (Algorithums 3.36) gegebenenfalls die Suchrichtung d_k durch die Richtung des steilsten Abstiegs ersetzen. Falls die Minimum-Norm-Lösung d_k des Gleichungssystem (3.45) keine Abstiegsrichtung ist, besteht eine weitere Methode darin – ähnlich wie in Fußnote 20 – statt dem Gleichungssystem (3.45) das Gleichungssystem

$$(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k) + \tau_k I_p)d_k = -J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(\eta(\vartheta_k) - y\right) \tag{3.49}$$

mit einem geeignetem $\tau_k > 0$ zu betrachten. Im Levenberg-Marquardt-Verfahren wird dieses alternative Gleichungssystem für jedes $k \in \mathbb{N}_0$ betrachtet. Man kann dieses Verfahren als Mittelung zwischem dem Verfahren des steilsten Abstiegs, also

$$d_k = -J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(\eta(\vartheta_k) - y\right),$$

und dem Gauß-Newton-Verfahren interpretieren. Für großes τ_k ergibt sich approximativ die Richtung des steilsten Abstiegs, für kleines $\tau_k > 0$ hingegen im Fall $\operatorname{rg}(J_\eta(\vartheta_k)) = p$ approximativ die Richtung aus dem Gauß-Newton-Verfahren. Mit $d_k = \vartheta^* - \vartheta_k$ kann man das Gleichungssystems (3.49) mit den Bezeichnungen der vorigen Bemerkung auch äquivalent durch das Kleinste-Quadrate-Problem

$$\vartheta^* = \operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^p} \|y - \mathfrak{l}_k(\vartheta)\|_\Sigma^2 + \|\vartheta - \vartheta_k\|^2_{\frac{1}{\tau_k}I_p}$$

ersetzen und interpretieren.
Im Gegensatz zu dem Vorgehen in Abschnitt 5.1.3 ist aber τ_k im Levenberg-Marquardt-Verfahrens nicht fix, sondern wird entsprechend adjustiert, typischerweise so, dass $\tau_k \to 0$ gilt. Für die lokale Konvergenz ergibt ein ähnliches Verhalten wie beim Gauß-Newton-Verfahren, das wir im folgenden diesbezüglich untersuchen werden. Bei den heute bevorzugt verwendeten Verfahren wird τ_k entsprechend einem Trust-Region-Ansatz eingestellt, vergleiche hierzu Nocedal u. Wright (1999, Chapter 10).

Zuletzt stellen wir noch einen Zusammenhang zwischen dem Levenberg-Marquardt-Verfahren und der Minimum-Norm-Wahl für d_k aus (3.48) her.
Sei $\Sigma^{-1} = U^{\mathrm{T}}U$, bezeichne

$$f : \Theta \to \mathbb{R}^N,\ f(\vartheta) = U(\eta(\vartheta) - y)$$

und für festes $\vartheta \in \operatorname{int}(\Theta)$ sei $J_f(\vartheta) = WDV^{\mathrm{T}}$ eine Singulärwertzerlegung der Jacobi-Matrix von f in ϑ mit $W \in \mathbb{R}^{N\times N}$ orthogonal, $V \in \mathbb{R}^{p\times p}$ orthogonal und $D \in \mathbb{R}^{N\times p}$,

einer Matrix, die nur auf der Hauptdiagonalen Einträge ungleich 0 hat. Dann ist

$$\begin{aligned}(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta))^+ J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta)-y\right) &= (J_f^{\mathrm{T}}(\vartheta)J_f(\vartheta))^+ J_f^{\mathrm{T}}(\vartheta)f(\vartheta)\\ = (VD^{\mathrm{T}}DV^{\mathrm{T}})^+ VD^{\mathrm{T}}W^{\mathrm{T}}f(\vartheta) &= VD^+(D^+)^{\mathrm{T}}V^{\mathrm{T}}VD^{\mathrm{T}}W^{\mathrm{T}}f(\vartheta)\\ &= VD^+W^{\mathrm{T}}f(\vartheta) = J_f^+(\vartheta)f(\vartheta)\end{aligned} \tag{3.50}$$

und

$$\begin{aligned}&\lim_{\substack{\tau\to 0\\ \tau>0}}(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)+\tau I_p)^{-1}J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta)-y\right)\\ &\qquad = \lim_{\substack{\tau\to 0\\ \tau>0}}(J_f^{\mathrm{T}}(\vartheta)J_f(\vartheta)+\tau I_p)^{-1}J_f^{\mathrm{T}}(\vartheta)f(\vartheta) = J_f^+(\vartheta)f(\vartheta),\end{aligned}$$

da für $A\in\mathbb{R}^{N\times p}$ gilt

$$A^+ = \lim_{\substack{\tau\to 0\\ \tau>0}}(A^{\mathrm{T}}A+\tau I_p)^{-1}A^{\mathrm{T}},$$

siehe Harville (1997, Theorem 20.7.1). Die Wahl von d_k als Minimum-Norm-Lösung gemäß (3.48) kann man also als Grenzfall der Bestimmung der Richtung beim Levenberg-Marquardt-Verfahren (3.49) für $\tau_k\to 0$ interpretieren.

Wir betrachten nun die Konvergenzeigenschaften des Gauß-Newton-Verfahrens bzw. des Gauß-Newton-Verfahrens mit Liniensuche.

Satz 3.44
Sei J_η Lipschitz-stetig auf einer offenen Umgebung der Subniveaumenge $U=\{\vartheta\in\mathbb{R}^p : S^{\eta,y}(\vartheta)\le S^{\eta,y}(\vartheta_0)\}$ und $\|J_\eta\|\le L$ auf U. Ferner gebe es ein $\gamma>0$, so dass

$$\|J_\eta(\vartheta)z\|\ge\gamma\|z\| \tag{3.51}$$

für alle $z\in\mathbb{R}^p$ und $\vartheta\in U$.
Dann gilt für jede Folge $(\vartheta_k)_{k\in\mathbb{N}_0}\subset\mathbb{R}^p$, die vom Gauß-Newton-Verfahren (3.45) mit inexakter Liniensuche in Suchrichtung d_k, die die Wolfe-Bedingungen erfüllt, erzeugt wird, dass

$$\lim_{k\to\infty}\operatorname{grad}S^{\eta,y}(\vartheta_k)=\mathbf{0}_p^{\mathrm{T}}.$$

Beweis: Wir weisen nach, dass die Voraussetzungen von Korollar 3.25 erfüllt sind. Wegen der Äquivalenz der Normen gibt es auch ein $\tilde{L}>0$, so dass $\|J_\eta(\vartheta)\|_\Sigma\le\tilde{L}$ für alle $\vartheta\in U$ bzw. ein $\tilde{\gamma}>0$, so dass $\|J_\eta(\vartheta)z\|_\Sigma\ge\tilde{\gamma}\|z\|$ für alle $z\in\mathbb{R}^p$, $\vartheta\in U$.
Wir zeigen nun, dass d_k Abstiegsrichtung und aufgrund der Wolfe-Bedingungen gilt daher induktiv $\vartheta_k\in U$ für alle $k\in\mathbb{N}_0$:

Für den Winkel ϕ_k zwischen d_k und $-\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta_k)$ gilt:

$$\begin{aligned}\cos(\phi_k) &= -\frac{\operatorname{grad} S^{\eta,y}(\vartheta_k) d_k}{\|\operatorname{grad} S^{\eta,y}(\vartheta_k)\| \, \|d_k\|} \\ &\overset{\text{Bem.}}{\underset{(3.45)}{=}} \frac{\|J_\eta(\vartheta_k) d_k\|_\Sigma^2}{\left\|J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1} J_\eta(\vartheta_k) d_k\right\| \, \|d_k\|} \\ &\geq \frac{\tilde{\gamma}^2 \|d_k\|^2}{\|J_\eta(\vartheta_k)\|_\Sigma^2 \, \|d_k\|^2} \geq \frac{\tilde{\gamma}^2}{\tilde{L}^2}.\end{aligned}$$

Wegen (3.51) gilt insbesondere, dass $\operatorname{rg}(J_\eta(\vartheta_k)) = p$. Daher ist nach Bemerkung 3.43 d_k für $k \in \mathbb{N}_0$ Abstiegsrichtung und wegen Bemerkung 3.26d) sind damit die Voraussetzungen von Korollar 3.25 erfüllt. □

Bemerkung 3.45
Die Bedingung (3.51) ist äquivalent dazu, dass für $\vartheta \in U$ sowohl $\operatorname{rg}(J_\eta(\vartheta)) = p$ gilt als auch die Singulärwerte $d_1(\vartheta) \geq \ldots \geq d_p(\vartheta)$ von $J_\eta(\vartheta)$ von 0 weg beschränkt sind, d.h. es gibt ein $\gamma > 0$, so dass $d_p(\vartheta) \geq \gamma$ für $\vartheta \in U$. Nach Satz 3.12 ist $d_p(\vartheta) = \min\{\|J_\eta(\vartheta) - C\| : C \in \mathbb{R}^{N\times p}, \operatorname{rg}(C) < \operatorname{rg}(J_\eta(\vartheta))\}$. Die Jacobi-Matrizen dürfen also in diesem Sinne nicht beliebig nahe an eine Matrix mit Rangdefekt kommen.

Wir widmen uns nun noch den lokalen Konvergenzeigenschaften bzw. der Konvergenzordnung. Es zeigt sich, dass die lokale Konvergenzordnung von der Güte der Approximation der Hesse-Matrix im Konvergenzpunkt abhängt.

Korollar 3.46
Sei η zweimal stetig differenzierbar auf einer offenen konvexen Menge D und $\operatorname{Hess} S^{\eta,y}$ *Lipschitz-stetig auf D. Zudem gelten für die Folgen $(\vartheta_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$, $(d_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$ und $(\lambda_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}_0^+$, die vom Gauß-Newton-Verfahren mit Liniensuche erzeugt werden, die Wolfe-Bedingungen* (3.19) *und* (3.20) *für ein $\alpha < 1/2$ und $(\vartheta_k)_{k\in\mathbb{N}_0}$ konvergiere gegen ein $\vartheta^* \in D$ mit* $\operatorname{rg}(J_\eta(\vartheta^*)) = p$. *Ferner gelte $H_\eta(\vartheta^*) = \mathbf{0}_{p\times p}$. Dann gilt* $\operatorname{grad} S^{\eta,y}(\vartheta^*) = \mathbf{0}_p^{\mathrm{T}}$ *und es gibt ein $k_0 \in \mathbb{N}_0$, so dass für $k \geq k_0$ die Wolfe-Bedingungen von $\tilde{\lambda}_k = 1$ erfüllt werden. Falls es zudem ein $s \in \mathbb{N}_0$ gibt, so dass $\lambda_k = 1$ für $k \geq s$, dann konvergiert $(\vartheta_k)_{k\in\mathbb{N}_0}$ superlinear gegen ϑ^*.*

Beweis: Da mit $G = S^{\eta,y}$

$$\lim_{k\to\infty} \frac{\left\|\operatorname{grad}^{\mathrm{T}} G(\vartheta_k) + \operatorname{Hess} G(\vartheta_k) d_k\right\|}{\|d_k\|} = \lim_{k\to\infty} \frac{\|2H_\eta(\vartheta_k) d_k\|}{\|d_k\|} = 0$$

und weil wegen $\operatorname{rg}(J_\eta(\vartheta^*)) = p$ die Hesse-Matrix $\operatorname{Hess} S^{\eta,y}(\vartheta^*) = 2J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)$ positiv definit ist, ist Satz 3.38 anwendbar. □

Bei Anwendung eines Backtracking-Algorithmus stimmen dann in der Nähe von ϑ^* die Gauß-Newton-Verfahren mit und ohne Liniensuche überein. Fordert man, dass J_η

Lipschitz-stetig und auf D beschränkt ist, dann erhält man sogar quadratische Konvergenz.

Satz 3.47
Sei η zweimal stetig differenzierbar auf einer offenen konvexen Menge D, J_η Lipschitz-stetig auf D und $\|J_\eta\| \leq \kappa$ auf D, $\vartheta^* \in D$ mit $\operatorname{grad} S^{\eta,y}(\vartheta^*) = 0$, $\operatorname{rg}(J_\eta(\vartheta^*)) = p$ *und $H_\eta(\vartheta^*) = \mathbf{0}_{p\times p}$. Dann gibt es eine Umgebung U von ϑ^*, so dass für alle $\vartheta_0 \in U$ die vom Gauß-Newton-Verfahren erzeugte Folge $(\vartheta_k)_{k\in\mathbb{N}_0}$ quadratisch gegen ϑ^* konvergiert.*

Beweis: Dennis u. Schnabel (1983, Theorem 10.2.1 und Corollary 10.2.2) □

Im Fall $\eta(\vartheta^*) = y$ oder im Fall[26] $\operatorname{Hess}\eta_i(\vartheta^*) = \mathbf{0}_{p\times p}$ für $i = 1,\ldots,N$ gilt offensichtlich $H_\eta(\vartheta^*) = \mathbf{0}_{p\times p}$. Naheliegend ist daher die Vermutung, dass für kleine Residuen $y - \eta(\vartheta^*) \approx 0$ oder wenn η in ϑ^* fast linear ist, auch noch gute lokale Konvergenzeigenschaften vorliegen. Dies präzisiert Satz 3.63. Wir benötigen für den Beweis noch einige Vorbereitungen.

Definition 3.48
Ein Paar $\mathcal{J} = (\mathcal{G}, D)$ mit $D \subseteq \mathbb{R}^p$, $D \neq \emptyset$ und $\mathcal{G} : D \to \mathbb{R}^p$ heißt (sequentieller, stationärer) **1-Schritt-Prozess**.
Ein Punkt $\vartheta^* \in \mathbb{R}^p$ heißt Grenzwert des Prozesses $\mathcal{J}$, falls es ein $\vartheta_0 \in D$ gibt, so dass die durch $\vartheta_{k+1} = \mathcal{G}(\vartheta_k)$ für $k \in \mathbb{N}_0$ rekursiv definierte Folge wohldefiniert ist (d.h. $\mathcal{G}(\vartheta_k) \in D$ für $k \in \mathbb{N}_0$) und Grenzwert ϑ^* hat. Die Menge aller solchen Folgen mit Grenzwert ϑ^* bezeichnen wir mit $\mathcal{C}(\mathcal{J}, \vartheta^*)$.

Definition 3.49

a) Sei $(\vartheta_k)_{k\in\mathbb{N}_0} \subseteq \mathbb{R}^p$ eine Folge, die gegen $\vartheta^* \in \mathbb{R}^p$ konvergiert. Dann heißt

$$R_1\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right) = \limsup_{k\to\infty} \|\vartheta_k - \vartheta^*\|^{1/k}$$

der **asymptotische Wurzel-Konvergenzfaktor der Folge** zu $q = 1$. Entsprechend definiert man für $q > 1$

$$R_q\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right) = \limsup_{k\to\infty} \|\vartheta_k - \vartheta^*\|^{1/q^k}.$$

b) Sei $\mathcal{J}$ ein 1-Schritt-Prozess und ϑ^* ein Grenzwert des Prozesses und $q \in [1,\infty)$. Dann heißt

$$R_q(\mathcal{J}, \vartheta^*) = \sup\{R_q\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right) \mid (\vartheta_k)_{k\in\mathbb{N}_0} \in \mathcal{C}(\mathcal{J}, \vartheta^*)\}$$

der **asymptotische Wurzel-Konvergenzfaktor des 1-Schritt-Prozesses** $\mathcal{J}$ in ϑ^* zu q.

[26] Es gilt insbesondere dann $\operatorname{Hess}\eta_i(\vartheta^*) = \mathbf{0}_{p\times p}$ für $i = 1,\ldots,N$, wenn η linear in einer Umgebung von ϑ^* ist. Nach Bemerkung 3.42 ist man dann in dieser Umgebung allerdings bereits nach einem Schritt im Optimum.

Bemerkung 3.50
Konvergiert die Folge $(\vartheta_k)_{k\in\mathbb{N}_0} \subseteq \mathbb{R}^p$ gegen ϑ^*, dann gibt es ein $\tilde{k} \in \mathbb{N}_0$, so dass $\|\vartheta_k - \vartheta^*\| < 1$ für $k \geq \tilde{k}$. Daher gilt $0 \leq R_q\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right) \leq 1$.

Lemma 3.51
Sei $\mathcal{J}$ ein 1-Schritt-Prozess und ϑ^* ein Grenzwert des Prozesses. Dann gilt genau eine der folgenden Aussagen:

a) $R_q(\mathcal{J}, \vartheta^*) = 0$ für alle $q \in [1, \infty)$.

b) Es gibt ein $q_0 \in [1, \infty)$, so dass $R_q(\mathcal{J}, \vartheta^*) = 0$ für alle $q \in [1, q_0)$ und $R_q(\mathcal{J}, \vartheta^*) = 1$ für alle $q \in (q_0, \infty)$.

c) $R_q(\mathcal{J}, \vartheta^*) = 1$ für alle $q \in [1, \infty)$.

Beweis: Ortega u. Rheinboldt (1970, Theorem 9.2.3) □

Definition 3.52
Sei $\mathcal{J}$ ein 1-Schritt-Prozess und ϑ^* ein Grenzwert des Prozesses. Dann heißt

$$O_R(\mathcal{J}, \vartheta^*) = \begin{cases} \infty, & \text{falls } R_q(\mathcal{J}, \vartheta^*) = 0 \text{ für alle } q \in [1, \infty), \\ \inf\{q \in [1, \infty) : R_q(\mathcal{J}, \vartheta^*) = 1\}, & \text{sonst,} \end{cases}$$

die **Wurzel-Ordnung** des Prozesses $\mathcal{J}$ in ϑ^*.

Bemerkung 3.53
Ist $R_q(\mathcal{J}, \vartheta^*) < 1$ für ein $q \in [1, \infty)$, dann gilt nach Lemma 3.51, dass $O_R(\mathcal{J}, \vartheta^*) \geq q$.
Ist $R_q(\mathcal{J}, \vartheta^*) > 0$ für ein $q \in [1, \infty)$, dann gilt nach Lemma 3.51, dass $O_R(\mathcal{J}, \vartheta^*) \leq q$.
Ist also $0 < R_q(\mathcal{J}, \vartheta^*) < 1$ für ein $q \in [1, \infty)$, dann gilt $O_R(\mathcal{J}, \vartheta^*) = q$.

Definition 3.54

a) Sei $(\vartheta_k)_{k\in\mathbb{N}_0} \subseteq \mathbb{R}^p$ eine Folge, die gegen ϑ^* konvergiert und sei $q \in [1, \infty)$. Dann heißt

$$Q_q\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right) = \begin{cases} 0, & \text{falls } \vartheta_k = \vartheta^* \text{ für } k \in \mathbb{N}_0 \setminus A \text{ mit } A \text{ endl.}, \\ \limsup\limits_{k\to\infty} \frac{\|\vartheta_{k+1} - \vartheta^*\|}{\|\vartheta_k - \vartheta^*\|^q}, & \text{falls } \vartheta_k \neq \vartheta^* \text{ für } k \in \mathbb{N}_0 \setminus B \text{ mit } B \text{ endl.}, \\ \infty, & \text{sonst}^{27}, \end{cases}$$

der **(Quotienten-)Konvergenzfaktor der Folge** $(\vartheta_k)_{k\in\mathbb{N}_0}$ zu q in ϑ^*.

[27] d.h. es gibt eine unendliche Menge $C \subset \mathbb{N}_0$, so dass $\mathbb{N}_0 \setminus C$ unendlich und $\vartheta_k = \vartheta^*$ für $k \in C$ und $\vartheta_k \neq \vartheta^*$ für $k \in \mathbb{N}_0 \setminus C$. In den von uns betrachteten iterativen Prozessen kann dieser letzte Fall nicht auftreten, da aufgrund der Stetigkeit der Abbildung $\mathcal{G}$ Grenzwerte stets Fixpunkte sind.

b) Sei $\mathcal{J}$ ein 1-Schritt-Prozess und ϑ^* ein Grenzwert des Prozesses und $q \in [1, \infty)$. Dann heißt

$$Q_q(\mathcal{J}, \vartheta^*) = \sup\{Q_q\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right) : (\vartheta_k)_{k\in\mathbb{N}_0} \in \mathcal{C}(\mathcal{J}, \vartheta^*)\}$$

der **(Quotienten-)Konvergenzfaktor des Prozesses** $\mathcal{J}$ zu q in ϑ^* .

Lemma 3.55
Sei $\mathcal{J}$ ein 1-Schritt-Prozess und ϑ^* ein Grenzwert des Prozesses. Dann gilt genau eine der folgenden Aussagen:

a) $Q_q(\mathcal{J}, \vartheta^*) = 0$ *für alle* $q \in [1, \infty)$.

b) Es gibt ein $q_0 \in [1, \infty)$, *so dass* $Q_q(\mathcal{J}, \vartheta^*) = 0$ *für alle* $q \in [1, q_0)$ *und* $Q_q(\mathcal{J}, \vartheta^*) = \infty$ *für alle* $q \in (q_0, \infty)$.

c) $Q_q(\mathcal{J}, \vartheta^*) = \infty$ *für alle* $q \in [1, \infty)$.

Beweis: Ortega u. Rheinboldt (1970, Theorem 9.1.3) □

Definition 3.56
Sei $\mathcal{J}$ ein 1-Schritt-Prozess und ϑ^* ein Grenzwert des Prozesses. Dann heißt

$$O_Q(\mathcal{J}, \vartheta^*) = \begin{cases} \infty, & \text{falls } Q_q(\mathcal{J}, \vartheta^*) = 0 \text{ für alle } q \in [1, \infty) \\ \inf\{q \in [1, \infty) : Q_q(\mathcal{J}, \vartheta^*) = \infty\}, & \text{sonst,} \end{cases}$$

die **(Quotienten-)Ordnung des Prozesses** $\mathcal{J}$ in ϑ^*.

Bemerkung 3.57

a) Gibt es ein $q \in [1, \infty)$ mit $Q_q(\mathcal{J}, \vartheta^*) < \infty$, so gilt nach Lemma 3.55, dass $O_Q(\mathcal{J}, \vartheta^*) \geq q$. Ist $Q_q(\mathcal{J}, \vartheta^*) > 0$ für ein $q \in [1, \infty)$, dann gilt nach Lemma 3.55, dass $O_Q(\mathcal{J}, \vartheta^*) \leq q$. Ist also $0 < Q_q(\mathcal{J}, \vartheta^*) < \infty$ für ein $q \in [1, \infty)$, dann gilt $O_Q(\mathcal{J}, \vartheta^*) = q$.

b) Falls $Q_1(\mathcal{J}, \vartheta^*) < 1$, dann nennt man den Prozess $\mathcal{J}$ **linear konvergent** in ϑ^*, gilt $Q_1(\mathcal{J}, \vartheta^*) = 0$, dann nennt man den Prozess $\mathcal{J}$ **superlinear konvergent** in ϑ^* und gilt $Q_2(\mathcal{J}, \vartheta^*) < 1$, dann nennt man den Prozess $\mathcal{J}$ **quadratisch konvergent** in ϑ^*. Dies deckt sich mit den bisherigen Bezeichnungen.

Lemma 3.58

a) Sei $(\vartheta_k)_{k\in\mathbb{N}_0} \subset \mathbb{R}^p$ *eine Folge, die gegen* ϑ^* *konvergiert. Dann gilt*

$$R_1\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right) \leq Q_1\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right).$$

b) Ist $\mathcal{J}$ ein 1-Schritt-Prozess und ϑ^* ein Grenzwert des Prozesses, dann gilt

$$O_Q(\mathcal{J}, \vartheta^*) \leq O_R(\mathcal{J}, \vartheta^*).$$

Beweis: Ortega u. Rheinboldt (1970, Theorem 9.3.1 und 9.3.2) □

Bemerkung 3.59
Hat man also eine Abschätzung der Form

$$\|\vartheta_{k+1} - \vartheta^*\| \leq c \|\vartheta_k - \vartheta^*\| \text{ für } k \geq k_0,$$

dann liefert $R_1\left((\vartheta_k)_{k\in\mathbb{N}_0}, \vartheta^*\right)$ eine untere Schranke für die Konstante c.

Satz 3.60 (Ostrowskis Theorem)
Sei $\mathcal{J} = (\mathcal{G}, D)$ ein 1-Schritt-Prozess, $\vartheta^* \in \operatorname{int}(D)$ ein Fixpunkt von $\mathcal{G}$ und $\mathcal{G}$ differenzierbar in ϑ^*. Es gelte

$$\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) < 1$$

für den Spektralradius von $J_{\mathcal{G}}(\vartheta^*)$.

*a) Dann ist ϑ^* **Attraktorpunkt des Prozesses** $\mathcal{J}$, d.h. es gibt eine offene Umgebung $U \subseteq D$ von ϑ^*, so dass für alle $\vartheta_0 \in U$ die durch den Prozess definierten Folgen in D liegen und gegen ϑ^* linear konvergieren.*

b) $R_1(\mathcal{J}, \vartheta^*) = \rho\left(J_{\mathcal{G}}(\vartheta^*)\right)$.

c) Ist $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) > 0$, dann gilt $O_R(\mathcal{J}, \vartheta^*) = O_Q(\mathcal{J}, \vartheta^*) = 1$.

Beweis: Ortega u. Rheinboldt (1970, Theorem 10.1.3 und 10.1.4) □

Wir benötigen als letzte Vorbereitung für die Konvergenzuntersuchungen zum Gauß-Newton-Verfahren noch eine Formel für die Ableitung der Moore-Penrose-Inversen.

Lemma 3.61
Sei $U \subseteq \mathbb{R}^p$ offen und $F : U \to \mathbb{R}^{m\times n}$ stetig differenzierbar. Es gelte $\operatorname{rg}(F)$ *ist konstant in einer Umgebung von $\vartheta \in U$. Dann ist F^+ in ϑ differenzierbar und es gilt für $l \in \{1, \ldots, p\}$, dass*

$$\begin{aligned}\frac{\partial F^+}{\partial \vartheta_l}(\vartheta) = &- F^+(\vartheta)\frac{\partial F}{\partial \vartheta_l}(\vartheta)F^+(\vartheta) \\ &+ F^+(\vartheta)\left(F^+(\vartheta)\right)^{\mathrm{T}}\left(\frac{\partial F}{\partial \vartheta_l}(\vartheta)\right)^{\mathrm{T}}\left(I_m - F(\vartheta)F^+(\vartheta)\right) \\ &+ \left(I_n - F^+(\vartheta)F(\vartheta)\right)\left(\frac{\partial F}{\partial \vartheta_l}(\vartheta)\right)^{\mathrm{T}}\left(F^+(\vartheta)\right)^{\mathrm{T}}F^+(\vartheta).\end{aligned}$$

Beweis: siehe Harville (1997, Theorem 20.8.2) □

Korollar 3.62
Sei $U \subseteq \mathbb{R}^p$ offen und $F : U \to \mathbb{R}^{n\times n}$ stetig differenzierbar in U, $\vartheta \in U$ und $F(\vartheta)$ sei regulär. Dann ist F^{-1} in ϑ differenzierbar und es gilt für $l \in \{1,\dots,p\}$, dass

$$\frac{\partial F^{-1}}{\partial \vartheta_l}(\vartheta) = -F^{-1}(\vartheta)\frac{\partial F}{\partial \vartheta_l}(\vartheta)F^{-1}(\vartheta).$$

Beweis: Spezialfall von Lemma 3.61 oder für einen direkten Beweis siehe Harville (1997, Section 15.8) □

Satz 3.63
Sei η zweimal differenzierbar und bezeichne $\mathcal{J} = (\mathcal{G}, D)$ den Gauß-Newton-Prozess, d.h.

$$\mathcal{G} : D \to \mathbb{R}^p,\ \mathcal{G}(\vartheta) = \vartheta - \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1} J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - y\right)$$

für $D \subseteq \mathbb{R}^p$. Ist $\vartheta^* \in \mathrm{int}(D)$ mit $\mathrm{rg}(J_\eta(\vartheta^*)) = p$, dann ist ϑ^* genau dann ein stationärer Punkt von $S^{\eta,y}$, wenn ϑ^* ein Fixpunkt von $\mathcal{G}$ ist.
Sei ϑ^* ein stationärer Punkt von $S^{\eta,y}$ mit $\mathrm{rg}(J_\eta(\vartheta^*)) = p$, dann gilt

$$J_{\mathcal{G}}(\vartheta^*) = -\left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)\right)^{-1} H_\eta(\vartheta^*).$$

Gilt $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) < 1$, dann ist ϑ^* Attraktorpunkt des Prozesses $\mathcal{J}$ und der Prozess ist linear konvergent und der Quotientenkonvergenzfaktor beträgt bestenfalls $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right)$.
Die Bedingung $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) = 0$ ist eine notwendige Bedingung dafür, dass der Prozess superlinear konvergent ist.

Beweis: Sei $\vartheta^* \in \mathrm{int}(D)$ mit $\mathrm{rg}(J_\eta(\vartheta^*)) = p$. Dann gilt

$$\begin{aligned}\mathcal{G}(\vartheta^*) = \vartheta^* &\Leftrightarrow \left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)\right)^{-1} J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}\left(\eta(\vartheta^*) - y\right) = \mathbf{0}_p \\ &\Leftrightarrow \mathrm{grad}^{\,\mathrm{T}} S^{\eta,y}(\vartheta^*) = 2J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}\left(\eta(\vartheta^*) - y\right) = \mathbf{0}_p,\end{aligned}$$

d.h. $\vartheta^* \in \mathrm{int}(D)$ ist genau dann ein stationärer Punkt von $S^{\eta,y}$, wenn ϑ^* ein Fixpunkt von $\mathcal{G}$ ist.
Für die Ableitung von $\mathcal{G}$ gilt für $l \in \{1,\dots,p\}$ wegen Korollar 3.62

$$\frac{\partial}{\partial \vartheta_l}\mathcal{G}(\vartheta) = e_l - \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1}\left(\frac{1}{2}\mathrm{Hess}\, S^{\eta,y}(\vartheta)\right)_l + \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1} b_l(\vartheta)$$

mit

$$b_l(\vartheta) = \left(\frac{\partial}{\partial \vartheta_l}\left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)\right)\left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1}\frac{1}{2}\mathrm{grad}\, S^{\eta,y}(\vartheta).$$

Mit $B(\vartheta) = [b_1(\vartheta),\dots,b_p(\vartheta)]$ folgt

$$\begin{aligned}J_{\mathcal{G}}(\vartheta) &= I_p - \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1}\left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + H_\eta(\vartheta)\right) + \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1} B(\vartheta) \\ &= \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1}\left(-H_\eta(\vartheta) + B(\vartheta)\right).\end{aligned}$$

Sei nun $\vartheta^* \in \operatorname{int}(D)$ ein stationärer Punkt von $S^{\eta,y}$. Dann gilt $B(\vartheta^*) = \mathbf{0}_{p\times p}$ und daher

$$J_{\mathcal{G}}(\vartheta^*) = -\left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)\right)^{-1} H_\eta(\vartheta^*).$$

Die restlichen Aussagen folgen mit Satz 3.60. □

Die Bedingung $\rho\left(\left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)\right)^{-1} H_\eta(\vartheta^*)\right) < 1$ ist schwächer[28] als die Bedingung

$$\left\|\left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)\right)^{-1} H_\eta(\vartheta^*)\right\| \ll 1,$$

die mit einem heuristischen Argument in Nocedal u. Wright (1999, Chapter 10) hergeleitet wurde, bzw. deckt sich mit den in Ramsin u. Wedin (1977) angegebenen Resultaten. Für eine geometrische Interpretation von $J_{\mathcal{G}}(\vartheta^*)$ mittels Normalkrümmungen siehe Wedin (1974) oder Björck (1996, Section 9.1.2). Es liegt also Konvergenz gegen einen stationären Punkt ϑ^* mit $\operatorname{rg}(J_\eta(\vartheta^*)) = p$ bei Startwerten in der Nähe von ϑ^* vor, falls η fast linear in ϑ^* ist, d.h. $\operatorname{Hess}\eta_j(\vartheta^*) \approx 0_{p\times p}$ gilt, bzw. falls das Residuum in ϑ^* klein ist, d.h. $\eta(\vartheta^*) \approx y$ gilt, weil in beiden Fällen dann $H_\eta(\vartheta^*) \approx 0_{p\times p}$ und daher dann $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) < 1$.

Ist $\operatorname{rg}(J_\eta(\vartheta^*)) < p$, so kann man wie in Bemerkung 3.43 stattdessen die Minimum-Norm-Lösung (3.48) der Normalgleichungen (3.44) betrachten, d.h. für entsprechendes D den 1-Schritt-Prozess $\mathcal{J} = (\mathcal{G}, D)$ mit

$$\mathcal{G} : D \to \mathbb{R}^p,\ \mathcal{G}(\vartheta) = \vartheta - (J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta))^{+} J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - y\right), \tag{3.52}$$

wobei mit $\Sigma^{-1} = U^{\mathrm{T}}U$ und $f(\vartheta) = U(\eta(\vartheta) - y)$ gilt, dass

$$G(\vartheta) = \vartheta - J_f^{+}(\vartheta)f(\vartheta), \tag{3.53}$$

siehe (3.50). 1-Schritt-Prozesse der Form (3.53) wurden auch von Levin u. Ben-Israel (2001) untersucht, die unter Bedingungen, dass J_f Lipschitz-stetig ist, dass die zweiten Ableitungen von f nicht „zu groß“ und dass die ersten Ableitungen von f nicht „zu klein“ sind, quadratische Konvergenz gegen einen stationären Punkt von

$$S^{\eta,y} = S^f : \mathbb{R}^p \to \mathbb{R},\ S^f(\vartheta) = \|f(\vartheta)\|^2 = \|\eta(\vartheta) - y\|_\Sigma^2$$

zeigen. Insbesondere die Voraussetzung, dass es ein $c > 0$ und ein $r > 0$ gibt, so dass

$$\|(J_f^{+}(\vartheta) - J_f^{+}(\tilde{\vartheta}))f(\vartheta)\| \le c\,\|\vartheta - \tilde{\vartheta}\|^2 \tag{3.54}$$

für alle $\vartheta, \tilde{\vartheta} \in B_r(\vartheta_0)$, wobei $B_r(\vartheta_0)$ die offene Kugel vom Radius r um den Startwert ϑ_0 bezeichnet, scheint aber kaum erfüllbar zu sein, da in der Regel die linke Seite von (3.54) nur durch einen linearen Term der Form $c\,\|\vartheta - \tilde{\vartheta}\|$ und nicht durch einen quadratischen Term nach oben abgeschätzt werden kann. Die in Levin u. Ben-Israel (2001) untersuchten Beispiele erfüllen jedenfalls die Voraussetzung (3.54) nicht.

[28] Für $A \in \mathbb{R}^{p\times p}$ gilt $\rho(A) \le \|A\|$, da insbesondere für den betragsgrößten Eigenwert λ und den zugehörigen Eigenvektor u gilt, dass $\|Au\| = |\lambda|\,\|u\|$.

Wir untersuchen, ob sich das Vorgehen im Beweis von Satz 3.63 auf den Fall $\operatorname{rg}(J_\eta(\vartheta^*)) < p$ und den Prozess (3.52) ausweiten lässt. Zum einen gilt:[29]

$$\begin{aligned}\mathcal{G}(\vartheta^*) = \vartheta^* &\Leftrightarrow J_f^+(\vartheta^*) f(\vartheta^*) = \mathbf{0}_p \Leftrightarrow f(\vartheta^*) \in \operatorname{Null}(J_f^+(\vartheta^*)) = \operatorname{Null}(J_f^{\mathrm{T}}(\vartheta^*)) \\ &\Leftrightarrow J_f^{\mathrm{T}}(\vartheta^*) f(\vartheta^*) = \mathbf{0}_p \Leftrightarrow \operatorname{grad} S^{\eta,y} = 2 J_\eta^{\mathrm{T}}(\vartheta) \Sigma^{-1} \left(\eta(\vartheta) - y\right) = \mathbf{0}_p,\end{aligned}$$

d.h. genau alle stationären Punkte sind Fixpunkte. Ferner gilt mit Lemma 3.61 für stationäre Punkte, in deren Umgebung der Rang von J_η konstant ist,

$$\begin{aligned}J_{\mathcal{G}}(\vartheta^*) &= I_p - \left(J_\eta^{\mathrm{T}}(\vartheta^*) \Sigma^{-1} J_\eta(\vartheta^*)\right)^+ \frac{1}{2} \operatorname{Hess} S^{\eta,y}(\vartheta^*) \\ &= I_p - P_{J_\eta^{\mathrm{T}}(\vartheta^*)} - \left(J_\eta^{\mathrm{T}}(\vartheta^*) \Sigma^{-1} J_\eta(\vartheta^*)\right)^+ H_\eta(\vartheta^*)\end{aligned}$$

und man kann damit den Satz 3.60 anwenden. Ist $\operatorname{rg}(J_\eta(\vartheta^*)) < p$ und $H_\eta(\vartheta^*) = 0_{p\times p}$, dann ist $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) = 1$. Im Fall $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) = 1$ kann man aber keine Aussage über die Konvergenz treffen, es kommt dann auf die quadratischen (oder höheren) Terme in der Taylorentwicklung von $\mathcal{G}$ in ϑ^* an, siehe Ortega u. Rheinboldt (1970, Exercise 10.1-2). Wir beenden damit unsere Untersuchungen zum Konvergenzverhalten des Gauß-Newton-Verfahrens und verweisen noch auf die ausführlichen numerischen Studien zu Problemen des Gauß-Newton-Verfahrens in Fraley (1989).

Im Falle großer Residuen oder stark nichtlinearer Modellfunktionen η empfiehlt sich – vergleiche Satz 3.63 – das Gauß-Newton-Verfahren nicht, sondern sind Quasi-Newton-Verfahren geeigneter. Wir werden mit dem BFGS-Verfahren das Quasi-Newton-Verfahren vorstellen, das von den meisten Autoren im Allgemeinen für das effektivste Quasi-Newton-Verfahren gehalten wird.

3.2.3. Quasi-Newton-Verfahren: BFGS und weitere Verfahren

Wir kehren zum allgemeinen Problem der Minimierung einer differenzierbaren Funktion $G : \mathbb{R}^p \to \mathbb{R}$ zurück. Bei Quasi-Newton-Verfahren wird wie beim Gauß-Newton-Verfahren (für die eingeschränkte Funktionenklasse) versucht, die Berechnung der Hesse-Matrizen durch eine Approximation zu vermeiden, die einfacher berechnet werden kann. Statt der Newton-Iterationsvorschrift

$$\vartheta_{k+1} = \vartheta_k - \lambda_k \left(\operatorname{Hess} G(\vartheta_k)\right)^{-1} \operatorname{grad}^{\mathrm{T}} G(\vartheta_k)$$

betrachtet man eine Iterationsvorschrift der Form

$$\vartheta_{k+1} = \vartheta_k - \lambda_k H_k \operatorname{grad}^{\mathrm{T}} G(\vartheta_k),$$

wobei H_k im $(k+1)$-ten Schritt also $(\operatorname{Hess} G(\vartheta_k))^{-1}$, die Inverse der Hesse-Matrix, ersetzt. Man erspart sich also gegenüber dem Newton-Verfahren bei geeigneter Definition von H_k sogar noch die Lösung linearer Gleichungssysteme. Wir stellen hier ein spezielles Quasi-Newton-Verfahren vor, das BFGS-Verfahren, das unabhängig voneinander von Broyden (1970), Fletcher (1970), Goldfarb (1970) und Shanno (1970) entwickelt wurde, und betrachten noch einige leichte Abwandlungen davon.

[29] Für $A \in \mathbb{R}^{m\times n}$ gilt $\operatorname{Null}(A^+) = \operatorname{Null}(A^{\mathrm{T}})$, siehe Harville (1997, Theorem 20.5.1).

Algorithmus 3.64 (BFGS-Verfahren ohne Liniensuche)
Input: $G : \mathbb{R}^p \to \mathbb{R}$ stetig differenzierbar,
Startpunkt $\vartheta_0 \in \mathbb{R}^p$,
symmetrische, positiv definite Matrix $H_0 \in \mathbb{R}^{p\times p}$ (üblich $H_0 = I_p$),
Abbruchkriterium
Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation der Minimumstelle von G)

1. Setze $k = 0$ und berechne $g_0 = \text{grad}^{\,\mathrm{T}} G(\vartheta_0)$.
2. Berechne $p_k = d_k = -H_k g_k$, $\vartheta_{k+1} = \vartheta_k + p_k$ und den neuen Gradienten $g_{k+1} = \text{grad}^{\,\mathrm{T}} G(\vartheta_{k+1})$.
3. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten berechne $r_k = g_{k+1} - g_k$,

$$v_k = \sqrt{r_k^{\mathrm{T}} H_k r_k}\left(\frac{1}{p_k^{\mathrm{T}} r_k}\, p_k - \frac{1}{r_k^{\mathrm{T}} H_k r_k}\, H_k r_k\right)$$

und

$$H_{k+1} = H_k + \frac{1}{p_k^{\mathrm{T}} p_k} p_k p_k^{\mathrm{T}} - \frac{1}{r_k^{\mathrm{T}} H_k r_k}(H_k r_k)(H_k r_k)^{\mathrm{T}} + v_k v_k^{\mathrm{T}}, \tag{3.55}$$

setze $k := k + 1$ und gehe zu Schritt 2.

Algorithmus 3.65 (BFGS-Verfahren mit Liniensuche)
Input: $G : \mathbb{R}^p \to \mathbb{R}$ stetig differenzierbar,
Startpunkt $\vartheta_0 \in \mathbb{R}^p$,
symmetrische, positiv definite Matrix $H_0 \in \mathbb{R}^{p\times p}$ (üblich $H_0 = I_p$)
Abbruchkriterium
Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation der Minimumstelle von G)

1. Setze $k = 0$ und berechne $g_0 = \text{grad}^{\,\mathrm{T}} G(\vartheta_0)$.
2. Bestimme λ_k durch eine (event. inexakte) Liniensuche in ϑ_k mit Suchrichtung $d_k = -H_k g_k$ und setze $p_k = \lambda_k d_k$ und $\vartheta_{k+1} = \vartheta_k + p_k$.
3. Berechne den neuen Gradienten $g_{k+1} = \text{grad}^{\,\mathrm{T}} G(\vartheta_{k+1})$.
4. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten berechne $r_k = g_{k+1} - g_k$,

$$v_k = \sqrt{r_k^{\mathrm{T}} H_k r_k}\left(\frac{1}{p_k^{\mathrm{T}} r_k}\, p_k - \frac{1}{r_k^{\mathrm{T}} H_k r_k}\, H_k r_k\right)$$

und

$$H_{k+1} = H_k + \frac{1}{p_k^{\mathrm{T}} p_k} p_k p_k^{\mathrm{T}} - \frac{1}{r_k^{\mathrm{T}} H_k r_k}(H_k r_k)(H_k r_k)^{\mathrm{T}} + v_k v_k^{\mathrm{T}}, \tag{3.56}$$

setze $k := k + 1$ und gehe zu Schritt 2.

Bemerkung 3.66
Die wesentliche Idee vieler Quasi-Newton-Verfahren liegt darin, wie in einem Newton-Verfahren mit Liniensuche vorzugehen, aber im k-ten Schleifendurchlauf die Hesse-Matrix $\operatorname{Hess} G(\vartheta_{k+1})$ für die nächste Iteration durch eine symmetrische (und möglichst noch positiv definite) Matrix B_{k+1} oder die inverse Hesse-Matrix $(\operatorname{Hess} G(\vartheta_{k+1}))^{-1}$ durch eine symmetrische (und möglichst noch positiv definite) Matrix H_{k+1} zu approximieren, indem man zum einen die Gradienten $g_{k+1} = \operatorname{grad}^{\mathrm{T}} G(\vartheta_{k+1})$ und $g_k = \operatorname{grad}^{\mathrm{T}} G(\vartheta_k)$ verwendet, um die Sekantengleichung

$$B_{k+1} p_k = r_k$$

mit $p_k = \vartheta_{k+1} - \vartheta_k$ und $r_k = g_{k+1} - g_k$ bzw.

$$H_{k+1} r_k = p_k$$

zu erfüllen. Die Sekantengleichung kann man so interpretieren, dass der Gradient der quadratischen Funktion

$$q_{k+1} : \mathbb{R}^p \to \mathbb{R},\ q_{k+1}(\vartheta) = G(\vartheta_{k+1}) + g_{k+1}^{\mathrm{T}} \vartheta + \frac{1}{2}\, \vartheta^{\mathrm{T}} B_{k+1} \vartheta$$

in ϑ_k und ϑ_{k+1} mit dem Gradienten von G übereinstimmt.

Zum anderen stellt man die Forderung an B_{k+1} bzw. H_{k+1}, dass die Matrix sich nicht zu stark von B_k bzw. H_k unterscheidet und das Update wenig aufwendig ist. So basieren einige Quasi-Newton-Verfahren darauf, dass $B_{k+1} - B_k$ bzw. $H_{k+1} - H_k$ eine Matrix vom Rang 1 ist, vergleiche z.B. Nocedal u. Wright (1999, Section 8.2). Beim DFP-Verfahren[30] ist $B_{k+1} - B_k$ eine Matrix vom Rang 2, die man als Lösung

$$B_{k+1} = \operatorname*{arg\,min}_{\substack{B \in \mathbb{R}^{p \times p} \text{ mit} \\ B = B^{\mathrm{T}} \text{ und } B p_k = r_k}} \|B - B_k\|_W$$

interpretieren kann mit

$$W = \left(\int_0^1 \operatorname{Hess} G(\vartheta_k + t p_k)\ \mathrm{d}t \right)^{-1}$$

und der gewichteten Frobenius-Norm

$$\|A\|_W = \left\| W^{1/2} A W^{1/2} \right\|_{\mathrm{Frob}} \quad \text{für } A \in \mathbb{R}^{p \times p}.$$

Beim BFGS-Verfahren ist stattdessen $H_{k+1} - H_k$ eine Matrix vom Rang 2, die man als Lösung

$$H_{k+1} = \operatorname*{arg\,min}_{\substack{H \in \mathbb{R}^{p \times p} \text{ mit} \\ H = H^{\mathrm{T}} \text{ und } H r_k = p_k}} \|H - H_k\|_W$$

[30] benannt nach Davidon (1959) und Fletcher u. Powell (1963)

auffassen kann.[31]

Gegenüber dem Newton-Verfahren spart man sich beim BFGS-Verfahren nicht nur die eventuell aufwendige Berechnung der Hesse-Matrix, sondern statt der Lösung eines Gleichungssystems fallen lediglich Vektor-Vektor- und Matrix-Vektor-Multiplikationen an, so dass bei jedem einzelnen Schritt $O(p^3)$ Rechenoperationen durch $O(p^2)$ Rechenoperationen ersetzt werden.

Lemma 3.67
Sei H_k positiv definit und es gelte

$$p_k^{\mathrm{T}} r_k > 0 \quad \textit{bzw. äquivalent} \quad \operatorname{grad} G(\vartheta_{k+1}) d_k > \operatorname{grad} G(\vartheta_k) d_k. \tag{3.57}$$

Dann ist auch H_{k+1} positiv definit.

Beweis: Dennis u. Schnabel (1983, Lemma 9.2.1) bzw. Bertsekas (1999, Proposition 1.7.1) □

Korollar 3.68
Ist H_0 positiv definit und verwendet man in Schritt 2 eine inexakte Liniensuche, die die Wolfe-Bedingungen oder die strengen Wolfe-Bedingungen einhält, dann ist H_k für alle $k \in \mathbb{N}_0$ positiv definit.

Beweis: Die Forderung (3.57) ist durch die Bedingung (3.34), also die 2. Bedingung aus den Wolfe-Bedingungen, erfüllt. □

Bemerkung 3.69
Eine naheliegende Alternative zu $H_0 = I_p$ ist $H_0 = (\operatorname{Hess} G(\vartheta_0))^{-1}$ zu verwenden, indem man die Hesse-Matrix durch Einsetzen in eine Formel oder durch numerisches Ableiten mittels finiter Differenzen berechnet, vgl. Dennis u. Schnabel (1983, Section 4.2 und Algorithmen A5.6.1 f.). Dieses Vorgehen kann nur empfohlen werden, wenn man sich schon in der Nähe des Minimums befindet. Zum einen, weil man nur dort sichergehen kann, dass man so eine positiv definite Matrix erhält und zum anderen, weil bei einer größeren Entfernung die Richtung des steilsten Abstiegs, die man im ersten Schritt bei $H_0 = I_p$ verwendet, in praktischen Anwendungen meist anfangs zu starken Verbesserungen in der Zielfunktion G führt.
Ferner sind sogenannte „Restarts“ populär, d.h. man unterbricht die Iterationsfolge für H_k nach einer fest vorgegebenen Schrittzahl und verwendet statt (3.55) bzw. (3.56) $H_{k+1} = I_p$, d.h. startet sozusagen mit dem aktuellen Wert für ϑ_{k+1} als neuem Startwert ϑ_0 neu, siehe z.B. Grötschel (1991, Kapitel 8). Ziel ist es, so numerische Probleme aufgrund von kumulierten Rundungsfehlern zu vermeiden, wie z.B. dass H_{k+1} zu nahe an einer singulären Matrix ist. Korollar 3.68 gilt dann entsprechend.

[31] Statt der durchschnittlichen Hesse-Matrix W kann man hierbei auch jede andere positiv definite Matrix W verwenden, die die Sekantengleichung $W p_k = r_k$ erfüllt.

Als Erstes geben wir die Eigenschaften des BFGS-Verfahrens und des DFP-Verfahrens für positiv definite quadratische Formen an. Man erreicht hier superlineare Konvergenz bei Verfahren mit inexakter Liniensuche.

Lemma 3.70
Sei $a \in \mathbb{R}^p$, $Q \in \mathbb{R}^{p\times p}$ symmetrisch und positiv definit und

$$G_{Q,a} : \mathbb{R}^p \to \mathbb{R},\ G_{Q,a}(\vartheta) = \frac{1}{2}\,\vartheta^{\mathrm{T}} Q\vartheta - a^{\mathrm{T}}\vartheta.$$

Dann gilt:

a) Das BFGS-Verfahren mit exakter Liniensuche[32] findet das Minimum $\vartheta^* = Q^{-1}a$ von $G_{Q,a}$ nach $k_0 \le p+1$ Schritten und es gilt $H_{k_0-1} = Q^{-1}$.
Analog findet das DFP-Verfahren mit exakter Liniensuche das Minimum nach $l_0 \le p+1$ Schritten und es gilt $B_{l_0-1} = Q$.

b) Das BFGS-Verfahren ohne Liniensuche und das DFP-Verfahren ohne Liniensuche sind zur Minimierung von $G_{Q,a}$ global konvergent und konvergieren superlinear, d.h. für alle Startpunkte ϑ_0 gilt

$$\limsup_{k\to\infty} \frac{\|\vartheta_{k+1} - \vartheta^*\|}{\|\vartheta_k - \vartheta^*\|} = 0.$$

Beweis: siehe Bertsekas (1999, Proposition 1.7.2) □

Wir formulieren abschließend einen Satz zur globalen Konvergenz und zur lokalen Konvergenzordnung nach Powell (1976). Unter Konvexitätsannahmen kann man hier auch superlineare Konvergenz zeigen.

Satz 3.71
Sei G zweimal stetig differenzierbar,

$$U_0 = \{\vartheta \in \mathbb{R}^p : G(\vartheta) \le G(\vartheta_0)\}$$

konvex und Hess G *positiv definit auf U_0, d.h. G ist streng konvex auf U_0. Dann konvergiert das BFGS-Verfahren mit Liniensuche, also Algorithmus 3.65, gegen das eindeutige Minimum auf U_0, wenn die Liniensuche die Wolfe-Bedingungen mit $\alpha < 1/2$ erfüllt und dabei $\lambda_k = 1$ gewählt wird, falls dies zulässig ist, also z.B. ein Backtracking-Algorithmus zum Einsatz kommt.*
Ist zudem Hess G *Lipschitz-stetig in einer Umgebung des Minimums, dann konvergiert das Verfahren superlinear und in der Nähe des Minimums ist $\lambda_k = 1$ stets zulässig.*

Beweis: siehe Powell (1976) oder Nocedal u. Wright (1999, Section 8.4) und Satz 3.38 □

[32] Wir nehmen also an, dass das Liniensuchproblem in Schritt 2 von Algorithmus 3.65 exakt gelöst wird und man sich nicht wie bei der inexakten Liniensuche z.B. lediglich mit den Wolfe-Bedingungen zufrieden gibt.

Bemerkung 3.72

a) Beim DFP-Verfahren kann man für konvexe Funktionen aus Satz 3.71 nur bei exakter Liniensuche Konvergenz zeigen, siehe Powell (1971, 1972). Bei inexakter Liniensuche basierend auf den Wolfe-Bedingungen sind keine Konvergenzaussagen bekannt.

b) Ist die Funktion G nicht konvex, so gibt Dai (2002) ein Beispiel an, bei dem das BFGS-Verfahren mit inexakter Liniensuche unter Einhaltung der Wolfe-Bedingungen nicht konvergiert. Für das BFGS-Verfahren mit exakter Liniensuche gibt Mascarenhas (2004) ein entsprechendes Beispiel an.

c) Um globale Konvergenz mit Hilfe und im Sinne von Korollar 3.25 für ein Verfahren mit inexakter Liniensuche unter Einhaltung der Wolfe-Bedingungen zeigen zu können, verändern Li u. Fukushima (2001) zu einer gegebenen streng monotonen Funktion $\phi : [0, \infty) \to [0, \infty)$ mit $\phi(0) = 0$ (etwa $\phi(x) = \delta x^\beta$ für festes $\delta, \beta > 0$) die Formel für H_{k+1} in (3.56) zu

$$H_{k+1} = \begin{cases} H_k, & \text{falls } \frac{r_k^{\mathrm{T}} p_k}{\|p_k\|^2} \geq \phi(\|g_k\|), \\ \text{gemäß (3.56)}, & \text{sonst.} \end{cases}$$

d) Um den nötigen Speicherbedarf bei hochdimensionalen Problemen zu begrenzen, ist die Modifikation L-BFGS entwickelt worden, die auch in vielen Software-Paketen zum Einsatz kommt, etwa bei der `optim`-Funktion in R, Version 2.9.0, oder den Minimierungsfunktionen in Mathematica[33], Version 6. Hierbei wird zur Berechnung der Matrix H_k nicht die Matrix H_{k-1} gespeichert, sondern für festes $m \in \mathbb{N}$, $m < p$, jeweils nur $\min(m-1, k-1)$ Vektoren-Paare (p_i, r_i), $i = \max(0, k-m), \ldots, k-2$, aus den letzten $\min(m-1, k-1)$ Rang-2-Updates für H_i gespeichert, (p_{k-1}, r_{k-1}) neu berechnet und damit H_k approximiert.[34] Mit

$$V_k = I_p - \lambda_k r_k p_k^{\mathrm{T}}$$

und einer Anfangsapproximation H_k^0 für die inverse Hesse-Matrix, etwa

$$H_k^0 = \frac{p_{k-1}^{\mathrm{T}} r_{k-1}}{r_{k-1}^{\mathrm{T}} r_{k-1}} I_p$$

setzt man (für $k \geq m$, ansonsten entsprechend verkürzt)

$$\begin{aligned} H_k = &(V_{k-1}^{\mathrm{T}} \cdots V_{k-m}^{\mathrm{T}}) H_k^0 (V_{k-m} \cdots V_{k-1}) \\ &+ \lambda_{k-m} (V_{k-1}^{\mathrm{T}} \cdots V_{k-m+1}^{\mathrm{T}}) p_{k-m} p_{k-m}^{\mathrm{T}} (V_{k-m+1} \cdots V_{k-1}) \\ &+ \lambda_{k-m+1} (V_{k-1}^{\mathrm{T}} \cdots V_{k-m+2}^{\mathrm{T}}) p_{k-m+1} p_{k-m+1}^{\mathrm{T}} (V_{k-m+2} \cdots V_{k-1}) \\ &+ \cdots \\ &+ \lambda_{k-1} p_{k-1} p_{k-1}^{\mathrm{T}}. \end{aligned}$$

[33] Die Standardvorgabe bei Quasi-Newton-Verfahren in Mathematica ist, für $p \leq 250$ das BFGS-Verfahren und für $p > 250$ das L-BFGS-Verfahren zu benutzen.

[34] Für $k \leq m$ stimmen die Matrizen bei gleicher Wahl der Anfangsapproximation noch mit dem BFGS-Verfahren überein.

Für weitere Details, wie der effizienten Berechnung der Suchrichtung $d_k = -H_k g_k$, ohne die Matrix H_k explizit bestimmen zu müssen, siehe Nocedal u. Wright (1999, Chapter 9).

Wir behandeln im Folgenden noch einige Verfahren, die speziell für Kleinste-Quadrate-Probleme entwickelt wurden.

Eine Synthese aus Gauß-Newton-Verfahren und Quasi-Newton-Verfahren wird im Algorithmus NL2SOL verwendet, der z.B. als eine der möglichen Methoden der `nls`-Funktion in R, Version 2.9.0, wählbar ist und von Dennis u. a. (1981) entwickelt wurde. Die Idee besteht darin, in der Summe

$$\operatorname{Hess} S^{\eta,y}(\vartheta_{k+1}) = 2J_\eta^{\mathrm{T}}(\vartheta_{k+1})\Sigma^{-1}J_\eta(\vartheta_{k+1}) + 2\sum_{j=1}^{N}\left(\Sigma^{-1}\left(\eta(\vartheta_{k+1}) - y\right)\right)_j \operatorname{Hess}\eta_j(\vartheta_{k+1})$$

den zweiten Summanden anders als beim Gauß-Newton-Verfahren nicht zu ignorieren, sondern mit einer ähnlichen Idee wie in Bemerkung 3.66 durch eine Matrix B_{k+1} zu approximieren.
Mit den Bezeichnungen

$$\begin{aligned} p_k &= \vartheta_{k+1} - \vartheta_k, \\ g_k &= 2J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}\left(\eta(\vartheta_k) - y\right), \\ r_k &= g_{k+1} - g_k, \\ y_k &= 2(J_\eta^{\mathrm{T}}(\vartheta_{k+1}) - J_\eta^{\mathrm{T}}(\vartheta_k))\Sigma^{-1}(\eta(\vartheta_{k+1}) - y) \end{aligned}$$

und der Sekantengleichung

$$B_{k+1}p_k = y_k$$

ergibt sich

$$B_{k+1} = B_k + \frac{1}{r_k^{\mathrm{T}}p_k}\left((y_k - B_kp_k)r_k^{\mathrm{T}} + r_k(y_k - B_kp_k)^{\mathrm{T}}\right) - \frac{(y_k - B_kp_k)^{\mathrm{T}}p_k}{(r_k^{\mathrm{T}}p_k)^2}\, r_kr_k^{\mathrm{T}}.$$

Die zugehörige neue Suchrichtung d_{k+1} wird dann aus dem Gleichungssystem

$$\left(2J_\eta^{\mathrm{T}}(\vartheta_{k+1})\Sigma^{-1}J_\eta(\vartheta_{k+1}) + B_{k+1}\right)d_{k+1} = -g_{k+1}$$

bestimmt. Als Startapproximation wird $B_0 = \mathbf{0}_{p\times p}$ verwendet. Die Wahl von B_{k+1} kann man, ähnlich wie beim DFP-Verfahren, als Lösung des Minimierungsproblem

$$B_{k+1} = \operatorname*{arg\,min}_{\substack{B\in\mathbb{R}^{p\times p} \text{ mit} \\ B=B^{\mathrm{T}} \text{ und } Bp_k=y_k}} \|B - B_k\|_W$$

für eine positiv definite Matrix W mit $Wp_k = r_k$ motivieren.
Neben dieser Grundidee verwendet NL2SOL unter anderem einen Trust-Region-Ansatz und eine Skalierung der Matrix B_k, die große Änderungen in den Residuen besser widerspiegeln soll. Hierzu und für weitere Details siehe Dennis u. a. (1981) und Dennis u. Schnabel (1983, Section 10.3).

Bemerkung 3.73
Ein weiterer Weg der Kombination von Gauß-Newton-Verfahren und BFGS-Methode wird bei bestimmten hybriden Verfahren benutzt, von denen das bekannteste von Fletcher u. Xu (1987) vorgeschlagen wurde.[35] Hierbei wird die Suchrichtung während des Programmdurchlaufs je nach vermutetem Problemtyp vom Algorithmus gewählt; im konkreten Fall wird die Gauß-Newton-Suchrichtung verwendet, wenn ein Kriterium erfüllt ist, das auf ein Problem mit kleinen Residuen deutet, ansonsten die Suchrichtung aus dem BFGS-Verfahren.

Für einige weitere Verfahren, die speziell für Kleinste-Quadrate-Probleme vorgeschlagen wurden, verweisen wir auf Seber u. Wild (1989, Section 14.2 und Section 14.3).

Wir widmen uns abschließend noch der Frage nach geeigneten Abbruchkriterien.

3.2.4. Abbruchkriterien

Ein naheliegendes Abbruchkriterium nach der Berechnung von ϑ_{k+1} im $(k+1)$-ten Schritt ist

$$\|\vartheta_{k+1} - \vartheta_k\| \leq \delta$$

für ein vorher festgelegtes $\delta > 0$. Mögliche Nachteile dieser Wahl sind jedoch, dass zum einen die einzelnen Komponenten des Parametervektors unterschiedlich skaliert sind können[36] und zum anderen, dass es unklar ist, wie δ zu wählen ist; wären alle Parameterkomponenten ungefähr von der gleichen Größenordnung, etwa

$$\big[\min_{i=1}^{p} |\vartheta_i|, \max_{i=1}^{p} |\vartheta_i|\big] \subseteq [u^s, u^{s+1}]$$

für moderates u, etwa $u = 10$, und hat man t gültige Stellen im Dezimalsystem, so wäre $\delta = u^s \cdot 10^{-t+1}$ eine nicht unvernünftige Wahl. Es bleibt jedoch die Kritik, dass man so hauptsächlich misst, ob keine wesentlichen Veränderungen mehr in den Parameterwerten geschehen und nicht, ob man sich in unmittelbarer Nähe eines (lokalen) Minimums befindet.
Nur bedingt geeignet wäre jedoch

$$\|\operatorname{grad} G(\vartheta_{k+1})\| \leq \delta$$

für ein $\delta > 0$, da diese Bedingung von der Skalierung von G und von ϑ abhängig ist und man genaue Informationen über geeignete Größenordnungen von $\operatorname{grad} G(\vartheta)$ meist nicht hat.
Eine verbreitete Alternative ist die Bedingung

$$\big\|\operatorname{grad} G(\vartheta_{k+1})\operatorname{Hess}^{-1} G(\vartheta_{k+1})\operatorname{grad} G(\vartheta_{k+1})\big\| \leq \delta.$$

[35] Der NL2SOL-Algorithmus verfolgt zusätzlich auch noch ähnliche hybride Strategien.
[36] Die annähernd gleiche Skalierung der Parameterkomponenten wäre jedoch auch für die Algorithmen selbst vorteilhaft, siehe Dennis u. Schnabel (1983, Section 7.1).

Zwar ist dieses Kriterium unabhängig unter linearen Transformationen des Parameterraums, aber abgesehen davon, dass man meist die Bestimmung von $\operatorname{Hess} G(\vartheta_{k+1})$ oder gar der Inversen hiervon vermeiden will, hat es den Nachteil, immer noch von der Skalierung von G abhängig zu sein.
Bezeichnet $\operatorname{typ}(G)$ die vom Benutzer spezifizierte typische Größe von G und $\operatorname{typ}(\vartheta_i)$ die typische Größe von ϑ_i, so empfehlen Dennis u. Schnabel (1983, Section 7.2), dass eine der beiden Bedingungen

$$\max_{i=1}^{p} \frac{|(\operatorname{grad} G(\vartheta_{k+1}))_i| \cdot \max(|(\vartheta_{k+1})_i|, \operatorname{typ}(\vartheta_i))}{\max(|G(\vartheta_{k+1})|, \operatorname{typ}(G))} \leq \delta_1$$

und

$$\max_{i=1}^{p} \frac{|(\vartheta_{k+1})_i - (\vartheta_k)_i|}{\max(|(\vartheta_{k+1})_i|, \operatorname{typ}(\vartheta_i))} \leq \delta_2$$

für einen Abbruch der Iteration erfüllt sein muss, und schlagen $\delta_1 = (\epsilon_{\text{Mach}})^{1/3}$ und $\delta_2 = 10^{-t}$ vor, wenn t signifikante Stellen für die Komponenten von ϑ erwünscht sind.

Speziell für Kleinste-Quadrate-Probleme[37] schlägt Dennis (1977) (im Fall von $\Sigma = I_N$) ein Kriterium vor, das sich an den Normalgleichungen (2.19) orientiert und das misst, welchen Winkel ϕ das Residuum $y - \eta(\vartheta_{k+1})$ zur Tangentialebene $\operatorname{Bild}(J_\eta(\vartheta_{k+1}))$ hat, und bei einem Winkel nahe $\pi/2$ abbricht. Da in dem von $<,>_\Sigma$-induzierten Hilbertraum

$$\cos(\phi) = \frac{\left\|\mathcal{P}_{J_\eta(\vartheta_{k+1}),\Sigma}(\eta(\vartheta_{k+1}) - y)\right\|_\Sigma}{\|\eta(\vartheta_{k+1}) - y\|_\Sigma}$$

gilt, wäre

$$\cos^2(\phi) \leq \delta$$

oder äquivalent

$$\frac{(\eta(\vartheta_{k+1}) - y)^{\mathrm{T}} \Sigma^{-1} J_\eta(\vartheta_{k+1}) (J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} J_\eta(\vartheta_{k+1}))^{+} J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} (\eta(\vartheta_{k+1}) - y)}{(\eta(\vartheta_{k+1}) - y)^{\mathrm{T}} \Sigma^{-1} (\eta(\vartheta_{k+1}) - y)} \leq \delta$$

ein entsprechendes Kriterium, wie es auch in der SAS Software, Version 8, als sogenanntes R-Kriterium[38] mit vom Nutzer spezifizierbarem $\delta > 0$ verwendet wird. Dennis (1977) selbst empfiehlt abzubrechen, wenn die Bedingung[39]

$$\|J_\eta(\vartheta_{k+1})\|_\Sigma \leq \delta_1 \quad \text{oder} \quad \max_{i=1}^{p} \frac{| < J_\eta^i(\vartheta_{k+1}), \eta(\vartheta_{k+1}) - y >_\Sigma |}{\left\|J_\eta^i(\vartheta_{k+1})\right\|_\Sigma \|\eta(\vartheta_{k+1}) - y\|_\Sigma} \leq \delta_2$$

für gegebenes $\delta_1 \geq 0$ und $\delta_2 > 0$ erfüllt ist. Für δ_1 rät er zu einem sehr kleinen Wert, eventuell sogar $\delta_1 = 0$, da durch die erste Bedingung vor allem abgesichert werden soll,

[37] Im Fall $\boldsymbol{B}$ kann man im folgenden stets Σ durch Σ_0 ersetzen.
[38] Im in ϑ_{k+1} linearisierten Modell ist $\cos^2(\phi) = R^2_{J_\eta(\vartheta_{k+1}), y - \eta(\vartheta_{k+1})}$, das R^2-Anpassungsmaß für die Residuen $y - \eta(\vartheta_{k+1})$.
[39] Hierbei bezeichne $J_\eta^i(\vartheta^*) = \frac{\partial \eta}{\partial \vartheta_i}(\vartheta^*)$ wieder den i-ten Spaltenvektor der Jacobi-Matrix $J_\eta(\vartheta^*) \in \mathbb{R}^{N \times p}$.

dass z.B. die zweite Größe überhaupt in den verwendeten Gleitkommazahlen berechnet werden kann. Bei 64-bit-Gleitkommazahlen („double precision“) rät er zu $\delta_2 = 10^{-6}$.

Bates u. Watts (1981, 1988) motivieren, basierend auf der Betrachtung von Konfidenzintervallen, siehe z.B. (2.22), die Wahl

$$\frac{\left\|P_{J_\eta(\vartheta_{k+1}),\Sigma}(y - \eta(\vartheta_{k+1}))\right\|_\Sigma^2}{\left\|(I_N - P_{J_\eta(\vartheta_{k+1}),\Sigma})(y - \eta(\vartheta_{k+1}))\right\|_\Sigma^2} \leq \frac{N-p}{p}\ \delta$$

mit $\delta = 10^{-6}$. Diese Bedingung wird bei der `nls`-Funktion in R, Version 2.9.0, verwendet.

Weitere Abbruchkriterien bei Kleinste-Quadrate-Problemen werden in Dennis u. a. (1981, Section 6), in Bates u. Watts (1988, Section 2.2.3) und in Seber u. Wild (1989, Section 14.4) diskutiert. Einen guten Einblick in eine praktische Implementierung liefert Gay (1990, Section 7).

Nicht behandelt haben wir in diesem Kapitel die Situation, dass $\Theta \neq \mathbb{R}^p$ gilt. Dies kann man in manchen Fällen durch eine Parametertransformation verhindern, z.B. wird die Bedingung $\vartheta_1 > 0$ durch die Parametertransformation $\tilde{\vartheta}_1 = \ln(\vartheta_1)$ automatisch erfüllt. Üblicher sind aber zum einen Penalty-Verfahren, bei denen die Zielfunktion abgeändert wird, so dass Parameterwerte außerhalb von Θ bestraft werden, und Barriere-Verfahren, bei denen Parameterwerte zu nahe am Rand von Θ bestraft werden, oder zum anderen Projektionsverfahren, bei denen die Suchrichtung entsprechend abgeändert wird und die Suchweite beschränkt wird, so dass man innerhalb von Θ bleibt. Wir verweisen hierzu auf die Literatur, z.B. Gill u. a. (1989, Chapter 5 und 6), Nocedal u. Wright (1999, Chapter 12) und Bertsekas (1999, Chapter 2 und 4).

Ferner finden die in den vergangenen Abschnitten behandelten Methoden für das nichtlineare Kleinste-Quadrate-Problem nur lokale Minima. Für Strategien zur Suche nach dem globalen Minimum, bei denen man verhindern will, dass man eventuell nur in einem lokalen Minimum landet, siehe etwa Törn u. Žilinskas (1989), Zhigljavsky u. Žilinskas (2008) und den Übersichtsartikel Neumaier (2004). Eine verwandte Fragestellung ist die Wahl geeigneter Startwerte ϑ_0 im nichtlinearen Problem. Hinweise hierzu liefern Ratkowsky (1983, Chapter 8), Bates u. Watts (1988, Section 3.3) und Seber u. Wild (1989, Section 15.2.1).

4. Schätzung bei a-priori-Information: Theorie

4.1. Bayes-Risiko und a-posteriori-Maximierung

Wir betrachten zuerst ein etwas allgemeineres Modell als das, was wir im Abschnitt 1.2 vorgestellt haben, nämlich

$$P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} = P_\vartheta$$

und

$$P^{\boldsymbol{\theta}} = Q$$

für eine Familie $\{P_\vartheta : \vartheta \in \Theta\}$ von W-Verteilungen auf[1] $(\mathbb{R}^N, \mathcal{B}^N)$, für einen Borelschen Parameterraum $\Theta \in \mathcal{B}^p$ und für eine a-priori-Verteilung Q auf $(\mathbb{R}^p, \mathcal{B}^p)$ mit Träger $\mathrm{supp}(Q) = \Theta$.
Speziell interessieren wir uns später für den Fall

$$P_\vartheta = \mathcal{N}(\eta(\vartheta), \Sigma), \tag{4.1}$$
$$Q = \mathcal{N}(\mu, B) \tag{4.2}$$

mit bekannter positiv definiter Kovarianzmatrix $\Sigma \in \mathbb{R}^{N\times N}$, bekanntem Parametervektor $\mu \in \mathbb{R}^p$ und bekannter positiv definiter Kovarianzmatrix $B \in \mathbb{R}^{p\times p}$ und für eine messbare Modellfunktion $\eta : (\mathbb{R}^p, \mathcal{B}^p) \to (\mathbb{R}^N, \mathcal{B}^N)$.

Definition 4.1
Die bedingte Verteilung

$$Q_y = P^{\boldsymbol{\theta}|\boldsymbol{Y}=y},$$

heißt **a-posteriori-Verteilung** für $\boldsymbol{\theta}$ zur Beobachtung y.

Unter den gegebenen Bedingungen existiert Q_y für $P^{\boldsymbol{Y}}$-fast alle $y \in \mathbb{R}^N$, siehe z.B. Witting (1985, Satz 1.122).

Sei im Folgenden $f_\vartheta : \mathbb{R}^N \to [0, \infty)$ eine Lebesgue-Dichte von P_ϑ für $\vartheta \in \Theta$, die Abbildung $(\vartheta, y) \mapsto f_\vartheta(y)$ sei messbar und sei q eine ν-Dichte von Q für ein σ-endliches Maß ν auf $(\mathbb{R}^p, \mathcal{B}^p)$.

[1] Mit $\mathcal{B}^N$ bezeichnen wir die Borelsche σ-Algebra im $\mathbb{R}^N$ und für $\Theta \in \mathcal{B}^p$ bezeichnen wir mit $\mathcal{B}^p_\Theta$ die Borelschen Teilmengen von Θ, d.h. $\mathcal{B}^p_\Theta = \{B \in \mathcal{B}^p : B \subseteq \Theta\}$.

Lemma 4.2

a)

$$h : \mathbb{R}^N \times \mathbb{R}^p \to \mathbb{R},\ h(y,\vartheta) = f_\vartheta(y) \cdot q(\vartheta)$$

ist eine $\boldsymbol{\lambda}^N \otimes \nu$*-Dichte von* $P^{(\boldsymbol{Y},\boldsymbol{\theta})}$*.*

b) Bezeichne für $y \in \mathbb{R}^N$

$$\bar{h}_Q(y) = \int_\Theta f_\vartheta(y) \cdot q(\vartheta)\,\mathrm{d}\nu(\vartheta) \in [0,\infty]$$

und

$$M_Q = \{y \in \mathbb{R}^N : 0 < \bar{h}_Q(y) < \infty\}.$$

Dann ist $\bar{h}_Q$ *eine Lebesgue-Dichte von* $P^{\boldsymbol{Y}}$ *und* $P(\boldsymbol{Y} \in M_Q) = 1$ *und setzt man für* $y \in M_Q$

$$q_y : \mathbb{R}^p \to \mathbb{R},\ q_y(\vartheta) = \frac{h(y,\vartheta)}{\bar{h}_Q(y)}$$

(und wählt man für $y \in \mathbb{R}^N \setminus M_Q$ *irgendeine* ν*-Dichte), dann ist* q_y *(*$P^{\boldsymbol{Y}}$*-fast sicher) eine* ν*-Dichte der a-posteriori-Verteilung* Q_y*.*

Beweis: siehe Witting (1985, Hilfssatz 2.137) □

Wir wollen $\vartheta \in \Theta$ oder allgemeiner $\gamma(\vartheta)$ für eine messbare Funktion

$$\gamma : (\Theta, \mathcal{B}^p_\Theta) \to (\Gamma, \mathcal{B}^s_\Gamma)$$

mit $\Gamma \in \mathcal{B}^s$ schätzen.

Definition 4.3
Zur Beurteilung der Güte von Schätzern sei eine **Verlustfunktion**

$$L : \Gamma \times \Theta \to \overline{\mathbb{R}}$$

gegeben, die nach oben oder nach unten beschränkt sei und die für jeden Schätzer $\delta : (\mathbb{R}^N, \mathcal{B}^N) \to (\Gamma, \mathcal{B}^s_\Gamma)$ erfüllt, dass $L(\delta(\cdot), \vartheta)$ Borel-messbar für jedes $\vartheta \in \Theta$ sei.

Beispiel 4.4
Beispiele für Verlustfunktionen, die im Folgenden eine Rolle spielen, sind

$$L^{(1)} : \Gamma \times \Theta \to \overline{\mathbb{R}},\ L^{(1)}(z,\vartheta) = \sum_{i=1}^{s} |z_i - \gamma_i(\vartheta)| \quad \text{(absoluter Verlust)},$$

$$L^{(2)} : \Gamma \times \Theta \to \overline{\mathbb{R}},\ L^{(2)}(z,\vartheta) = \|z - \gamma(\vartheta)\|^2 \quad \text{(quadratischer Verlust)}$$

oder für $\alpha > 0$

$$L^\alpha : \Gamma \times \Theta \to \overline{\mathbb{R}},\ L^\alpha(z,\vartheta) = \begin{cases} -\frac{1}{(2\alpha)^s}, & \text{falls } \max\limits_{i=1}^{s} |z_i - \gamma_i(\vartheta)| \le \alpha, \\ 0, & \text{sonst.} \end{cases}$$

Definition 4.5
Mit den obigen Voraussetzungen und Bezeichnungen heißt für $\delta : (\mathbb{R}^N, \mathcal{B}^N) \to (\Gamma, \mathcal{B}^s_\Gamma)$ die Funktion

$$\overline{L}_\delta : \Theta \to \overline{\mathbb{R}},\ \overline{L}_\delta(\vartheta) = \mathrm{E}(L(\delta(\cdot), \vartheta)|\boldsymbol{\theta} = \vartheta) = \int_{\mathbb{R}^N} L(\delta(y), \vartheta)) f_\vartheta(y)\, \mathrm{d}\boldsymbol{\lambda}^N(y)$$

die **Risiko-Funktion** von δ bezüglich der Verlustfunktion L.
Ist ferner noch

$$\overline{L}_\delta : (\Theta, \mathcal{B}^p_\Theta) \to (\overline{\mathbb{R}}, \overline{\mathcal{B}^1}),$$

dann heißt

$$\overline{L}_{\delta,Q} = \mathrm{E}_Q(\overline{L}_\delta) = \int_\Theta \overline{L}_\delta(\vartheta)\, \mathrm{d}Q(\vartheta) = \int_\Theta \overline{L}_\delta(\vartheta) q(\vartheta)\, \mathrm{d}\nu(\vartheta) \in \overline{\mathbb{R}}$$

das **Bayes-Risiko** von δ bezüglich der Verlustfunktion L und der a-priori-Verteilung Q.

Definition 4.6
Sei Δ eine Menge von Schätzern für $\gamma(\vartheta)$. Dann heißt δ^* **Bayes-Schätzer** in Δ bezüglich der Verlustfunktion L und der a-priori-Verteilung Q, falls

$$\overline{L}_{\delta^*,Q} = \min_{\delta \in \Delta} \overline{L}_{\delta,Q}.$$

Sei im Folgenden die Verlustfunktion sogar produkt-messbar, d.h. es gilt

$$L : (\Gamma \times \Theta, \mathcal{B}^s_\Gamma \otimes \mathcal{B}^p_\Theta) \to (\overline{\mathbb{R}}, \overline{B^1}). \tag{4.3}$$

Die Verlustfunktionen aus Beispiel 4.4 sind alle produkt-messbar. Die Produkt-Messbarkeit (4.3) impliziert die bisherigen Messbarkeitsvoraussetzungen in Definition 4.3 und 4.5 und es gilt

$$\begin{aligned}
\overline{L}_{\delta,Q} &= \int_\Theta \overline{L}_\delta\, \mathrm{d}Q(\vartheta) = \int_\Theta \int_{\mathbb{R}^N} L(\delta(y), \vartheta)) f_\vartheta(y)\, \mathrm{d}\boldsymbol{\lambda}^N(y)\, \mathrm{d}Q(\vartheta) \\
&= \int_{\mathbb{R}^N} \left(\int_\Theta L(\delta(y), \vartheta)) f_\vartheta(y)\, \mathrm{d}Q(\vartheta) \right) \mathrm{d}\boldsymbol{\lambda}^N(y) \\
&= \int_{\mathbb{R}^N} \left(\int_\Theta L(\delta(y), \vartheta))\, \mathrm{d}Q_y(\vartheta) \right) \bar{h}_Q(y)\, \mathrm{d}\boldsymbol{\lambda}^N(y).
\end{aligned} \tag{4.4}$$

Für den inneren Integranden des letzten Ausdrucks in (4.4) führt man folgende Bezeichnung ein.

Definition 4.7
Zur Beobachtung $y \in M_Q$ heißt

$$L_{z,Q_y} = \mathrm{E}_{Q_y}(L(z, \cdot)) = \int_\Theta L(z, \vartheta)\, \mathrm{d}Q_y(\vartheta)$$

das **a-posteriori-Risiko** der Schätzung z zur Verlustfunktion L und zur a-posteriori-Verteilung Q_y.

Es gilt also mit (4.4) und Definition 4.7, dass

$$\overline{L}_{\delta,Q} = \int_{\mathbb{R}^N} L_{\delta(y),Q_y} \bar{h}_Q(y)\,\mathrm{d}\boldsymbol{\lambda}^N(y). \tag{4.5}$$

Das nachfolgende Lemma, das aus der Darstellung (4.5) folgt, besagt im Wesentlichen, dass man in der Klasse der Schätzer den Bayes-Schätzer punktweise über die Schätzung mit dem minimalen a-posteriori-Risiko erhält.

Lemma 4.8
Sei $L : (\Gamma \times \Theta, \mathcal{B}^s_\Gamma \otimes \mathcal{B}^p_\Theta) \to (\overline{\mathbb{R}}, \overline{B^1})$ *messbar. Ist* $\delta^* : (\mathbb{R}^N, \mathcal{B}^N) \to (\Gamma, \mathcal{B}^s_\Gamma)$ *und gilt für* $P^{\boldsymbol{Y}}$*-fast alle* $y \in M_Q$*, dass*

$$L_{\delta^*(y),Q_y} = \min_{z\in\Gamma} L_{z,Q_y}, \tag{4.6}$$

dann ist δ^* *Bayes-Schätzer in der Klasse*

$$\Delta = \{\delta : (\mathbb{R}^N, \mathcal{B}^N) \to (\Gamma, \mathcal{B}^s_\Gamma)\}$$

aller Schätzer.

Beweis: siehe Witting (1985, Satz 2.138a) □

Satz 4.9
Sei $\Gamma = \mathbb{R}^s$.

a) Gilt $\mathrm{E}_{Q_y}(\|\gamma\|^2) < \infty$ *für* $P^{\boldsymbol{Y}}$*-fast alle* $y \in M_Q$*, dann ist der Bayes-Schätzer* δ^* *bezüglich quadratischem Verlust* $L^{(2)}$ *in der Klasse der messbaren Schätzer* $P^{\boldsymbol{Y}}$*-fast sicher durch*

$$\delta^*(y) = \mathrm{E}_{Q_y}(\gamma)$$

gegeben.

b) Sei $\mathrm{E}_{Q_y}(\|\gamma\|) < \infty$ *für* $P^{\boldsymbol{Y}}$*-fast alle* $y \in M_Q$ *und* $\delta^* : (\mathbb{R}^N, \mathcal{B}^N) \to (\Gamma, \mathcal{B}^s_\Gamma)$*. Gilt für* $P^{\boldsymbol{Y}}$*-fast alle* $y \in M_Q$ *und für* $i = 1, \ldots, s$*, dass* $\delta_i^*(y)$ *ein Median von* ${Q_y}^{\gamma_i}$ *ist, dem Bildmaß der a-posteriori-Verteilung* Q_y *unter* γ_i*, dann ist* δ^* *Bayes-Schätzer bezüglich absolutem Verlust* $L^{(1)}$ *in der Klasse der messbaren Schätzer.*

Beweis:

a) Es gilt

$$L_{z,Q_y} = \sum_{i=1}^{s} \int_{\mathbb{R}^s} (z_i - \gamma_i(\vartheta))^2 \,\mathrm{d}Q_y(\vartheta) = \sum_{i=1}^{s} \int_{\mathbb{R}} (z_i - u_i)^2 \,\mathrm{d}Q_y^{\gamma_i}(u_i).$$

Das Minimum in (4.6) wird daher angenommen, falls $z_i = \mathrm{E}[Q_y^{\gamma_i}] = \mathrm{E}_{Q_y}(\gamma_i)$ für $i = 1, \ldots, s$. Zur Messbarkeit siehe Witting (1985, Satz 1.126).

b) Es gilt

$$L_{z,Q_y} = \sum_{i=1}^{s} \int_{\mathbb{R}^s} |z_i - \gamma_i(\vartheta)| \, \mathrm{d}Q_y(\vartheta) = \sum_{i=1}^{s} \int_{\mathbb{R}} |z_i - u_i| \, \mathrm{d}Q_y^{\gamma_i}(u_i).$$

Das Minimum in (4.6) wird daher angenommen, falls z_i ein Median von $Q_y^{\gamma_i}$ für $i = 1, \ldots, s$ ist.

□

Bemerkung 4.10
Will man $\vartheta \in \Theta$ schätzen und ist $\Theta \neq \mathbb{R}^p$ nicht konvex, so kann man Satz 4.9 mit der Injektion $\gamma : \Theta \to \mathbb{R}^p$, $\gamma(\vartheta) = \vartheta$ anwenden, hat dann aber das Problem, dass der Erwartungswert und der Median eventuell nicht in Θ liegen. Beschränkt man sich auf die Klasse

$$\Delta = \{\delta : (\mathbb{R}^N, \mathcal{B}^N) \to (\Theta, \mathcal{B}_\Theta^p)\}$$

der messbaren Schätzer mit Schätzwerten in Θ, so hat man für die Verlustfunktion $L^{(1)}$ das Minimierungsproblem

$$\arg\min_{z \in \Theta} \sum_{i=1}^{s} \int_{\mathbb{R}^s} |z_i - \gamma(\vartheta)| \, \mathrm{d}Q_y(\vartheta)$$

bzw. für die Verlustfunktion $L^{(2)}$

$$\arg\min_{z \in \Theta} \int_{\mathbb{R}^s} \|z - \gamma(\vartheta)\|^2 \, \mathrm{d}Q_y(\vartheta) \tag{4.7}$$

zu lösen. Einige theoretische Überlegungen für Probleme der Form (4.7) für äußerst spezielle Mannigfaltigkeiten Θ finden sich in Hendriks u. Landsman (1998).

Für unsere Betrachtungen zu den Verlustfunktionen L^α, $\alpha > 0$, benötigen wir den folgenden Satz. Hierbei bezeichnen wir für $x \in \mathbb{R}^k$ und $r > 0$ mit $B_r(x) = \{y \in \mathbb{R}^k : \|x - y\| < r\}$ den offenen Ball vom Radius r um x.

Satz 4.11
Für alle $x \in \mathbb{R}^k$ sei $(E_i(x))_{i \in \mathbb{N}}$ eine Folge von Borelmengen im $\mathbb{R}^k$. Ferner gebe es für alle $x \in \mathbb{R}^k$ ein $a = a(x) > 0$ und eine Folge $(r_i)_{i \in \mathbb{N}}$ mit $r_i > 0$ und $\lim_{i \to \infty} r_i = 0$ und $E_i(x) \subseteq B_{r_i}(x)$ und $\boldsymbol{\lambda}^k(E_i(x)) \geq a\boldsymbol{\lambda}^k(B_{r_i}(x))$ für alle $i \in \mathbb{N}$. Ist $f : \mathbb{R}^k \to \mathbb{R}$ Lebesgue-integrierbar, dann gilt für $\boldsymbol{\lambda}^k$-fast alle $x \in \mathbb{R}^k$

$$f(x) = \lim_{i \to \infty} \frac{1}{\boldsymbol{\lambda}^k(E_i(x))} \int_{E_i(x)} f(y) \, \mathrm{d}\boldsymbol{\lambda}^k(y)$$

Beweis: siehe Rudin (1987, Theorem 7.10) □

Sei $\nu = \boldsymbol{\lambda}^p$. Wir betrachten nun die Verlustfunktionen L^α und wollen $\gamma(\vartheta) = \vartheta$ schätzen. Dann gilt für das a-posteriori-Risiko einer Schätzung z bei a-priori-Verteilung Q und Beobachtung $y \in M_Q$ mit $B^1_\alpha(z) = \{x \in \mathbb{R}^p : \max_{i=1}^p |z_i - x_i| \leq \alpha\}$

$$\begin{aligned} L^\alpha_{z,Q_y} &= \int_{\mathbb{R}^p} L^\alpha(z,\vartheta)\,\mathrm{d}Q_y(\vartheta) \\ &= -\frac{1}{(2\alpha)^p}\int_{B^1_\alpha(z)} q_y(\vartheta)\,\mathrm{d}\boldsymbol{\lambda}^p(\vartheta) = -\frac{1}{\boldsymbol{\lambda}^p(B^1_\alpha(z))}\int_{B^1_\alpha(z)} q_y(\vartheta)\,\mathrm{d}\boldsymbol{\lambda}^p(\vartheta); \end{aligned}$$

insbesondere gilt nach Satz 4.11, dass (für fast alle $z \in \mathbb{R}^p$)

$$\lim_{\alpha \searrow 0} L^\alpha_{z,Q_y} = -q_y(z). \tag{4.8}$$

Definition 4.12
Ein Schätzer

$$\delta^* : (\mathbb{R}^N, \mathcal{B}^N) \to (\Theta, \mathcal{B}^p_\Theta)$$

mit

$$q_y(\delta^*(y)) = \max_{z \in \mathbb{R}^p} q_y(z) \quad \text{für } (P^{\boldsymbol{Y}}\text{-fast alle}) \; y \in \mathbb{R}^N \tag{4.9}$$

heißt **Modalwert-Schätzer**.

Da $P^{\boldsymbol{Y}}$-fast sicher $\operatorname{supp}(Q_y) \subset \operatorname{supp}(Q) = \Theta$ gilt, folgt

$$\max_{z \in \mathbb{R}^p} q_y(z) = \max_{z \in \Theta} q_y(z),$$

d.h. man kann das Maximierungsproblem (4.9) auf Θ einschränken. Der Modalwertschätzer ist ein Analogon zum Maximum-Likelihood-Schätzer und wegen (4.8) und (4.9) gilt für kleines $\alpha > 0$, dass das a-posteriori-Risiko des Modalwert-Schätzers annähernd gleich dem a-posteriori-Risiko eines Bayes-Schätzers zu L^α ist. In diesem Sinne kann man also den Modalwert-Schätzer approximativ als Bayes-Schätzer zu einer Verlustfunktion L^α für kleines $\alpha > 0$ ansehen. Es gibt also eine große Belohnung, wenn man sehr nahe am „wahren“ Parameterwert liegt, ansonsten gibt es keine Belohnung. Bettet man den Raum der Verlustfunktionen zur Entscheidung z in den Raum der Distributionen ein, so konvergiert $L^\alpha(z,\cdot)$ gegen die negative Dirac-Distribution in z, siehe Forster (1984, Satz 1, S. 177). Aufgrund dessen kann man dann den Modalwert-Schätzer als Bayes-Schätzer bezüglich dieser „Verlustdistribution“ auffassen.

Bemerkung 4.13
Sei $g : \mathbb{R}^s \to [0,\infty)$ eine Lebesgue-integrierbare Funktion mit $\int_{\mathbb{R}^s} g(x)\,\mathrm{d}\boldsymbol{\lambda}^s(x) = 1$ und beschränktem Träger. Die Folge der Verlustfunktionen L^α kann man bei obigen Überlegungen zum Modalwert-Schätzer zumindest in den Punkten, in denen q_y stetig ist, durch die Verlustfunktionen

$$L^{g,\alpha} : \Gamma \times \Theta \to \overline{\mathbb{R}},\; L^{g,\alpha}(z,\vartheta) = -\frac{1}{\alpha^s}\, g\left((\vartheta - z)/\alpha\right)$$

ersetzen, wie man ebenfalls unter Anwendung von Forster (1984, Satz 1, S. 177) sieht.

4.2. Bayes-Schätzung im linearen Modell

Seien $B \in \mathbb{R}^{p\times p}$ und $\Sigma \in \mathbb{R}^{N\times N}$ symmetrisch und positiv definit, $X \in \mathbb{R}^{N\times p}$, $y_0 \in \mathbb{R}^N$, $\mu \in \mathbb{R}^p$ und $\vartheta_0 \in \mathbb{R}^p$ bekannt. Wir betrachten hier das affin-lineare Modell

$$\begin{aligned} P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathcal{N}(\eta(\vartheta), \Sigma), \\ P^{\boldsymbol{\theta}} &= \mathcal{N}(\mu, B) \end{aligned} \tag{4.10}$$

mit

$$\eta : \mathbb{R}^p \to \mathbb{R}^N,\ \eta(\vartheta) = y_0 + X \cdot (\vartheta - \vartheta_0), \tag{4.11}$$

das zuerst von Lindley u. Smith (1972) untersucht wurde.

Lemma 4.14
Im affin-linearen Modell (4.10) *und* (4.11) *gilt für die Verteilung von* $\boldsymbol{Y}$*, dass*

$$P^{\boldsymbol{Y}} = \mathcal{N}(X(\mu - \vartheta_0) + y_0, \Sigma + XBX^{\mathrm{T}}).$$

Bezeichne

$$\begin{aligned} b &= X^{\mathrm{T}}\Sigma^{-1}(y - y_0 + X\vartheta_0) + B^{-1}\mu, \\ C &= (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}. \end{aligned} \tag{4.12}$$

Dann ist die a-posteriori-Verteilung für $\boldsymbol{\theta}$ *zur Beobachtung* y *gegeben durch*

$$P^{\boldsymbol{\theta}|\boldsymbol{Y}=y} = \mathcal{N}(Cb, C).$$

Beweis: siehe Lindley u. Smith (1972, Section 2) □

Korollar 4.15
Im affin-linearen Modell (4.10) *und* (4.11) *ist der Schätzer*

$$\widehat{\vartheta}_{Bayes}(\boldsymbol{Y}) = (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}\left(X^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - y_0 + X\vartheta_0) + B^{-1}\mu\right) \tag{4.13}$$

zugleich der Modalwert-Schätzer und der Bayes-Schätzer bezüglich absolutem Verlust und bezüglich quadratischem Verlust.

Aufgrund der Normalgleichungen folgt im Fall $\mathrm{rg}(X) = p$

$$\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y}) = (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}\left(X^{\mathrm{T}}\Sigma^{-1}X\widehat{\vartheta}(\boldsymbol{Y}) + B^{-1}\mu\right), \tag{4.14}$$

bzw. im Fall $\mathrm{rg}(X) < p$

$$\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y}) = (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}\left(X^{\mathrm{T}}\Sigma^{-1}X\widehat{\vartheta}_{MP}(\boldsymbol{Y}) + B^{-1}\mu\right),$$

d.h. der Bayes-Schätzer $\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y})$ ist in diesem Sinne ein gewichtetes Mittel aus dem Kleinste-Quadrate-Schätzer $\widehat{\vartheta}(\boldsymbol{Y})$ und dem Erwartungswert μ der a-priori-Verteilung.

An Stelle der Formeln (4.12) und (4.13) sind einige alternative Formulierungen möglich, indem man z.B. die Sherman-Morrison-Woodbury-Formel verwendet, siehe Higham (2002, Exercise 13.9), woraus

$$C = (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1} = B - BX^{\mathrm{T}}(\Sigma + XBX^{\mathrm{T}})^{-1}XB \tag{4.15}$$

folgt oder indem man

$$(X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}X^{\mathrm{T}}\Sigma^{-1} = BX^{\mathrm{T}}(\Sigma + XBX^{\mathrm{T}})^{-1}$$

verwendet. Dies führt zu

$$\begin{aligned}\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y}) &= (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}X^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - y_0) \\ &\quad + (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}X^{\mathrm{T}}\Sigma^{-1}X\vartheta_0 + (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}B^{-1}\mu \\ &= (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}(X^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - y_0) + B^{-1}(\mu - \vartheta_0)) + \vartheta_0 \end{aligned} \tag{4.16}$$

oder zu

$$\begin{aligned}\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y}) &= (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}X^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y} - y_0 + X\vartheta_0) \\ &\quad - (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}X^{\mathrm{T}}\Sigma^{-1}X\mu + \mu \\ &= BX^{\mathrm{T}}(\Sigma + XBX^{\mathrm{T}})^{-1}(\boldsymbol{Y} - y_0 + X(\vartheta_0 - \mu)) + \mu. \end{aligned} \tag{4.17}$$

Die letztgenannte Formel (4.17) hat dabei den Vorteil, auch noch den Bayes-Schätzer im affin-linearen Modell anzugeben, falls B nur positiv semidefinit ist.

Bemerkung 4.16
Es gilt folgendes Analogon zum Satz 2.2 von Gauß-Markov, wenn man die Normalverteilungsannahmen im affin-linearen Modell zu folgenden Annahmen über die ersten beiden Momente abschwächt: Seien

$$\begin{aligned}\mathrm{E}(\boldsymbol{Y}|\boldsymbol{\theta} = \vartheta) &= \eta(\vartheta), \\ \mathrm{D}(\boldsymbol{Y}|\boldsymbol{\theta} = \vartheta) &= \Sigma,\end{aligned}$$

wobei η affin-linear gemäß (4.11) und Σ positiv definit seien, und $\boldsymbol{\theta} \sim Q$ mit

$$\begin{aligned}\mathrm{E}(\boldsymbol{\theta}) &= \mu, \\ \mathrm{D}(\boldsymbol{\theta}) &= B,\end{aligned}$$

wobei B positiv semidefinit sei. Wir wollen $\gamma(\vartheta) = L\vartheta$ für $L \in \mathbb{R}^{s\times p}$. schätzen und betrachten die Klasse

$$\Delta = \{U\boldsymbol{Y} + b : U \in \mathbb{R}^{s\times N}, b \in \mathbb{R}^s\}$$

von affin-linearen Schätzern für $L\vartheta$. Betrachtet man nun zu einem Schätzer $\delta : (\mathbb{R}^N, \mathcal{B}^N) \to (\mathbb{R}^s, \mathcal{B}^s)$ statt der Risikofunktion zum quadratischen Verlust die Streumatrix R_δ um $L\vartheta$, d.h.

$$R_\delta : \mathbb{R}^p \to \mathbb{R}^{p\times p},\ R_\delta(\vartheta) = \mathrm{E}\big((\delta(\cdot) - C\vartheta)(\delta(\cdot) - C\vartheta)^{\mathrm{T}}|\boldsymbol{\theta} = \vartheta\big),$$

und die mittlere Streumatrix

$$\overline{R}_{\delta,Q} = \int_{\mathbb{R}^p} R_\delta(\vartheta)\,\mathrm{d}Q(\vartheta),$$

dann gilt für

$$\delta^*(\boldsymbol{Y}) = C\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y}) = C(BX^{\mathrm{T}}(\Sigma + XBX^{\mathrm{T}})^{-1}(\boldsymbol{Y} - y_0 + X(\vartheta_0 - \mu)) + \mu) \in \Delta,$$

dass

$$\overline{R}_{\delta^*,Q} \leq_{\mathfrak{L}} \overline{R}_{\delta,Q} \quad \text{für alle } \delta \in \Delta;$$

insbesondere ist für jedes $d \in \mathbb{R}^s$ daher $d^{\mathrm{T}}\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y})$ Bayes-Schätzer für $d^{\mathrm{T}}\vartheta$ bezüglich quadratischen Verlusts in der Klasse der affin-linearen Schätzer und damit ist auch $\widehat{\vartheta}_{\mathrm{Bayes}}(\boldsymbol{Y})$ Bayes-Schätzer für ϑ bezüglich quadratischen Verlusts in der Klasse der affin-linearen Schätzer.

Bemerkung 4.17
Bilden die Verteilungen $P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta}$ für $\vartheta \in \Theta$ eine Exponentialfamilie und entspricht der Parameterraum Θ dem natürlichen Parameterraum und hat dieser nichtleeres Inneres, dann bildet die konjugierte Exponentialfamilie Π eine angenehme Klasse von a-priori-Verteilungen auf dem Inneren des natürlichen Parameterraums, denn für alle a-priori-Verteilungen $Q \in \Pi$ sind dann für $P^{\boldsymbol{Y}}$-fast alle Beobachtungen y die a-posteriori-Verteilungen $Q_y \in \Pi$, siehe Lehmann u. Casella (1998, S. 244 f.), Witting (1985, Aufgabe 1.43) oder Robert (2007, Section 3.3).

Da im Fall des linearen Modells (4.10) und (4.11) die konjugierte Exponentialfamilie ebenfalls von Normalverteilungen gebildet wird, macht dies die Angelegenheit so einfach.[2] Dies scheitert jedoch im nichtlinearen Modell, da der natürliche Parameterraum im Allgemeinen kein nichtleeres Inneres hat, vergleiche die Ausführungen auf Seite 10.

4.3. Der Modalwert-Schätzer im nichtlinearen Modell

Wir betrachten das Modell

$$\begin{aligned} P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathcal{N}(\eta(\vartheta), \Sigma), \\ P^{\boldsymbol{\theta}} &= Q \end{aligned} \tag{4.18}$$

für eine a-priori-Verteilung Q auf $(\mathbb{R}^p, \mathcal{B}^p)$ mit ν-Dichte q für ein σ-endliches Maß ν auf $(\mathbb{R}^p, \mathcal{B}^p)$ mit $\mathrm{supp}(Q) = \Theta \subseteq \mathbb{R}^p$ und eine Borel-messbare Modellfunktion $\eta : (\Theta, \mathcal{B}^p_\Theta) \to$

[2] Es gibt auch noch Varianten zu dem hier betrachteten linearen Modell, indem man etwa für die Kovarianzmatrix von $P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta}$ eine Zerlegung der Form $\Sigma = \sigma^2\Sigma_0$ annimmt und zusätzlich a-priori-Annahmen für den unbekannten Parameter σ^2 trifft. Dies führt zum „Normal-Gamma-Modell“, siehe Pukelsheim (1993, Chapter 11) und Rencher u. Schaalje (2008, Chapter 11), bei dem die Theorie ebenfalls entsprechend einfach ist. Die Wishart-Verteilung als a-priori-Verteilung für Σ^{-1} wird in Lindley u. Smith (1972) behandelt.

$(\mathbb{R}^N, \mathcal{B}^N)$. Wir bezeichnen mit

$$f_{\mathcal{N}(\eta(\vartheta),\Sigma)} : \mathbb{R}^N \to \mathbb{R},\ f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) = \frac{1}{\sqrt{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2} \|y - \eta(\vartheta)\|_\Sigma^2\right)$$

eine Lebesgue-Dichte von $\mathcal{N}(\eta(\vartheta), \Sigma)$. Dann gilt für die ν-Dichte q_y der a-posteriori-Verteilung Q_y nach Lemma 4.2

$$q_y(\vartheta) = \frac{1}{\bar{h}_Q(y)} f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) \cdot q(\vartheta).$$

Ist η nichtlinear oder ist q keine Normalverteilungsdichte, so ist in der Regel q_y nicht mehr Lebesgue-Dichte einer Normalverteilung. Die Bestimmung des Bayes-Schätzers bezüglich quadratischen Verlusts erfordert dann die Bestimmung des Erwartungswerts von Q_y, die Bestimmung des Bayes-Schätzers bezüglich absoluten Verlusts erfordert die Bestimmung der Mediane der Randverteilungen von Q_y. Beide Schätzprobleme müssen daher meist mit Methoden der numerischen Integration gelöst werden, auf die wir hier nicht eingehen werden und für die wir auf Robert (2007, Chapter 6) für eine Einführung aus Bayesscher Sicht, auf Dahlquist u. Björck (2008, Chapter 5) für eine Einführung in deterministische Verfahren zur Integration und auf Müller-Gronbach u. a. (2010) für stochastische Verfahren zur Integration verweisen.

Wir betrachten hier den Modalwert-Schätzer, also zu gegebener Beobachtung y das Maximierungsproblem

$$\arg\max_{\vartheta\in\Theta} q_y(\vartheta). \tag{4.19}$$

Wegen

$$\ln\left(f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) \cdot q(\vartheta)\right) = -\frac{1}{2} N \ln(2\pi) - \frac{1}{2} \ln(\det(\Sigma)) - \frac{1}{2} \|y - \eta(\vartheta)\|_\Sigma^2 + \ln(q(\vartheta))$$

ist das Maximierungsproblem (4.19) äquivalent zum Minimierungsproblem

$$\arg\min_{\vartheta\in\Theta} \|y - \eta(\vartheta)\|_\Sigma^2 - 2\ln(q(\vartheta)). \tag{4.20}$$

Gilt speziell $\Theta = \mathbb{R}^p$ und $Q = \mathcal{N}(\mu, B)$, so wird das Minimierungsproblem (4.20) zu

$$\arg\min_{\vartheta\in\mathbb{R}^p} \|y - \eta(\vartheta)\|_\Sigma^2 + \|\vartheta - \mu\|_B^2 . \tag{4.21}$$

Ist $\Theta \neq \mathbb{R}^p$ mit $\boldsymbol{\lambda}^p(\Theta) > 0$, so erscheint statt $Q = \mathcal{N}(\mu, B)$ die bedingte Verteilung $\tilde{Q} = Q(\cdot|\Theta)$ als a-priori-Verteilung plausibel. Für die Lebesgue-Dichte $f_{\tilde{Q}}$ von $\tilde{Q}$ gilt dann

$$f_{\tilde{Q}} : \mathbb{R}^p \to [0,\infty),\ f_{\tilde{Q}}(\vartheta) = \begin{cases} \frac{1}{Q(\Theta)} f_{\mathcal{N}(\mu,B)}(\vartheta), & \text{für } \vartheta \in \Theta, \\ 0, & \text{sonst.} \end{cases}$$

Damit ergibt sich für den Modalwert-Schätzer das Minimierungsproblem

$$\arg\min_{\vartheta\in\Theta} \|y - \eta(\vartheta)\|_\Sigma^2 + \|\vartheta - \mu\|_B^2 .$$

Gegenüber dem Kleinste-Quadrate-Verfahren (2.15) tritt beim Modalwert-Schätzproblem noch der „Strafterm“ $-2\ln(q(\vartheta))$ bzw. $\|\vartheta-\mu\|_B^2$ auf, der groß wird, falls die Dichte q in ϑ klein bzw. falls ϑ weit von dem a-priori-Modalwert μ entfernt ist. Außerdem erkennt man, dass der Ansatz, Nichtwissen durch eine konstante a-priori-Dichte zu modellieren, zum Least-Squares-Verfahren führt, siehe hierzu auch Bemerkung 4.19. Daher ist der Modalwert-Schätzer in der Literatur auch als „**penalized-least-Squares**-Schätzer“ bzw. **extended-weighted-least-Squares**-Schätzer bekannt, siehe auch Isenberg (1979).

Formal kann man bei einer Normalverteilung $Q=\mathcal{N}(\mu,B)$ als a-priori-Verteilung das Kleinste-Quadrate-Problem mit Strafterm (4.21) auch als Kleinste-Quadrate-Problem auffassen, da

$$\|y-\eta(\vartheta)\|_\Sigma^2+\|\vartheta-\mu\|_B^2=\left\|\begin{pmatrix}y\\ \mu\end{pmatrix}-\begin{pmatrix}\eta(\vartheta)\\ \vartheta\end{pmatrix}\right\|^2_{\left(\begin{smallmatrix}\Sigma & 0\\ 0 & B\end{smallmatrix}\right)}. \tag{4.22}$$

Es ergibt sich also das gleiche Minimierungsproblem wie bei der Kleinste-Quadrate-Schätzung in einem Modell aus Abschnitt 2.1.2 zu einer Modellfunktion $\tilde\eta$ mit $\tilde\eta(\vartheta)=\binom{\eta(\vartheta)}{\vartheta}$, zu einer „Beobachtung“ $\tilde y=\binom{y}{\mu}$ und zu einer Kovarianzmatrix $\tilde\Sigma=\left(\begin{smallmatrix}\Sigma & 0\\ 0 & B\end{smallmatrix}\right)$.

In Fragen der Existenz und Eindeutigkeit können somit direkt die Resultate aus Abschnitt 2.1.2 angewendet werden. Auch die entsprechenden Nichtlinearitätsmaße (vgl. die Referenzen in Abschnitt 2.2.2) zu übertragen bietet sich direkt an. Im nächsten Kapitel diskutieren wir die Anwendung der numerischen Verfahren aus Abschnitt 3.2 und weitere numerische Verfahren für die Minimierungsprobleme (4.20) und (4.21).

Bemerkung 4.18
Da

$$J_{\tilde\eta}(\vartheta)=\begin{pmatrix}J_\eta(\vartheta)\\ I_p\end{pmatrix},$$

gilt insbesondere $\operatorname{rg}(J_{\tilde\eta}(\vartheta))=p$ für alle $\vartheta\in\Theta$, d.h. es liegt ein reguläres Modell vor, auch wenn $\operatorname{rg}(J_{\tilde\eta}(\vartheta))<p$ ist. Dies ist auch der Grund, warum Anwender und Numeriker unter dem Stichwort „**Tichonov-Regularisierung**“ das Minimierungsproblem

$$\arg\min_{\vartheta\in\Theta}\|y-\eta(\vartheta)\|^2$$

durch das Minimierungsproblem

$$\arg\min_{\vartheta\in\Theta}\|y-\eta(\vartheta)\|^2+\|\vartheta-\mu\|_B^2$$

mit $\mu=0$, $B=\tau I_p$ und $\tau>0$ ersetzen, siehe Bickel u. Li (2006), Eriksson u. a. (2005) und Engl u. a. (1996).

Bemerkung 4.19
Um zu modellieren, dass man für bestimmte Parameterkomponenten keine Vorinformation hat, lässt man auch sogenannte **uneigentliche a-priori-Verteilungen** zu.[3]

[3] „A-priori-Maß“ wäre eine bessere Bezeichnung.

Unterteilt man für $s \in \{0, \dots, p\}$ etwa

$$\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)})$$

mit einer $(\mathbb{R}^s, \mathcal{B}^s)$-wertigen Zufallsvariablen $\boldsymbol{\theta}^{(1)}$ und einer $(\mathbb{R}^{p-s}, \mathcal{B}^{p-s})$-wertigen Zufallsvariablen $\boldsymbol{\theta}^{(2)}$, so ist mit $\mu_1 \in \mathbb{R}^s$ und positiv definiter Matrix $B_1 \in \mathbb{R}^{s\times s}$

$$\boldsymbol{\theta} \sim \mathcal{N}(\mu_1, B_1) \otimes \boldsymbol{\lambda}^{p-s}$$

ein entsprechendes a-priori-Maß, das modellieren soll, dass man nur Vorinformationen über die ersten s Komponenten des Parametervektors hat. Zu solchen a-priori-Maßen betrachtet man dann eine Folge von a-priori-Verteilungen, die zwar bezüglich der schwach-$*$-Topologie divergent ist, von der man aber hofft, dass die entsprechenden Bayes-Risiken bzw. a-posteriori-Verteilungen konvergieren, siehe Witting (1985, S. 326).[4]
In unserem Fall würde man demgemäß als Lebesgue-Dichte q_y der a-posteriori-Verteilung den Grenzwert

$$\begin{aligned} q_y(\vartheta) &= \lim_{n\to\infty} \frac{f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) f_{\mathcal{N}(\mu_1,B_1)}(\vartheta^{(1)}) \mathbf{1}_{[-n,n]^{p-s}}(\vartheta^{(2)})}{\int_{\mathbb{R}^p} f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) f_{\mathcal{N}(\mu_1,B_1)}(\vartheta^{(1)}) \mathbf{1}_{[-n,n]^{p-s}}(\vartheta^{(2)}) \,\mathrm{d}\boldsymbol{\lambda}^p(\vartheta)} \\ &= \frac{f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) f_{\mathcal{N}(\mu_1,B_1)}(\vartheta^{(1)})}{\int_{\mathbb{R}^p} f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) f_{\mathcal{N}(\mu_1,B_1)}(\vartheta^{(1)}) \,\mathrm{d}\boldsymbol{\lambda}^p(\vartheta)} \end{aligned} \tag{4.23}$$

ansehen, falls das Integral im Nenner (für fast alle $y \in \mathbb{R}^N$) endlich ist. Maximierung der a-posteriori-Dichte führt dann zu dem anschaulichen Minimierungsproblem

$$\arg\min_{\vartheta\in\Theta} \|y - \eta(\vartheta)\|_\Sigma^2 + \|\vartheta^{(1)} - \mu_1\|_{B_1}^2 .$$

Allerdings hat zum einen dieses Konzept der „Uninformiertheit“ seine Schwächen, da es nicht invariant gegen Parametertransformationen ist, siehe Robert (2007, Section 1.5, 3.5 und Chapter 9) und Berger (1985, Section 3.3), zum anderen muss das Integral im Nenner von (4.23) nicht endlich sein, wie man an dem Beispiel

$$\eta : \mathbb{R} \to \mathbb{R},\ \eta(\vartheta) = \sqrt{\ln(\sqrt{2+\vartheta^2})},$$

$\Sigma = 1$, $\boldsymbol{\theta} \sim \boldsymbol{\lambda}^1$ und beliebiges $y \in \mathbb{R}$ sieht.

[4] Mit den Bezeichnungen von Abschnitt 4.1 auf S. 99 geht man wie folgt vor: Ist ν ein σ-endliches Maß auf $(\mathbb{R}^p, \mathcal{B}^p)$ und das a-priori-Maß $Q = g \cdot \nu$ ein nicht endliches Maß auf $(\mathbb{R}^p, \mathcal{B}^p)$, so betrachtet man eine Folge $(\nu_n)_{n\in\mathbb{N}}$ von endlichen Maßen $\nu_n = g_n \cdot \nu$ mit $\lim_{n\to\infty} g_n(\vartheta) = g(\vartheta)$ für ν-fast alle $\vartheta \in \mathbb{R}^p$. Die zugehörigen a-priori-Verteilungen sind dann $Q_n = \frac{1}{\nu_n(\mathbb{R}^p)} \cdot \nu_n$ für $n \in \mathbb{N}$. Existiert

$$q_y(\vartheta) = \lim_{n\to\infty} \frac{f_\vartheta(y) g_n(\vartheta)}{\int_{\mathbb{R}^p} f_\vartheta(y) g_n(\vartheta) \,\mathrm{d}\nu(\vartheta)}$$

für fast alle $y \in \mathbb{R}^N$, so heißt $q_y \cdot \nu$ die (limitierte) a-posteriori-Verteilung.

4.4. Empirische Bayes-Schätzung

Wir betrachten jetzt Modelle, in denen in der a-priori-Verteilung unbekannte Parameter ξ und ρ auftreten.

Seien $\eta : (\mathbb{R}^p, \mathcal{B}^p) \to (\mathbb{R}^N, \mathcal{B}^N)$, $\xi \in W \subseteq \mathbb{R}^w$, $\mu : (W, \mathcal{B}_W^w) \to (\mathbb{R}^p, \mathcal{B}^p)$, $\Sigma \in \mathbb{R}^{N\times N}$ symmetrisch und positiv definit, $\rho \in V \subseteq \mathbb{R}^v$, und $B : (V, \mathcal{B}_V^v) \to (\mathbb{R}^{p\times p}, \mathcal{B}^{p\times p})$ eine Funktion mit Werten in den symmetrischen und positiv definiten Matrizen. Gegeben sei das Modell[5]

$$\begin{aligned} P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathcal{N}(\eta(\vartheta), \Sigma), \\ P^{\boldsymbol{\theta}} &= \mathcal{N}(\mu(\xi), B(\rho)). \end{aligned} \tag{4.24}$$

Das übliche Vorgehen in einem solchen Modell besteht darin, die Lebesgue-Dichte der Verteilung von $\boldsymbol{Y}$ zu bestimmen, d.h.

$$\bar{h}_Q(y) = \int_{\mathbb{R}^p} f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) \cdot f_{\mathcal{N}(\mu(\xi),B(\rho))}(\vartheta)\, \mathrm{d}\boldsymbol{\lambda}^p(\vartheta).$$

Zu gegebener Beobachtung y werden anschließend hieraus Schätzungen $\widehat{\xi}(y)$ für ξ und $\widehat{\rho}(y)$ für ρ bestimmt, wobei meist die Maximum-Likelihood-Methode verwendet wird. Anschließend ergibt sich die Schätzung für ϑ gemäß dem Verfahren bei vollständiger a-priori-Information unter Verwendung der Schätzungen $\widehat{\mu} = \mu(\widehat{\xi}(y))$ und $\widehat{B} = B(\widehat{\rho}(y))$.

Ein Problem besteht nun darin, daß im Allgemeinen die Integration in der Formel für die Dichte $\bar{h}_Q$ von $\boldsymbol{Y}$ nicht explizit lösbar ist und nicht wie im linearen Modell die Dichte einer Normalverteilung ist. Ein mögliches Vorgehen im nichtlinearen Fall wäre nun, das Integral für die Randdichte jeweils numerisch auszuwerten, z.B. mit einer Monte-Carlo-Methode. Für die von uns anvisierten Anwendungen ist dieses Verfahren allerdings sehr aufwendig, da hierbei für viele verschiedene ϑ-Werte $\eta(\vartheta)$ bestimmt werden muß, also z.B. viele strukturelle Eigenwert-Probleme gelöst werden müssen. Wir werden uns im Abschnitt 5.2 mit Alternativen beschäftigen.

Wie in Lemma 4.14 bereits festgestellt, lässt sich im affin-linearen Fall, d.h. für

$$\eta : \mathbb{R}^p \to \mathbb{R}^N,\ \eta(\vartheta) = y_0 + X \cdot (\vartheta - \vartheta_0),$$

die Verteilung von $\boldsymbol{Y}$ explizit bestimmen, und man erhält

$$P^{\boldsymbol{Y}} = \mathcal{N}(X(\mu(\xi) - \vartheta_0) + y_0, \Sigma + XB(\rho)X^{\mathrm{T}}). \tag{4.25}$$

Mit einer Konstanten $c \in \mathbb{R}$ gilt somit für die zugehörige Dichte $\bar{h}_Q$

$$\begin{aligned} -2\ln(\bar{h}_Q(y)) = {} & c + \ln\det(\Sigma + XB(\rho)X^{\mathrm{T}}) \\ & + (y - y_0 + X\vartheta_0 - X\mu(\xi))^{\mathrm{T}}(\Sigma + XB(\rho)X^{\mathrm{T}})^{-1}(y - y_0 + X\vartheta_0 - X\mu(\xi)). \end{aligned}$$

[5] Ist η nur auf $\Theta \subsetneq \mathbb{R}^p$ definiert, kann man wie im vorhergehenden Abschnitt stattdessen als a-priori-Verteilung eine bedingte Normalverteilung betrachten.

Man hat daher zu einer Beobachtung y das Minimierungsproblem

$$\operatorname*{arg\,min}_{\xi\in W,\rho\in V} \ln\det(\Sigma + XB(\rho)X^{\mathrm{T}}) + \|y - y_0 + X\vartheta_0 - X\mu(\xi)\|^2_{\Sigma+XB(\rho)X^{\mathrm{T}}}$$

zu betrachten. Für solche Minimierungsprobleme wurden in der Varianzkomponentenanalyse Verfahren entwickelt, siehe z.B. Rao u. Kleffe (1988) und Searle u. a. (1992). Hat man hiermit Schätzwerte $\widehat{\xi} = \widehat{\xi}(y)$ und $\widehat{\rho} = \widehat{\rho}(y)$ bestimmt, erhält man im affin-linearen Modell als Schätzung für ϑ somit

$$\begin{aligned}\widehat{\vartheta}_{EB}(y) &= (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1}(\widehat{\rho}))^{-1}\left(X^{\mathrm{T}}\Sigma^{-1}(y - y_0 + X\vartheta_0) + B^{-1}(\widehat{\rho})\mu(\widehat{\xi})\right) &(4.26)\\ &= \vartheta_0 + (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1}(\widehat{\rho}))^{-1}\left(X^{\mathrm{T}}\Sigma^{-1}(y - y_0) + B^{-1}(\widehat{\rho})(\mu(\widehat{\xi}) - \vartheta_0)\right). &(4.27)\end{aligned}$$

Für weitere Hintergründe zur empirischen Bayes-Schätzung verweisen wir auf Lehmann u. Casella (1998, Section 4.6), Berger (1985, Section 4.5) und Carlin u. Louis (1996, Chapter 3).

5. Schätzung bei a-priori-Information: numerische Verfahren

Wir präsentieren in diesem Kapitel numerische Verfahren zur Berechnung des Modalwert-Schätzers und der empirischen Bayes-Schätzung im nichtlinearen Modell.

5.1. Numerische Verfahren für den Modalwert-Schätzer

In diesem Abschnitt stellen wir Verfahren zur Bestimmung des Modalwert-Schätzers im nichtlinearen Modell (4.18) mit $\Theta = \mathbb{R}^p$ vor, genauer zur Lösung der Minimierungsprobleme[1]

$$\arg\min_{\vartheta\in\mathbb{R}^p} \|y-\eta(\vartheta)\|_\Sigma^2 - 2\ln(q(\vartheta)) \tag{5.1}$$

bzw. spezieller

$$\arg\min_{\vartheta\in\mathbb{R}^p} \|y-\eta(\vartheta)\|_\Sigma^2 + \|\vartheta-\mu\|_B^2\,, \tag{5.2}$$

vergleiche (4.20) und (4.21). Mit

$$Z^y(\vartheta) = \|y-\eta(\vartheta)\|_\Sigma^2 - 2\ln(q(\vartheta))$$

gilt dann für eine differenzierbare a-priori-Dichte q

$$\begin{aligned}\operatorname{grad}^{\mathrm{T}} Z^y(\vartheta) &= 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta)-y\right) - 2\operatorname{grad}^{\mathrm{T}}(\ln(q))(\vartheta)\\ &= 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta)-y\right) - \frac{2}{q(\vartheta)}\operatorname{grad}^{\mathrm{T}} q(\vartheta)\end{aligned}$$

und, wenn q zweimal differenzierbar ist, so gilt

$$\begin{aligned}\operatorname{Hess} Z^y(\vartheta) &= 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + 2H_\eta(\vartheta) - 2\operatorname{Hess}(\ln(q))(\vartheta)\\ &= 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + 2\sum_{j=1}^{N}\left(\Sigma^{-1}\left(\eta(\vartheta)-y\right)\right)_j \operatorname{Hess}\eta_j(\vartheta)\\ &\quad + \frac{2}{q^2(\vartheta)}\operatorname{grad}^{\mathrm{T}} q(\vartheta)\operatorname{grad} q(\vartheta) - \frac{2}{q(\vartheta)}\operatorname{Hess} q(\vartheta).\end{aligned}$$

Für

$$W^y(\vartheta) = \|y-\eta(\vartheta)\|_\Sigma^2 + \|\vartheta-\mu\|_B^2$$

[1] Für Minimierungsprobleme mit eingeschränktem Parameterbereich $\Theta \subsetneq \mathbb{R}^p$ verweisen wir auf die Anmerkungen auf S. 98.

gilt

$$\operatorname{grad}{}^{\mathrm{T}} W^y(\vartheta) = 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - y\right) + 2B^{-1}(\vartheta - \mu) \quad \text{und}$$
$$\operatorname{Hess} W^y(\vartheta) = 2B^{-1} + 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + 2\sum_{j=1}^{N}\left(\Sigma^{-1}\left(\eta(\vartheta) - y\right)\right)_j \operatorname{Hess}\eta_j(\vartheta).$$

5.1.1. Der Algorithmus von Collins u.a.

Ein Verfahren zur Lösung des Minimierungsproblems (5.2) wurde von Collins u. a. (1974) vorgeschlagen.

Die Idee des Verfahrens liegt darin, beginnend mit $\vartheta_0 = \mu$ und $B_0 = B$ im $(k+1)$-ten Schritt die Modellfunktion in ϑ_k zu linearisieren, d.h.

$$\mathfrak{l}_k : \mathbb{R}^p \to \mathbb{R}^N,\ \mathfrak{l}_k(\vartheta) = J_\eta(\vartheta_k)\cdot(\vartheta - \vartheta_k) + \eta(\vartheta_k)$$

an Stelle von η zu verwenden. Hiermit wird dann das affin-lineare Modell

$$P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} = \mathcal{N}(\mathfrak{l}_k(\vartheta), \Sigma),$$
$$P^{\boldsymbol{\theta}} = \mathcal{N}(\vartheta_k, B_k)$$

betrachtet und ϑ_{k+1} auf den Mittelwert und B_{k+1} auf die Kovarianzmatrix der a-posteriori-Verteilung zur Beobachtung y gesetzt, d.h. die alte a-posteriori-Verteilung wird die neue a-priori-Verteilung und für das neue Modell wird die Modellfunktion im Punkt ϑ_k, der Bayes-Schätzung im alten Modells, linearisiert. Dies wird für $k = 0, 1, 2, \ldots$ fortgesetzt, bis Konvergenz eintritt. Nutzt man hierbei die Formeln (4.15) und (4.17), so ergibt sich folgender Algorithmus, wie er von Collins u. a. (1974) formuliert wurde.

Algorithmus 5.1

Input: $\Sigma \in \mathbb{R}^{N\times N}$ positiv definit,
$B \in \mathbb{R}^{p\times p}$ positiv semidefinit,
$\mu \in \mathbb{R}^p$, $y \in \mathbb{R}^N$,
$\eta : \mathbb{R}^p \to \mathbb{R}^N$ stetig differenzierbar,
Abbruchkriterium

Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation der Minimumstelle von (5.2))

1. Setze $k = 0$, $\vartheta_0 = \mu$, $B_0 = B$.

2. Setze

$$\vartheta_{k+1} = \vartheta_k + B_k J_\eta^{\mathrm{T}}(\vartheta_k)\left(\Sigma + J_\eta(\vartheta_k)B_k J_\eta^{\mathrm{T}}(\vartheta_k)\right)^{-1}\left(y - \eta(\vartheta_k)\right), \tag{5.3}$$
$$B_{k+1} = B_k - B_k J_\eta^{\mathrm{T}}(\vartheta_k)\left(\Sigma + J_\eta(\vartheta_k)B_k J_\eta^{\mathrm{T}}(\vartheta_k)\right)^{-1} J_\eta(\vartheta_k)B_k. \tag{5.4}$$

3. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten setze $k := k + 1$ und gehe zu Schritt 2.

Wir bezeichnen mit PSD(p) die Menge der positiv semidefiniten Matrizen aus $\mathbb{R}^{p\times p}$. Sei $\operatorname{rg}(J_\eta(\vartheta)) = p$ für alle $\vartheta \in \mathbb{R}^p$. Betrachtet man den zum Algorithmus 5.1 zugehörigen Prozess $\mathcal{J} = (\mathcal{G}, D)$ mit $D = \mathbb{R}^p \times \mathrm{PSD}(p)$ und

$$\mathcal{G} : D \to D,\ \mathcal{G}\begin{pmatrix}\vartheta\\ B\end{pmatrix} = \begin{pmatrix}\vartheta + BJ_\eta^{\mathrm{T}}(\vartheta)\left(\Sigma + J_\eta(\vartheta)BJ_\eta^{\mathrm{T}}(\vartheta)\right)^{-1}(y-\eta(\vartheta))\\ B - BJ_\eta^{\mathrm{T}}(\vartheta)\left(\Sigma + J_\eta(\vartheta)BJ_\eta^{\mathrm{T}}(\vartheta)\right)^{-1}J_\eta(\vartheta)B\end{pmatrix},$$

dann folgt

$$\mathcal{G}\begin{pmatrix}\vartheta\\ B\end{pmatrix} = \begin{pmatrix}\vartheta\\ B\end{pmatrix} \Longleftrightarrow B = \mathbf{0}_{p\times p},$$

d.h. es gibt in der Regel viele Fixpunkte $\left(\begin{smallmatrix}\vartheta\\ B\end{smallmatrix}\right)$ des Prozesses, bei denen ϑ keine Lösung des Minimierungsproblems (5.2) ist. Ein weiteres Problem zeigt der folgende Satz.

Satz 5.2
Seien $B \in \mathbb{R}^{p\times p}$ und $\Sigma \in \mathbb{R}^{N\times N}$ symmetrisch und positiv definit, $y_0 \in \mathbb{R}^N$, $\mu \in \mathbb{R}^p$, $\tilde{\vartheta}_0 \in \mathbb{R}^p$, $X \in \mathbb{R}^{N\times p}$ mit $\operatorname{rg}(X) = p$ und

$$\eta : \mathbb{R}^p \to \mathbb{R}^N,\ \eta(\vartheta) = y_0 + X\cdot(\vartheta - \tilde{\vartheta}_0).$$

Dann gilt für den Algorithmus 5.1

$$\lim_{k\to\infty} \vartheta_k = \widehat{\vartheta}(y),$$

d.h. für alle $y \in \mathbb{R}^N$ konvergiert die generierte Folge gegen die Kleinste-Quadrate-Schätzung.

Beweis: Wir nehmen o.B.d.A. $\tilde{\vartheta}_0 = \mu$ an, da man ansonsten y_0 durch $\tilde{y}_0 = y_0 + X(\mu - \tilde{\vartheta}_0)$ ersetzt. Wegen (4.14) und (4.15) folgt mit der Kleinste-Quadrate-Schätzung

$$\widehat{\vartheta}(y) = \vartheta_0 + (X^{\mathrm{T}}\Sigma^{-1}X)^{-1}X^{\mathrm{T}}\Sigma^{-1}(y - y_0),$$

dass

$$\begin{aligned}\vartheta_{k+1} &= B_{k+1}(B_k^{-1}\vartheta_k + X^{\mathrm{T}}\Sigma^{-1}X\hat{\vartheta}(y)),\\ B_{k+1} &= (B_k^{-1} + X^{\mathrm{T}}\Sigma^{-1}X)^{-1}.\end{aligned}$$

Damit gilt

$$\begin{aligned}B_k &= (B^{-1} + kX^{\mathrm{T}}\Sigma^{-1}X)^{-1} \overset{k\to\infty}{\longrightarrow} \mathbf{0}_{p\times p},\\ \vartheta_k &= B_k(B^{-1}\mu + kX^{\mathrm{T}}\Sigma^{-1}X\widehat{\vartheta}(y)) \qquad (5.5)\\ &= \Big(\frac{1}{k}B^{-1} + X^{\mathrm{T}}\Sigma^{-1}X\Big)^{-1}\Big(\frac{1}{k}B^{-1}\mu + X^{\mathrm{T}}\Sigma^{-1}X\widehat{\vartheta}(y)\Big) \overset{k\to\infty}{\longrightarrow} \widehat{\vartheta}(y).\end{aligned}$$

□

Der Algorithmus löst also im affin-linearen Fall das falsche Minimierungsproblem!

Es gelten die Voraussetzungen von Satz 5.2 und es sei $\mu \neq \widehat{\vartheta}(y)$. Wegen (5.5) gilt dann

$$\frac{\left\|\vartheta_{k+1} - \widehat{\vartheta}(y)\right\|}{\left\|\vartheta_k - \widehat{\vartheta}(y)\right\|} = \frac{\left\|\left(\frac{1}{k+1}B^{-1} + X^{\mathrm{T}}\Sigma^{-1}X\right)^{-1}\frac{1}{k+1}B^{-1}(\mu - \widehat{\vartheta}(y))\right\|}{\left\|\left(\frac{1}{k}B^{-1} + X^{\mathrm{T}}\Sigma^{-1}X\right)^{-1}\frac{1}{k}B^{-1}(\mu - \widehat{\vartheta}(y))\right\|} \overset{k\to\infty}{\longrightarrow} 1,$$

d.h. es liegt nicht einmal lineare Konvergenz gegen die Kleinste-Quadrate-Schätzung vor. Es wird also weder das richtige Minimierungsproblem gelöst, noch ist der Algorithmus aufgrund der schlechten Konvergenzordnung ein geeignetes Verfahren zur Lösung des Kleinste-Quadrate-Problems.

Bemerkung 5.3
Interessant ist ferner, dass dieses unerwünschte Ergebnis auch noch gilt, falls man lediglich ϑ_k, $k \in \mathbb{N}_0$, gemäß (5.3) definiert und $B_k \equiv B$ konstant lässt, d.h. mit $\vartheta_0 = \mu$ induktiv

$$\vartheta_{k+1} = \vartheta_k + BJ_\eta^{\mathrm{T}}(\vartheta_k)\left(\Sigma + J_\eta(\vartheta_k)BJ_\eta^{\mathrm{T}}(\vartheta_k)\right)^{-1}(y - \eta(\vartheta_k))$$

setzt, d.h. den 1-Schritt-Prozess $\mathcal{J} = (\mathcal{G}, \mathbb{R}^p)$ mit

$$\mathcal{G} : \mathbb{R}^p \to \mathbb{R}^p,\ \mathcal{G}(\vartheta) = \vartheta + BJ_\eta^{\mathrm{T}}(\vartheta)\left(\Sigma + J_\eta(\vartheta)BJ_\eta^{\mathrm{T}}(\vartheta)\right)^{-1}(y - \eta(\vartheta))$$

betrachtet. Überträgt man (4.17) mit $X = J_\eta(\vartheta)$, $y_0 = \eta(\vartheta)$, $\mu = \vartheta$ und $\vartheta_0 = \vartheta$, so gilt

$$\mathcal{G}(\vartheta) = (J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + B^{-1})^{-1}(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}(y - \eta(\vartheta) + J_\eta(\vartheta)\cdot\vartheta) + B^{-1}\vartheta).$$

Daher folgt

$$\begin{aligned}\mathcal{G}(\vartheta) = \vartheta &\iff (J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta) + B^{-1})\vartheta = (J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}(y - \eta(\vartheta) + J_\eta(\vartheta)\cdot\vartheta) + B^{-1}\vartheta)\\ &\iff J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}(y - \eta(\vartheta)) = \mathbf{0}_p,\end{aligned}$$

d.h. ϑ ist Fixpunkt von $\mathcal{G}$ genau dann, wenn $\operatorname{grad}^{\mathrm{T}} S^{\eta,y}(\vartheta) = \mathbf{0}_p$. Daher sind die Fixpunkte genau die stationären Punkte von $S^{\eta,y} : \mathbb{R}^p \to \mathbb{R}$, $S^{\eta,y}(\vartheta) = \|\eta(\vartheta) - y\|_\Sigma^2$. Es wird somit bei Konvergenz auch hier das falsche Minimierungsproblem gelöst. Im affin-linearen Fall kann man wieder Konvergenz gegen die Kleinste-Quadrate-Schätzung zeigen. Unter den Voraussetzungen von Satz 5.2 gilt

$$\mathcal{G}(\vartheta) = (X^{\mathrm{T}}\Sigma^{-1}X + B^{-1})^{-1}\left(X^{\mathrm{T}}\Sigma^{-1}X\widehat{\vartheta}(y) + B^{-1}\vartheta\right),$$

vergleiche (4.14). Mit

$$A = (BX^{\mathrm{T}}\Sigma^{-1}X + I_p)^{-1}$$

folgt für $k \in \mathbb{N}_0$

$$\vartheta_k = (A + A^2 + \cdots + A^k)BX^{\mathrm{T}}\Sigma^{-1}X\widehat{\vartheta}(y) + A^k\mu = (I_p - A^k)\widehat{\vartheta}(y) + A^k\mu,$$

da

$$ABX^{\mathrm{T}}\Sigma^{-1}X = I_p - A.$$

Da B und $X^{\mathrm{T}}\Sigma^{-1}X$ positiv definit sind, hat das Produkt $BX^{\mathrm{T}}\Sigma^{-1}X$ nur positive Eigenwerte, siehe Huppert (1990, S. 192), A^{-1} nur Eigenwerte größer als 1 und folglich A nur Eigenwerte kleiner als 1, womit der Spektralradius $\rho(A)$ kleiner als 1 ist. Daher gilt $\lim_{k\to\infty} A^k = 0$ und wieder

$$\lim_{k\to\infty} \vartheta_k = \widehat{\vartheta}(y).$$

Da

$$\frac{\left\|\vartheta_{k+1} - \widehat{\vartheta}(y)\right\|}{\left\|\vartheta_k - \widehat{\vartheta}(y)\right\|} = \frac{\left\|A^{k+1}(\mu - \widehat{\vartheta}(y))\right\|}{\left\|A^k(\mu - \widehat{\vartheta}(y))\right\|}$$

und $\rho(A) < 1$, ist nach Satz 3.60 der Prozess $\mathcal{G}$ im affin-linearen Fall mit $\mathrm{rg}(X) = p$ linear konvergent gegen die Kleinste-Quadrate-Schätzung, so dass dieses abgewandelte Verfahren selbst im affin-linearen Fall keine besonders gute Konvergenzrate zeigt und daher wohl auch im Allgemeinen kein geeignetes Verfahren zur Bestimmung der Kleinste-Quadrate-Schätzung ist.

Die naheliegende Idee, die a-priori-Verteilung fest zu lassen und nur den Linearisierungspunkt in jeder Iteration zu ändern, verfolgen wir in Abschnitt 5.1.3.

5.1.2. Der Algorithmus von Friswell

Ein weiteres Verfahren zur Lösung des Minimierungsproblems (5.2) stammt von Friswell (1989), der zu den Startwerten $\vartheta_0 = \mu$, $B_0 = B$ und $D_0 = \mathbf{0}_{p\times N}$ folgende Iterationsformeln vorschlug, wobei wir hier $J_k = J_\eta(\vartheta_k)$ abkürzen:

$$\begin{aligned}
\vartheta_{k+1} &= \vartheta_k + (B_k J_k^{\mathrm{T}} - D_k)\left(J_k B_k J_k^{\mathrm{T}} - J_k D_k - D_k^{\mathrm{T}} J_k^{\mathrm{T}} + \Sigma\right)^{-1}(y - \eta(\vartheta_k)),\\
B_{k+1} &= B_k - (B_k J_k^{\mathrm{T}} - D_k)\left(J_k B_k J_k^{\mathrm{T}} - J_k D_k - D_k^{\mathrm{T}} J_k^{\mathrm{T}} + \Sigma\right)^{-1}(B_k J_k^{\mathrm{T}} - D_k)^{\mathrm{T}},\\
D_{k+1} &= D_k - (B_k J_k^{\mathrm{T}} - D_k)\left(J_k B_k J_k^{\mathrm{T}} - J_k D_k - D_k^{\mathrm{T}} J_k^{\mathrm{T}} + \Sigma\right)^{-1}(J_k D_k - \Sigma).
\end{aligned}$$

Diese Formeln kann man wie folgt interpretieren, siehe unsere Diskussion in Zehn u. a. (1999). In der $(k+1)$-ten Iteration linearisiert man wieder die Modellfunktion η in ϑ_k und betrachtet

$$\mathfrak{l}_k : \mathbb{R}^p \to \mathbb{R}^N,\ \mathfrak{l}_k(\vartheta) = J_\eta(\vartheta_k)\cdot(\vartheta - \vartheta_k) + \eta(\vartheta_k).$$

Im Modell

$$\begin{aligned}
P^{\boldsymbol{\theta}} &= \mathcal{N}(\vartheta_k, B_k),\\
P^{\mathbf{e}} &= \mathcal{N}(\mathbf{0}_N, \Sigma) \quad \text{mit}\\
D_k &= -\mathrm{Cov}(\boldsymbol{\theta}, \mathbf{e}) = -\mathrm{E}(\boldsymbol{\theta}\mathbf{e}^{\mathrm{T}})
\end{aligned}$$

mit $\boldsymbol{\theta}$ und $\mathbf{e}$ gemeinsam normalverteilt und mit der Zufallsvariablen $\boldsymbol{Y} = \mathfrak{l}_k(\boldsymbol{\theta}) + \mathbf{e}$ gilt

$$\begin{aligned}\vartheta_{k+1} &= \mathrm{E}[P^{\boldsymbol{\theta}|\boldsymbol{Y}=y}],\\ B_{k+1} &= \mathrm{D}[P^{\boldsymbol{\theta}|\boldsymbol{Y}=y}],\\ D_{k+1} &= -\mathrm{Cov}[P^{(\boldsymbol{\theta},\mathbf{e})|\boldsymbol{Y}=y}].\end{aligned}$$

Man beachte ferner, dass

$$\begin{aligned}\mathrm{D}(\boldsymbol{Y}) &= J_\eta(\vartheta_k)B_kJ_\eta^{\mathrm{T}}(\vartheta_k) - J_\eta(\vartheta_k)D_k - D_k^{\mathrm{T}}J_\eta^{\mathrm{T}}(\vartheta_k) + \Sigma \quad \text{und}\\ \mathrm{D}[P^{\mathbf{e}|\boldsymbol{Y}=y}] &= \Sigma - (D_k^{\mathrm{T}}J_k^{\mathrm{T}} - \Sigma)(\mathrm{D}(\boldsymbol{Y}))^{-1}(J_kD_k - \Sigma) =: \Sigma_{k+1},\end{aligned}$$

gilt. Daher ist

$$\mathrm{D}[P^{(\boldsymbol{\theta},\mathbf{e})|\boldsymbol{Y}=y}] = \begin{pmatrix} B_{k+1} & -D_{k+1}\\ -D_{k+1} & \Sigma_{k+1}\end{pmatrix}$$

positiv semidefinit. Da $\Sigma_{k+1} \leq_{\mathfrak{L}} \Sigma$ gilt, folgt leicht, dass auch

$$\begin{pmatrix} B_{k+1} & -D_{k+1}\\ -D_{k+1} & \Sigma\end{pmatrix}$$

positiv semidefinit ist.

Wir setzen

$$\mathcal{D} = \left\{(B,D) \in \mathrm{PSD}(p) \times \mathbb{R}^{p\times N} : \begin{pmatrix} B & -D\\ -D & \Sigma\end{pmatrix}\right\} \in \mathrm{PSD}(p+N)$$

und untersuchen den (aufgrund obiger Überlegung wohldefinierten) zugehörigen Prozess $\mathcal{J} = \big(\mathcal{G}, \mathbb{R}^p \times \mathcal{D}\big)$ mit

$$\mathcal{G} : \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}^p \times \mathcal{D},\ \mathcal{G}\begin{pmatrix}\vartheta\\ B\\ D\end{pmatrix} =$$

$$\begin{pmatrix} \vartheta + (BJ_\eta^{\mathrm{T}}(\vartheta) - D)\left(J_\eta(\vartheta)BJ_\eta^{\mathrm{T}}(\vartheta) - J_\eta(\vartheta)D - D^{\mathrm{T}}J_\eta^{\mathrm{T}}(\vartheta) + \Sigma\right)^{-1}(y - \eta(\vartheta))\\ B - (BJ_\eta^{\mathrm{T}}(\vartheta) - D)\left(J_\eta(\vartheta)BJ_\eta^{\mathrm{T}}(\vartheta) - J_\eta(\vartheta)D - D^{\mathrm{T}}J_\eta^{\mathrm{T}}(\vartheta) + \Sigma\right)^{-1}(BJ_\eta^{\mathrm{T}}(\vartheta) - D)^{\mathrm{T}}\\ D - (BJ_\eta^{\mathrm{T}}(\vartheta) - D)\left(J_\eta(\vartheta)BJ_\eta^{\mathrm{T}}(\vartheta) - J_\eta(\vartheta)D - D^{\mathrm{T}}J_\eta^{\mathrm{T}}(\vartheta) + \Sigma\right)^{-1}(J_\eta(\vartheta)D - \Sigma)\end{pmatrix}$$

auf Fixpunkte. Es gilt

$$\mathcal{G}\begin{pmatrix}\vartheta\\ B\\ D\end{pmatrix} = \begin{pmatrix}\vartheta\\ B\\ D\end{pmatrix} \Longleftrightarrow BJ_\eta^{\mathrm{T}}(\vartheta) = D.$$

Es gibt also in der Regel viele Fixpunkte $\left(\begin{smallmatrix}\vartheta\\ B\\ D\end{smallmatrix}\right)$ des Prozesses, bei denen ϑ keine Lösung des Minimierungsproblems (5.2) ist.

Im affin-linearen Modell verhält sich das Verfahren allerdings gut und es wird nach einem Schritt der Modalwertschätzer bestimmt.

Satz 5.4
Seien $\Sigma \in \mathbb{R}^{N\times N}$ symmetrisch und positiv definit und $B \in \mathbb{R}^{p\times p}$ symmetrisch und positiv semidefinit, $y_0 \in \mathbb{R}^N$, $\mu \in \mathbb{R}^p$, $\tilde{\vartheta}_0 \in \mathbb{R}^p$, $X \in \mathbb{R}^{N\times p}$ und

$$\eta : \mathbb{R}^p \to \mathbb{R}^N,\ \eta(\vartheta) = y_0 + X \cdot (\vartheta - \tilde{\vartheta}_0).$$

Dann gilt für die vom Verfahren von Friswell erzeugte Folge $(\vartheta_k)_{k\in\mathbb{N}_0}$ mit Startwerten $\vartheta_0 = \mu$, $B_0 = B$ und $D_0 = \mathbf{0}_{p\times N}$

$$\vartheta_k = \widehat{\vartheta}_{Bayes}(y) = \mu + BX^{\mathrm{T}}(\Sigma + XBX^{\mathrm{T}})^{-1}(y - y_0 + X(\tilde{\vartheta}_0 - \mu)) \quad \textit{für } k \geq 1.$$

Beweis: Es gilt

$$\begin{aligned}
\vartheta_1 &= \widehat{\vartheta}_{\mathrm{Bayes}}(y) = \mu + BX^{\mathrm{T}}(\Sigma + XBX^{\mathrm{T}})^{-1}(y - y_0 + X(\tilde{\vartheta}_0 - \mu)) \\
D_1 &= BX^{\mathrm{T}}\left(XBX^{\mathrm{T}} + \Sigma\right)^{-1}\Sigma \\
B_1 &= B - BX^{\mathrm{T}}\left(XBX^{\mathrm{T}} + \Sigma\right)^{-1}XB
\end{aligned}$$

und damit $B_1X^{\mathrm{T}} - D_1 = \mathbf{0}_{p\times N}$. Somit folgt $\vartheta_2 = \vartheta_1$, $B_2 = B_1$ und $D_2 = D_1$ und daher die Behauptung. □

Bemerkung 5.5
Im linearen Fall terminiert also das Iterationsverfahren nach dem 2. Schritt. Man kann also hoffen, dass bei fast-linearen Modellen schnelle Konvergenz eintritt. Die Tatsache, dass es viele Fixpunkte des Prozesses gibt, die nicht Lösung des gewünschten Minimierungsproblems sind, ist jedoch bedenklich. Wir werden im Abschnitt 5.1.6 an einem einfachen Beispiel sehen, dass das Verfahren sehr ungünstige Eigenschaften zeigt.

5.1.3. Ein Algorithmus basierend auf dem Gauß-Newton-Verfahren

Wir untersuchen nun, wie am Ende des Abschnitts 5.1.1 angesprochen, die neue Idee, ausgehend von $\vartheta_0 = \mu$ im $(k+1)$-ten Schritt das Modell in ϑ_k zu linearisieren und für ϑ_{k+1} die Bayes-Schätzung zur originalen a-priori-Verteilung im linearisierten Modell zu wählen, d.h. anders als bei den bisherigen Verfahren in diesem Kapitel die a-priori-Verteilung im gesamten Verfahren fix zu lassen. Wir betrachten also im $(k+1)$-ten Schritt das Modell

$$\begin{aligned}
P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathcal{N}(\mathfrak{l}_k(\vartheta), \Sigma), \\
P^{\boldsymbol{\theta}} &= \mathcal{N}(\mu, B)
\end{aligned}$$

und setzen ϑ_{k+1} auf die Bayes-Schätzung in diesem Modell zur Beobachtung y. Es ergibt sich mit $\vartheta_0 = \mu$ nach (4.17)

$$\vartheta_{k+1} = \mu + BJ_\eta^{\mathrm{T}}(\vartheta_k)\left(\Sigma + J_\eta(\vartheta_k)BJ_\eta^{\mathrm{T}}(\vartheta_k)\right)^{-1}(y - \eta(\vartheta_k) + J_\eta(\vartheta_k)(\vartheta_k - \mu)) \tag{5.6}$$

bzw. äquivalent, vergleiche (4.16),

$$\vartheta_{k+1} = \vartheta_k + (J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k) + B^{-1})^{-1}\big(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}(y-\eta(\vartheta_k)) + B^{-1}(\mu-\vartheta_k)\big). \tag{5.7}$$

Es zeigt sich der folgende überraschende Zusammenhang mit dem Gauß-Newton-Verfahren, wenn wir (4.22) verwenden.

Verwendet man das Gauß-Newton-Verfahren ohne Liniensuche (Algorithmus 3.40) für

$$W^y(\vartheta) = \left\| \begin{pmatrix} y \\ \mu \end{pmatrix} - \begin{pmatrix} \eta(\theta) \\ \theta \end{pmatrix} \right\|^2_{\left(\begin{smallmatrix} \Sigma & 0 \\ 0 & B \end{smallmatrix}\right)} = \|\tilde{y} - \tilde{\eta}(\vartheta)\|^2_{\tilde{\Sigma}},$$

wobei wir wieder $\tilde{\eta}(\vartheta) = \binom{\eta(\vartheta)}{\vartheta}$, $\tilde{y} = \binom{y}{\mu}$ und $\tilde{\Sigma} = \left(\begin{smallmatrix} \Sigma & 0 \\ 0 & B \end{smallmatrix}\right)$ setzen, so ergibt sich mit

$$\begin{aligned} d_k &= -(J_{\tilde{\eta}}^{\mathrm{T}}(\vartheta_k)\tilde{\Sigma}^{-1}J_{\tilde{\eta}}(\vartheta_k))^{-1}J_{\tilde{\eta}}^{\mathrm{T}}(\vartheta_k)\tilde{\Sigma}^{-1}\,(\tilde{\eta}(\vartheta_k) - y) \\ &= -(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k) + B^{-1})^{-1}\big(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}(\eta(\vartheta_k) - y) + B^{-1}(\vartheta_k - \mu)\big) \end{aligned} \tag{5.8}$$

$$\vartheta_{k+1} = \vartheta_k + d_k = \vartheta_k + (J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k) + B^{-1})^{-1}\big(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}(y-\eta(\vartheta_k)) + B^{-1}(\mu-\vartheta_k)\big).$$

Daher stimmen die Iterationsvorschriften (5.6) bzw. (5.7) mit dem Gauß-Newton-Verfahren für $W^y(\vartheta) = \|\tilde{y} - \tilde{\eta}(\vartheta)\|^2_{\tilde{\Sigma}}$ überein!

5.1.4. Der Algorithmus von Pázman

Zur numerischen Berechnung des Modalwert-Schätzers hat Pázman (1992, 1993), ein Verfahren vorgeschlagen, das sich wie folgt motivieren lässt:
Zum einen wird im $(k+1)$-ten Schritt wie z.B. beim Gauß-Newton-Verfahren (siehe Bemerkung 3.42) η durch die Linearisierung $\mathfrak{l}_k$ ersetzt, so dass Z^y durch Z_k^y mit

$$Z_k^y(\vartheta) = \|y - \mathfrak{l}_k(\vartheta)\|^2_\Sigma - 2\ln(q(\vartheta))$$

approximiert wird und zum anderen wird in der Gleichung

$$\operatorname{grad}^{\mathrm{T}} Z_k^y(\vartheta) = \mathbf{0}_p \iff J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\,(\mathfrak{l}_k(\vartheta) - y) = \operatorname{grad}^{\mathrm{T}}(\ln(q))(\vartheta) \tag{5.9}$$

die rechte Seite durch den Wert in ϑ_k ersetzt, d.h.

$$J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\,(\mathfrak{l}_k(\vartheta) - y) = \operatorname{grad}^{\mathrm{T}}(\ln(q))(\vartheta_k)$$

betrachtet. Dieses lineare Gleichungssystem für ϑ wird von $\vartheta_{k+1} = \vartheta_k + d_k$ mit

$$d_k = (J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k))^{-1}\Big(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}(y - \eta(\vartheta_k)) + \operatorname{grad}^{\mathrm{T}}(\ln(q))(\vartheta_k)\Big)$$

gelöst, falls $\operatorname{rg}(J_\eta^{\mathrm{T}}(\vartheta_k)) = p$.

Insgesamt ergibt sich folgender Algorithmus.

Algorithmus 5.6

Input: $\Sigma \in \mathbb{R}^{N\times N}$ positiv definit,
$q : \mathbb{R}^p \to \mathbb{R}$ differenzierbar,
$y \in \mathbb{R}^N$, Startpunkt ϑ_0
$\eta : \mathbb{R}^p \to \mathbb{R}^N$ stetig differenzierbar,
Abbruchkriterium

Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation der Minimumstelle von (5.2))

1. Setze $k = 0$.
2. Löse das lineare Gleichungssystem
$$J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k)d_k = J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}(y - \eta(\vartheta_k) + \operatorname{grad}^{\mathrm{T}}(\ln(q))(\vartheta_k) \tag{5.10}$$
für d_k und setze $\vartheta_{k+1} = \vartheta_k + d_k$.
3. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten setze $k := k + 1$ und gehe zu Schritt 2.

Pázman schlägt ferner vor, dieses Verfahren nach Schritt 2 um eine Liniensuche zu ergänzen, d.h. anstatt $\vartheta_{k+1} = \vartheta_k + d_k$ zu setzen, bestimmt man λ_k durch eine (event. inexakte) Liniensuche zu Z^y in ϑ_k mit Suchrichtung d_k und setzt $\vartheta_{k+1} = \vartheta_k + \lambda_k d_k$.

Wir untersuchen nun den zugehörigen Prozess $\mathcal{J} = (\mathcal{G}, \mathbb{R}^p)$ mit

$$\mathcal{G} : \mathbb{R}^p \to \mathbb{R}^p,\ \mathcal{G}(\vartheta) = \vartheta + (J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta))^{-1}\Big(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}(y - \eta(\vartheta)) + \operatorname{grad}^{\mathrm{T}}(\ln(q))(\vartheta)\Big).$$

Es gilt

$$\mathcal{G}(\vartheta) = \vartheta \iff \operatorname{grad}^{\mathrm{T}} Z^y(\vartheta) = 2J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}\left(\eta(\vartheta) - y\right) - 2\operatorname{grad}^{\mathrm{T}}(\ln(q))(\vartheta) = \mathbf{0}_p,$$

und die Fixpunkte des Prozesses sind somit genau die stationären Punkte von Z^y. Ferner bestimmen wir analog zu Satz 3.63 die lokalen Konvergenzeigenschaften. Für die Ableitung von $\mathcal{G}$ gilt für $l \in \{1, \dots, p\}$ wegen Korollar 3.62

$$\frac{\partial}{\partial\vartheta_l}\mathcal{G}(\vartheta) = e_l - \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1}\left(\frac{1}{2}\operatorname{Hess} Z^y(\vartheta)\right)_l + \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1} c_l(\vartheta)$$

mit

$$c_l(\vartheta) = \frac{\partial}{\partial\vartheta_l}\left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)\ \left(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta)\right)^{-1}\frac{1}{2}\operatorname{grad} Z^y(\vartheta),$$

wobei $\left(\frac{1}{2}\operatorname{Hess} Z^y(\vartheta)\right)_l$ die l-te Spalte von $\frac{1}{2}\operatorname{Hess} Z^y(\vartheta)$ bezeichne. Für stationäre Punkte ϑ^* von $Z^y(\vartheta)$ folgt daher

$$J_{\mathcal{G}}(\vartheta^*) = -\left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)\right)^{-1}\left(H_\eta(\vartheta^*) - \operatorname{Hess}(\ln(q))(\vartheta^*)\right).$$

Gilt $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) < 1$, dann ist nach Satz 3.60 ϑ^* Attraktorpunkt des Prozesses $\mathcal{J}$, der Prozess ist linear konvergent und der Quotientenkonvergenzfaktor beträgt bestenfalls

$\rho\left(J_{\mathcal{G}}(\vartheta^*)\right)$. In der Regel wird daher nur dann lineare Konvergenz eintreten, falls $\ln(q)$ und η nicht zu stark nichtlinear in ϑ sind. Zudem ist $\rho\left(J_{\mathcal{G}}(\vartheta^*)\right) = 0$ eine notwendige Bedingung dafür, dass der Prozess superlinear konvergent ist, was in der Regel nur dann gegeben ist, wenn η und $\ln(q)$ linear in ϑ sind und daher $H_\eta(\vartheta^*) = \mathbf{0}_{p\times p}$ $\operatorname{Hess}\left(\ln(q)\right)(\vartheta^*) = \mathbf{0}_{p\times p}$.

Im Fall $q = f_{\mathcal{N}(\mu,B)}$, d.h. zur Lösung des Minimierungsproblems (5.2), ergibt sich speziell

$$d_k = (J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta))^{-1}(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}(y-\eta(\vartheta_k)) + B^{-1}(\mu-\vartheta_k)).$$

Im Unterschied zum Algorithmus basierend auf dem Gauß-Newton-Verfahren ist der erste Term $(J_\eta^{\mathrm{T}}(\vartheta)\Sigma^{-1}J_\eta(\vartheta))^{-1}$ statt $(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k) + B^{-1})^{-1}$. In diesem Fall kann man die Gleichung (5.9) direkt nach ϑ auflösen und daher auf das Einsetzen des aktuellen Punktes verzichten. Dies liefert dann wieder die Formeln des Gauß-Newton-Verfahrens. Wir wollen im Fall $q = f_{\mathcal{N}(\mu,B)}$ noch ferner das lokale Konvergenzverhalten des Algorithmus von Pázman untersuchen. Es ergibt sich dann

$$J_{\mathcal{G}}(\vartheta^*) = -\left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*)\right)^{-1}\left(H_\eta(\vartheta^*) + B^{-1}\right),$$

während sich für den Prozess $\tilde{\mathcal{J}} = (\tilde{\mathcal{G}}, \mathbb{R}^p)$ basierend auf dem Gauß-Newton-Verfahren, siehe Abschnitt 5.1.3, mit Satz 3.63

$$J_{\tilde{\mathcal{G}}}(\vartheta^*) = -\left(J_{\tilde{\eta}}^{\mathrm{T}}(\vartheta^*)\tilde{\Sigma}^{-1}J_{\tilde{\eta}}(\vartheta^*)\right)^{-1} H_{\tilde{\eta}}(\vartheta^*) = -\left(J_\eta^{\mathrm{T}}(\vartheta^*)\Sigma^{-1}J_\eta(\vartheta^*) + B^{-1}\right)^{-1} H_\eta(\vartheta^*)$$

ergibt. In der Regel wird daher der Prozess basierend auf dem Gauß-Newton-Verfahren bessere Konvergenzeigenschaften zeigen, weil der asymptotische Wurzelkonvergenzfaktor kleiner ist.

5.1.5. Weitere Verfahren und Abbruchkriterien

Eine naheliegende Alternative zur Minimierung von Z^y oder W^y besteht darin, ein Quasi-Newton-Verfahren aus Abschnitt 3.2.3 zu verwenden, z.B. das BFGS-Verfahren. Im Fall des Minimierungsproblems W^y bietet sich durch die Schreibweise (4.22) als Kleinste-Quadrate-Problem speziell der Algorithmus NL2SOL von Dennis u. a. (1981) an, siehe S. 95. Aber auch im Fall von Z^y ist je nach Art der a-priori-Dichte q eine Kombination aus der Gauß-Newton-Idee und Quasi-Newton-Verfahren interessant, z.B. könnte man bei leichter Verfügbarkeit der Hesse-Matrix von $\ln(q)$ nur für $\|y-\eta(\vartheta)\|_\Sigma^2$ eine Gauß-Newton-Approximation für die Hesse-Matrix vornehmen. Oder man nimmt für die Hesse-Matrix von $\ln(q)$ eine Quasi-Newton-Approximation, also mittels einer Sekantenmethode, vor und koppelt diese mit einer Gauß-Newton-Approximation zu $\|y-\eta(\vartheta)\|_\Sigma^2$.

Wir schließen hiermit die Vorstellung einiger Verfahren ab und geben noch die Übertragung der Abbruchbedingung gemäß des Orthogonalitätskriteriums aus Abschnitt 3.2.4 an.

In der augenblicklichen Problemstellung der Minimierung von W^y würde man also messen, welchen Winkel ϕ das „Residuum“ $\tilde{y} - \tilde{\eta}(\vartheta_{k+1})$ zur Tangentialebene Bild$(J_{\tilde{\eta}}(\vartheta_{k+1}))$ hat, und bei einem Winkel nahe $\pi/2$ abbrechen. Da in dem von $<,>_{\tilde{\Sigma}}$-induzierten Hilbertraum

$$\cos^2(\phi) = \frac{\left\|\mathcal{P}_{J_{\tilde{\eta}(\vartheta_{k+1})},\tilde{\Sigma}}(\tilde{\eta}(\vartheta_{k+1}) - \tilde{y})\right\|^2_{\tilde{\Sigma}}}{\|\tilde{\eta}(\vartheta_{k+1}) - \tilde{y}\|_{\tilde{\Sigma}}} = \frac{\left\|\mathcal{P}_{J_{\tilde{\eta}(\vartheta_{k+1})},\tilde{\Sigma}}(\tilde{\eta}(\vartheta_{k+1}) - \tilde{y})\right\|^2_{\tilde{\Sigma}}}{\|y - \eta(\vartheta)\|^2_{\Sigma} + \|\vartheta - \mu\|^2_B}$$

gilt, würde man also entsprechend wieder bei $\cos^2(\phi) \leq \delta$ für ein vorher festgelegtes $\delta > 0$ abbrechen. Hierbei gilt

$$\begin{aligned}\mathcal{P}_{J_{\tilde{\eta}(\vartheta_{k+1})},\tilde{\Sigma}} &= \begin{pmatrix} J_\eta(\vartheta_{k+1}) \\ I_p \end{pmatrix} \left((J_\eta^{\mathrm{T}}(\vartheta_{k+1}), I_p) \begin{pmatrix} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & B^{-1} \end{pmatrix} \begin{pmatrix} J_\eta(\vartheta_{k+1}) \\ I_p \end{pmatrix} \right)^{-1} \\ &\quad \cdot (J_\eta^{\mathrm{T}}(\vartheta_{k+1}), I_p) \begin{pmatrix} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & B^{-1} \end{pmatrix} \\ &= \begin{pmatrix} J_\eta(\vartheta_{k+1}) A(\vartheta_{k+1}) J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} & J_\eta(\vartheta_{k+1}) A(\vartheta_{k+1}) B^{-1} \\ A(\vartheta_{k+1}) J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} & A(\vartheta_{k+1}) B^{-1} \end{pmatrix}\end{aligned}$$

mit der Abkürzung

$$A(\vartheta_{k+1}) = \left(J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} J_\eta(\vartheta_{k+1}) + B^{-1} \right)^{-1}.$$

Daraus erhalten wir

$$\begin{aligned}\mathcal{P}_{J_{\tilde{\eta}},\tilde{\Sigma}}(\tilde{\eta}(\vartheta_{k+1}) - \tilde{y}) &= \begin{pmatrix} J_\eta(\vartheta_{k+1}) A(\vartheta_{k+1}) \left(J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} (\eta(\vartheta_{k+1}) - y) + B^{-1}(\vartheta_{k+1} - \mu) \right) \\ A(\vartheta_{k+1}) \left(J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} (\eta(\vartheta_{k+1}) - y) + B^{-1}(\vartheta_{k+1} - \mu) \right) \end{pmatrix} \\ &= \begin{pmatrix} -J_\eta(\vartheta_{k+1}) d_{k+1} \\ -d_{k+1} \end{pmatrix},\end{aligned}$$

wobei

$$d_{k+1} = -(J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} J_\eta(\vartheta_{k+1}) + B^{-1})^{-1} \left(J_\eta^{\mathrm{T}}(\vartheta_{k+1}) \Sigma^{-1} (\eta(\vartheta_{k+1}) - y) + B^{-1}(\vartheta_{k+1} - \mu) \right)$$

die Suchrichtung aus dem $(k+2)$-ten Schritt des Algorithmus basierend auf dem Gauß-Newton-Verfahren ist, siehe (5.8). Insgesamt ergibt sich damit

$$\cos^2(\phi) = \frac{\|J_\eta(\vartheta_{k+1}) d_{k+1}\|^2_{\Sigma} + \|d_{k+1}\|^2_B}{\|\eta(\vartheta_{k+1}) - y\|^2_{\Sigma} + \|\vartheta_{k+1} - \mu\|^2_B}.$$

Somit ist dieses Kriterium wie geschaffen für den Algorithmus basierend auf dem Gauß-Newton-Verfahren, da alle Größen leicht zur Verfügung stehen.

Auch die Kriterien von Dennis (1977) und Bates u. Watts (1981, 1988), siehe Abschnitt 3.2.4, übertragen sich entsprechend.

5.1.6. Vergleich der Verfahren in einem nichtlinearen Fall

Wir untersuchen die besprochenen Algorithmen an einem einfachen nichtlinearen Modell, das wir bereits in Abschnitt 1.3.1 und Abschnitt 1.4 angesprochen haben:

$$\begin{aligned} P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathcal{N}(\eta(\vartheta), \sigma_0^2 I_N), \\ P^{\boldsymbol{\theta}} &= \mathcal{N}(\mu, 1) \end{aligned} \tag{5.11}$$

mit

$$\eta : \mathbb{R} \to \mathbb{R}^N,\ \eta(\vartheta) = \frac{1}{2}\,\mathbf{1}_N\,\vartheta^2,$$

$\sigma_0^2 > 0$ bekannt und $\mu \in \mathbb{R}$ bekannt.

Mit

$$g_{s,t,\mu} : \mathbb{R} \to \mathbb{R},\ g_{s,t,\mu}(\vartheta) = \exp\left(-\frac{1}{8s}\,\vartheta^4 + (\frac{t}{2s} - \frac{1}{2})\vartheta^2 + \mu\vartheta\right)$$

ergibt sich für die Dichte q_y der a-posteriori-Verteilung

$$q_y : \mathbb{R} \to \mathbb{R}, q_y(\vartheta) = \frac{1}{\int_{\mathbb{R}} g_{\sigma_0^2/N,\bar{y},\mu}(\vartheta)\,\mathrm{d}\boldsymbol{\lambda}^1(\vartheta)}\, g_{\sigma_0^2/N,\bar{y},\mu}(\vartheta).$$

Hierbei bezeichnet $\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i$ den empirischen Mittelwert von $y \in \mathbb{R}^N$. Für eine Beobachtung y mit $\bar{y} = 1$, für $\mu = 1$ und für einige Werte von σ_0^2/N sind die zugehörigen a-posteriori-Dichten q_ϑ in Abbildung 5.1 dargestellt. Die Dichten sind linksschief und zum Teil sogar mehrgipflig ($\sigma_0^2/N = 0.1, 0.2$), was sich auch dadurch ausdrückt, dass Median und Erwartungswert kleiner als der Modalwert der Dichte sind.

Hätte man keine a-priori-Annahme für den unbekannten Parameter, so würde man im klassischen Modell die Parametertransformation $\beta = \vartheta^2$ vornehmen und hätte ein lineares Modell und die Kleinste-Quadrate-Schätzung $\widehat{\beta}(y) = 2\bar{y}$ für β. Da die Modellfunktion nicht injektiv ist, ist der zugehörige ϑ-Wert nicht eindeutig bestimmt und es ergeben sich für eine Kleinste-Quadrate-Schätzung von ϑ die beiden möglichen Werte $\widehat{\vartheta}(y) = \pm\sqrt{2\bar{y}}$. Führt man hingegen in unserem Modell mit a-priori-Annahme diese Parametertransformation durch, so muss man die a-priori-Verteilung ebenfalls entsprechend transformieren und es ergibt sich als transformierte a-priori-Verteilung eine nicht-zentrale χ^2-Verteilung mit Nichtzentralitätsparameter μ^2 und einem Freiheitsgrad ergeben. Die Transformation in ein lineares Modell wird also durch eine kompliziertere a-priori-Verteilung erkauft und wir betrachten deshalb das Modell (5.11).

In Abbildung 5.2 sind die Ergebnisse der besprochenen Algorithmen für eine Beobachtung y mit $\bar{y} = 1$ in Abhängigkeit von σ_0^2/N aufgetragen.
Das Ergebnisse des Verfahrens von Collins u. a. (1974) stimmen auch in diesem nichtlinearen Modell stets mit einer Kleinste-Quadrate-Schätzung überein, hier konvergiert das Verfahren jeweils gegen den Wert $\vartheta = \sqrt{2}$. Zudem ist die Konvergenz äußerst langsam.
Ein äußerst unangenehmes Verhalten zeigt sich beim Schätzer von Friswell: Zum einen sind im Bereich $0.25 < \sigma_0^2/N < 0.75$ große Stabilitätsprobleme zu erkennen. Die Rechnungen wurden dabei mit unterschiedlichen Abbruchkriterien und unterschiedlichen Genauigkeiten (mit bis zu 50-stelligen Rechnungen im Dezimalsystem) vollzogen und es

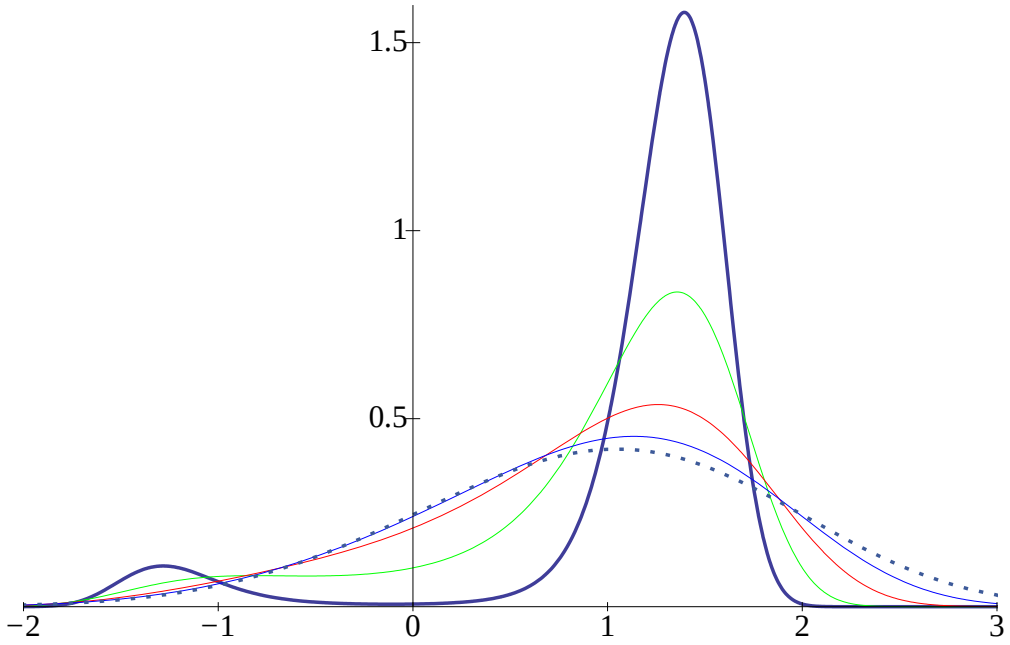


Abbildung 5.1.: Dichten der a-posteriori-Verteilung im Modell (5.11) mit $\bar{y} = 1$, $\mu = 1$ und für $\sigma_0^2/N = 0.1$ (fett), 0.3, 1, 3, 10 (gepunktet)

zeigte sich jedes Mal das gleiche Bild. Zum anderen ist das Verhalten für $\sigma_0^2/N \to 0$ nicht gemäß des Verhaltens des Modalwertes.

In Abbildung 5.3 finden sich noch weitere Rechnungen für $\sigma_0^2/N = 0.5$ bzw. $\sigma_0^2/N = 0.25$ und jeweils variablem μ. Wieder zeigen sich große Probleme. Daher können wir das Verfahren von Friswell nicht empfehlen. Der Algorithmus basierend auf dem Gauß-Newton-Verfahren und das Verfahren von Pázman liefern beide in allen Fällen den Modalwert.

Ferner sind noch die Schätzungen aufgetragen, die man erhält, wenn man den Median der a-posteriori-Verteilung berechnet, der optimal bezüglich des absoluten Fehlers ist, und die Schätzungen, die man erhält, wenn man den Erwartungswert der a-posteriori-Verteilung berechnet, der optimal bezüglich des quadratischen Fehlers ist. Diese sind numerisch allerdings deutlich schwerer zu bestimmen. In der folgenden Tabelle haben wir noch die Anzahl der nötigen Iterationen angegeben, die für den Algorithmus basierend auf dem Gauß-Newton-Verfahren (GN) und für das Verfahren von Pázman nötig waren, um eine bestimmte Genauigkeit zu erzielen: Zu einer Beobachtung y mit $\bar{y} = 1$ und für $\mu = 0.25, 0.50, \ldots, 2.00$ ist jeweils

$$\min\{k : |\vartheta_k - \widehat{\vartheta}_{\text{Modal}}(y)| < 10^{-7}\}$$

angegeben, wobei $\widehat{\vartheta}_{\text{Modal}}(y)$ die Modalwert-Schätzung für $\sigma_0^2/N = 0.1$, für $\sigma_0^2/N = 0.25$, für $\sigma_0^2/N = 0.5$ und für $\sigma_0^2 = 1$ ist.

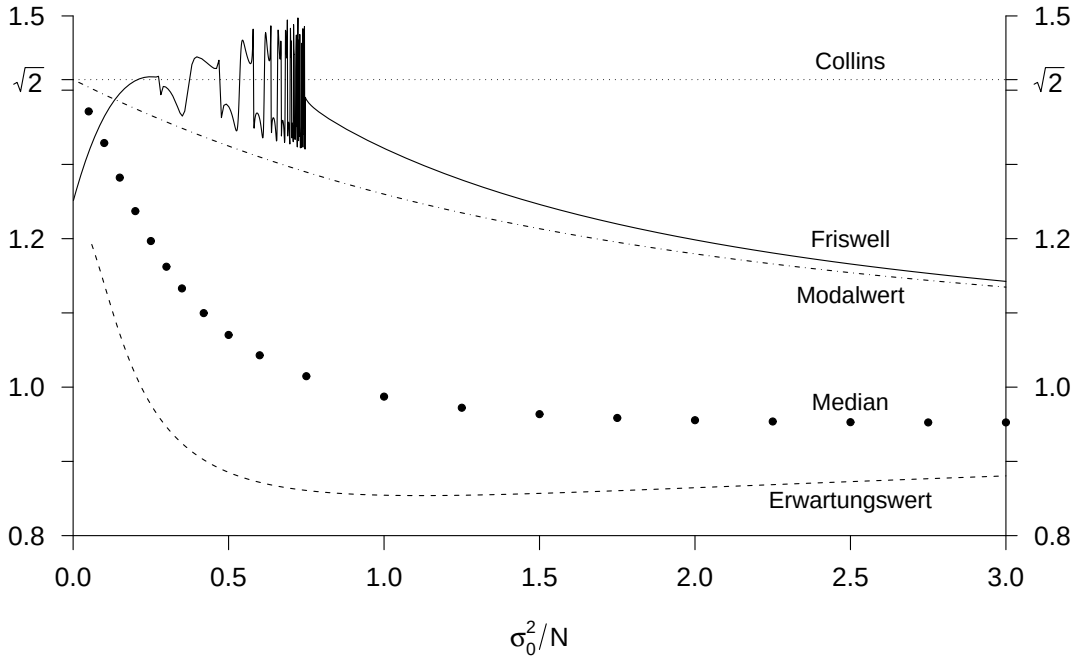


Abbildung 5.2.: Median, Erwartungswert und Modalwert der a-posteriori-Verteilung im Modell (5.11) mit $\bar{y} = 1$ und $\mu = 1$ in Abhängigkeit von σ_0^2/N und Ergebnisse des Verfahrens von Collins und des Verfahrens von Friswell; die Verfahren von Pazman, das Verfahren basierend auf Gauß-Newton und ein BFGS-Verfahren liefern jeweils den Modalwert

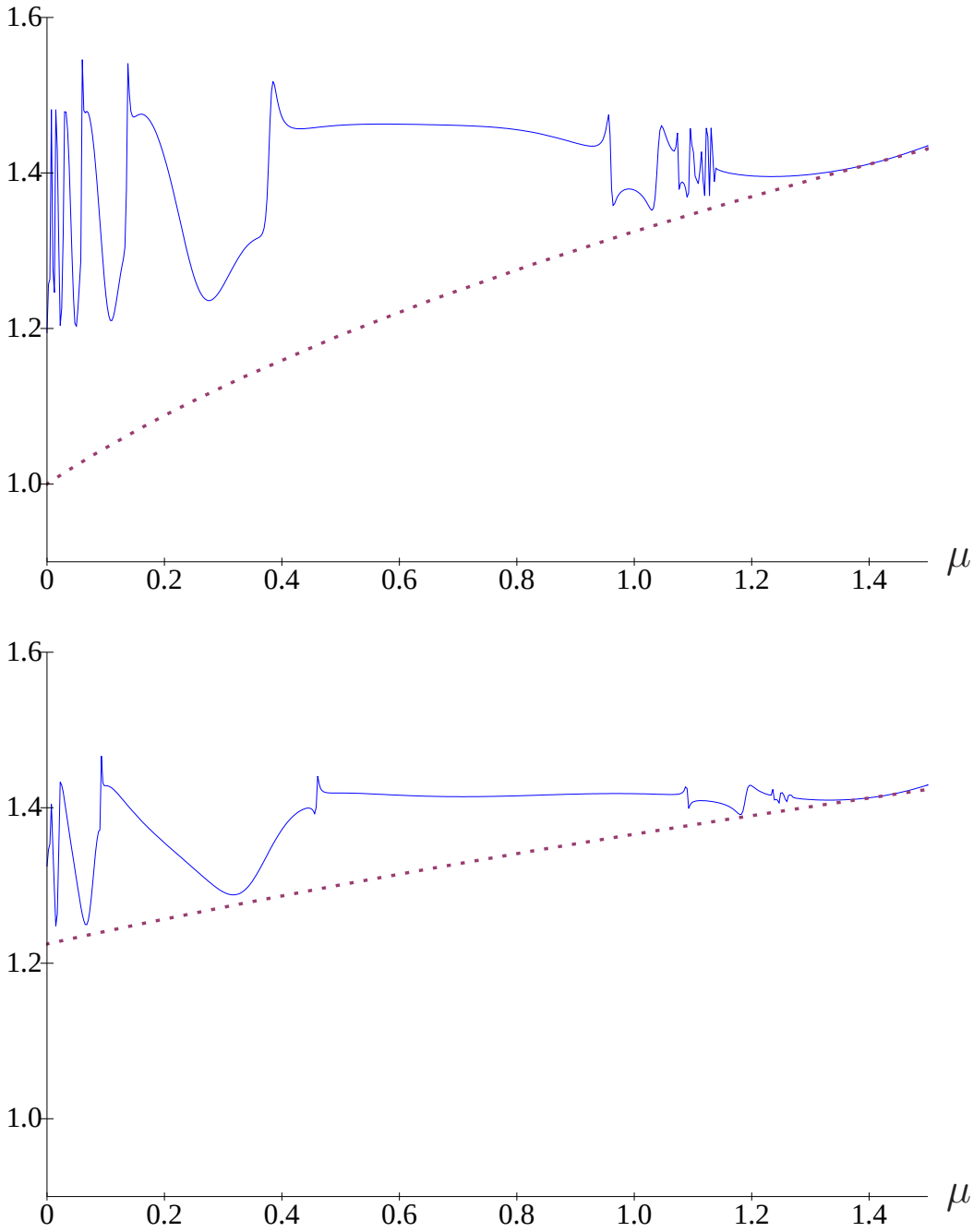


Abbildung 5.3.: Modalwert der a-posteriori-Verteilung (gepunktet) im Modell (5.11) und Ergebnisse des Verfahrens von Friswell (durchgezogen) mit $\bar{y} = 1$ und $\sigma_0^2/N = 0.5$ (oben) bzw. $\sigma_0^2/N = 0.25$ (unten) jeweils in Abhängigkeit von μ

$\sigma_0^2/N = 0.1$:

μ	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
GN	7	6	6	5	4	3	4	5
Pázman	7	6	6	5	5	5	6	6

$\sigma_0^2/N = 0.25$:

μ	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
GN	8	7	6	5	4	3	5	5
Pázman	8	7	7	7	7	7	7	8

$\sigma_0^2/N = 0.5$:

μ	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
GN	12	9	6	5	4	3	5	6
Pázman	9	9	9	10	10	10	11	11

$\sigma_0^2/N = 1$:

μ	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00
GN	19	12	8	6	4	3	5	6
Pázman	24	23	21	22	21	20	21	21

Der Algorithmus basierend auf dem Gauß-Newton-Verfahren erzielt hierbei in der Regel deutlich schneller die gewünschte Genauigkeit, so dass wir dem Algorithmus von Pázman auch keine Empfehlung aussprechen. Das verwendete BFGS-Verfahren und das Gauß-Newton-Verfahren benötigen jeweils fast identische Schrittanzahlen.

5.2. Numerische Verfahren zur Empirischen Bayes-Schätzung

Seien $\eta : (\mathbb{R}^p, \mathcal{B}^p) \to (\mathbb{R}^N, \mathcal{B}^N)$, $\xi \in W \subseteq \mathbb{R}^w$, $\mu : (W, \mathcal{B}_W^w) \to (\mathbb{R}^p, \mathcal{B}^p)$, $\Sigma \in \mathbb{R}^{N\times N}$ symmetrisch und positiv definit, $\rho \in V \subseteq \mathbb{R}^v$ und $B : (V, \mathcal{B}_V^v) \to (\mathbb{R}^{p\times p}, \mathcal{B}^{p\times p})$ eine Funktion mit Werten in den symmetrischen und positiv definiten Matrizen. Wir betrachten hier zwei neue Verfahren zur approximativen Berechnung der empirischen Bayes-Schätzung im Modell aus Abschnitt 4.4, d.h. für

$$\begin{aligned} P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathcal{N}(\eta(\vartheta), \Sigma), \\ P^{\boldsymbol{\theta}} &= \mathcal{N}(\mu(\xi), B(\rho)), \end{aligned}$$

die beide auf einer Approximation der Lebesgue-Dichte der Randverteilung von $\boldsymbol{Y}$, also von

$$\bar{h}_Q(y) = \int_{\mathbb{R}^p} f_{\mathcal{N}(\eta(\vartheta),\Sigma)}(y) \cdot f_{\mathcal{N}(\mu(\xi),B(\rho))}(\vartheta)\, \mathrm{d}\boldsymbol{\lambda}^p(\vartheta) \tag{5.12}$$

$$= \frac{1}{(\sqrt{2\pi})^{N+p}\sqrt{\det\Sigma}\sqrt{\det(B(\rho))}} \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}\left\|\eta(\vartheta) - y\right\|_\Sigma^2 - \frac{1}{2}\left\|\vartheta - \mu(\xi)\right\|_{B(\rho)}^2\right) \mathrm{d}\boldsymbol{\lambda}^p(\vartheta)$$

beruhen. Die Dichte aus (5.12) ist zwar im affin-linearen Fall, wie in Abschnitt 4.4 gezeigt, eine Normalverteilungsdichte. Im nichtlinearen Fall gilt dies in der Regel nicht mehr. Die im folgenden präsentierten neue Verfahren basieren auf einer schrittweisen Linearisierung der Modellfunktion bzw. einer Laplace-Approximation des Integrals.

5.2.1. Ein Algorithmus zur approximativen Bestimmung der empirischen Bayes-Schätzung mittels Linearisierung des Modells

Angelehnt an das erfolgreiche Vorgehen mittels Linearisierung im Abschnitt 5.1.3 betrachten wir im $(k+1)$-ten Schritt wieder statt η die Linearisierung

$$\mathfrak{l}_k : \mathbb{R}^p \to \mathbb{R}^N,\ \mathfrak{l}_k(\vartheta) = J_\eta(\vartheta_k) \cdot (\vartheta - \vartheta_k) + \eta(\vartheta_k)$$

und ersetzen die Verteilungsannahme für $P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta}$ durch

$$P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} = \mathcal{N}(\mathfrak{l}_k(\vartheta), \Sigma).$$

Wie bereits in Abschnitt 4.4 festgestellt, gilt für die entsprechende Randverteilung von $\boldsymbol{Y}$ dann

$$P^{\boldsymbol{Y}} = \mathcal{N}(J_\eta(\vartheta_k)(\mu(\xi) - \vartheta_k) + \eta(\vartheta_k), \Sigma + J_\eta(\vartheta_k)B(\rho)J_\eta^{\mathrm{T}}(\vartheta_k))$$

und für die zugehörige Randdichte $\bar{h}_Q$ gilt mit einer Konstanten c

$$\begin{aligned} -2\ln(\bar{h}_Q(y)) = c &+ \ln\det(\Sigma + J_\eta(\vartheta_k)B(\rho)J_\eta^{\mathrm{T}}(\vartheta_k)) \\ &+ \|J_\eta(\vartheta_k)\mu(\xi) - J_\eta(\vartheta_k)\vartheta_k + \eta(\vartheta_k) - y\|^2_{\Sigma + J_\eta(\vartheta_k)B(\rho)J_\eta^{\mathrm{T}}(\vartheta_k)}. \end{aligned}$$

Nach der Bestimmung von Schätzwerten $\widehat{\xi} = \widehat{\xi}(y)$ und $\widehat{\rho} = \widehat{\rho}(y)$ aus der Maximierung von $-2\ln(\bar{h}_Q(y))$ verwenden wir dann die Bayes-Schätzung gemäß (4.27) und setzen

$$\begin{aligned} \vartheta_{\mathrm{Bayes}}(y) = \vartheta_k &+ (J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k) + B^{-1}(\widehat{\rho}))^{-1} \\ &\cdot \left(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}(y - \eta(\vartheta_k)) + B^{-1}(\widehat{\rho})(\mu(\widehat{\xi}) - \vartheta_k)\right). \end{aligned}$$

Dies ergibt folgenden Algorithmus:

Algorithmus 5.7

Input: $\Sigma \in \mathbb{R}^{N\times N}$ positiv definit,
$B : (V, \mathcal{B}_V^v) \to (\mathbb{R}^{p\times p}, \mathcal{B}^{p\times p})$,
$\eta : \mathbb{R}^p \to \mathbb{R}^N$ stetig differenzierbar,
$\mu : (W, \mathcal{B}_W^w) \to (\mathbb{R}^p, \mathcal{B}^p)$,
Startwert $\xi_0 \in W$,
Beobachtung $y \in \mathbb{R}^N$,
Abbruchkriterium.

Output: $\vartheta^* \in \mathbb{R}^p$ (Approximation der empirischen Bayes-Schätzung)

1. Setze $k = 0$ und $\vartheta_0 = \mu(\xi_0)$.
2. Berechne $(\widehat{\xi}, \widehat{\rho}) = (\widehat{\xi}_{(k+1)}, \widehat{\rho}_{(k+1)})$ aus folgendem Minimierungsproblem

$$\begin{aligned} \min_{(\xi,\rho)\in W\times V} &\ln\det(\Sigma + J_\eta(\vartheta_k)B(\rho)J_\eta^{\mathrm{T}}(\vartheta_k)) \\ &+ \|J_\eta(\vartheta_k)\mu(\xi) - J_\eta(\vartheta_k)\vartheta_k + \eta(\vartheta_k) - y\|^2_{\Sigma + J_\eta(\vartheta_k)B(\rho)J_\eta^{\mathrm{T}}(\vartheta_k)}. \end{aligned}$$

3. Löse das lineare Gleichungssystem

$$(J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}J_\eta(\vartheta_k) + B^{-1}(\widehat{\rho}))d_k = J_\eta^{\mathrm{T}}(\vartheta_k)\Sigma^{-1}(y - \eta(\vartheta_k)) + B^{-1}(\widehat{\rho})(\mu(\widehat{\xi}) - \vartheta_k)$$

für d_k und setze $\vartheta_{k+1} = \vartheta_k + d_k$.

4. Falls das Abbruchkriterium erfüllt ist, dann STOP mit Output $\vartheta^* = \vartheta_{k+1}$, ansonsten setze $k := k + 1$ und gehe zu Schritt 2.

Dieser Algorithmus, der von uns erstmals in Offinger (2000) vorgeschlagen wurde, ist ähnlich zu dem Vorgehen, das Lindstrom u. Bates (1990) im nichtlinearen Gemischte-Effekte-Modell vorgeschlagen haben und das von Pinheiro u. Bates (1980) an einigen Beispielen numerisch untersucht wurde und dabei gute Ergebnisse erzielte. Für unseren Algorithmus sind noch keine theoretischen Ergebnisse bekannt, erste numerische Untersuchungen finden sich in Abschnitt 5.2.3.

5.2.2. Zwei Verfahren zur approximativen Bestimmung der empirischen Bayes-Schätzung mittels Laplace-Approximation

Die Methode von Laplace wird meist durch den folgenden Satz motiviert.

Satz 5.8
Seien $A \in \mathcal{B}^p$ mit $\mathrm{int}(A) \neq \emptyset$ und $g, h : (A, \mathcal{B}_A^p) \to (\mathbb{R}, \mathcal{B})$. Weiterhin sei $\mu \in \mathrm{int}(A)$ und es gelte:

a) Für alle $r > 0$ existiert ein $\delta = \delta(r) > 0$, so dass für alle $x \in A$ mit $|x - \mu| \geq r$ gilt

$$g(x) \geq g(\mu) + \delta.$$

b) g sei zweimal in μ differenzierbar und $\mathrm{Hess}\, g(\mu)$ sei positiv definit.

c) h sei stetig in μ und auf A integrierbar.

Dann folgt

$$\lim_{n\to\infty} \sqrt{\left(\frac{n}{2\pi}\right)^p \det(\mathrm{Hess}\, g(\mu))}\, \exp(ng(\mu)) \int_A \exp(-ng(x))h(x)\, \mathrm{d}\boldsymbol{\lambda}^p(x) = h(\mu).$$

Beweis: siehe Witting u. Müller-Funk (1995, Satz B5.8) □

Man kann also unter schwachen Voraussetzungen folgern, dass für eine Minimumstelle μ von g Integrale der Form $\int_A \exp(-ng(x))h(x)\, \mathrm{d}\boldsymbol{\lambda}^p(x)$ asymptotisch nur von $g(\mu)$, $h(\mu)$ und $\det(\mathrm{Hess}\, g(\mu))$ abhängen und für großes n gilt

$$\int_A \exp(-ng(x))h(x)\, \mathrm{d}\boldsymbol{\lambda}^p(x) \approx h(\mu)\exp(-ng(\mu))\sqrt{\frac{(2\pi)^p}{n^p \det(\mathrm{Hess}\, g(\mu))}}\,.$$

Will man diese Approximation nun allerdings für $h \equiv 1$ (und $n = 1$!) anwenden, so ist die Voraussetzung c) nicht erfüllt. Wir erachten daher folgende Motivation (und daher eher die Bezeichnung „Normalverteilungsapproximation“) als geeigneter: Wir betrachten hierzu zunächst $A = \mathbb{R}^p$ nehmen an, dass die zweimal stetig differenzierbare Funktion $g : \mathbb{R}^p \to \mathbb{R}$ ein Minimum in μ besitzt, wobei $\operatorname{Hess} g(\mu)$ positiv definit sei. Sei q die quadratische Approximation von g um die Minimumsstelle μ, d.h.

$$q(x) = g(\mu) + \frac{1}{2}\,(x-\mu)^{\mathrm{T}}\operatorname{Hess} g(\mu)(x-\mu) \quad \text{für } x \in \mathbb{R}^p.$$

Dann gilt

$$\begin{aligned}\int_{\mathbb{R}^p} \exp(-q(x))\,\mathrm{d}\boldsymbol{\lambda}^p(x) &= \exp(-g(\mu)) \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}\,(x-\mu)^{\mathrm{T}}\operatorname{Hess} g(\mu)(x-\mu)\right)\mathrm{d}\boldsymbol{\lambda}^p(x)\\ &= \exp(-g(\mu))\sqrt{\frac{(2\pi)^p}{\det(\operatorname{Hess} g(\mu))}}.\end{aligned}$$

Ist ferner $\mathcal{N}(g(\mu), \operatorname{Hess} g(\mu))(A) \approx 1$, dann ist eine Approximation

$$\int_A \exp(-g(x))\,\mathrm{d}\boldsymbol{\lambda}^p(x) \approx \exp(-g(\mu))\sqrt{\frac{(2\pi)^p}{\det(\operatorname{Hess} g(\mu))}} \tag{5.13}$$

plausibel.

Wir wollen nun eine Approximation wie in (5.13) für die Randdichte $\bar{h}_Q$ durchführen. Bezeichne hierzu

$$W^y_{(\xi,\rho)}(\vartheta) = \|y - \eta(\vartheta)\|^2_\Sigma + \|\vartheta - \mu(\xi)\|^2_{B(\rho)}\,,$$

um die Abhängigkeit von (ξ, ρ) deutlich zu machen, und

$$\vartheta^*(\xi,\rho) = \operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^p} W^y_{(\xi,\rho)}(\vartheta).$$

Dann ist

$$\begin{aligned}&\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}\,\|\eta(\vartheta) - y\|^2_\Sigma - \frac{1}{2}\,\|\vartheta - \mu(\xi)\|^2_{B(\rho)}\right)\mathrm{d}\boldsymbol{\lambda}^p(\vartheta)\\ &\qquad\approx \frac{(2\pi)^{p/2}}{\sqrt{\det\left(\frac{1}{2}\operatorname{Hess} W^y_{(\xi,\rho)}(\vartheta^*(\xi,\rho))\right)}}\,\exp\big(-\frac{1}{2}W^y_{(\xi,\rho)}(\vartheta^*(\xi,\rho))\big)\end{aligned}$$

und somit

$$\bar{h}_Q(y) \approx \frac{\exp\big(-\frac{1}{2}W^y_{(\xi,\rho)}(\vartheta^*(\xi,\rho))\big)}{\sqrt{(2\pi)^N}\sqrt{\det(\Sigma)}\sqrt{\det(B(\rho))}\sqrt{\det\left(\frac{1}{2}\operatorname{Hess} W^y_{(\xi,\rho)}\vartheta^*(\xi,\rho)\right)}}$$

bzw. für eine Konstante c

$$-2\ln(\bar{h}_Q(y)) \approx c + \ln(\det(\Sigma)) + \ln\det(B(\rho)) + \ln\big(\det\big(\frac{1}{2}\mathrm{Hess}\, W^y_{(\xi,\rho)}(\vartheta^*(\xi,\rho))\big)\big) + W^y_{(\xi,\rho)}(\vartheta^*(\xi,\rho)).$$

Es ergibt sich daher das Minimierungsproblem

$$\min_{(\xi,\rho)\in W\times V} \ln\det(B(\rho)) + \ln\det\big(B^{-1}(\rho) + J_\eta^{\mathrm{T}}(\vartheta^*(\xi,\rho))\Sigma^{-1}J_\eta(\vartheta^*(\xi,\rho)) + H_\eta(\vartheta^*(\xi,\rho))\big) + \|y-\eta(\vartheta^*(\xi,\rho))\|^2_\Sigma + \|\vartheta^*(\xi,\rho)-\mu(\xi)\|^2_{B(\rho)}$$

bzw. äquivalent

$$\min_{(\xi,\rho)\in W\times V} \ln\det\big(I_p + B(\rho)J_\eta^{\mathrm{T}}(\vartheta^*(\xi,\rho))\Sigma^{-1}J_\eta(\vartheta^*(\xi,\rho)) + B(\rho)H_\eta(\vartheta^*(\xi,\rho))\big) + \|y-\eta(\vartheta^*(\xi,\rho))\|^2_\Sigma + \|\vartheta^*(\xi,\rho)-\mu(\xi)\|^2_{B(\rho)}\,. \tag{5.14}$$

Verwendet man noch die Approximation aus der Gauß-Newton-Idee und vernachlässigt $H_\eta(\vartheta^*(\xi,\rho))$, so erhalten wir insgesamt das Minimierungsproblem

$$\min_{(\xi,\rho)\in W\times V} \ln\det\big(I_p + B(\rho)J_\eta^{\mathrm{T}}(\vartheta^*(\xi,\rho))\Sigma^{-1}J_\eta(\vartheta^*(\xi,\rho))\big) + \|y-\eta(\vartheta^*(\xi,\rho))\|^2_\Sigma + \|\vartheta^*(\xi,\rho)-\mu(\xi)\|^2_{B(\rho)}\,, \tag{5.15}$$

wobei

$$\vartheta^*(\xi,\rho) = \operatorname*{arg\,min}_{\vartheta\in\mathbb{R}^p} \|y-\eta(\vartheta)\|^2_\Sigma + \|\vartheta-\mu(\xi)\|^2_{B(\rho)}\,. \tag{5.16}$$

Mit der Lösung $(\widehat{\xi},\widehat{\rho}) = (\widehat{\xi}(y),\widehat{\rho}(y)) \in W\times V$ dieses Minimierungsproblems ergibt sich dann die empirische Bayes-Schätzung zu $\vartheta^*(\widehat{\xi},\widehat{\rho})$.

Für diese Approximationen[2] sind ebenfalls noch keine theoretischen Ergebnisse bekannt, erste numerische Untersuchungen finden sich in dem nächsten Abschnitt 5.2.3.

Die Laplace-Approximation wird häufig in der Bayesschen Analyse der a-posteriori-Verteilung verwendet, siehe z.B. Tierney u. Kadane (1986) und Robert (2007). In den numerischen Untersuchungen von Pinheiro u. Bates (1980) in verschiedenen nichtlinearen Gemischte-Effekte-Modellen erzielte die Laplace-Approximation ebenfalls gute Ergebnisse. Über Probleme für hohe Dimension berichten allerdings Shun u. McCullagh (1995).

Bemerkung 5.9
Neben den vorgestellten approximativen Verfahren erscheint uns ein Zugang über sphärische Integration, siehe Mysovskikh (1980), Monahan u. Genz (1997) und Monahan (2001,

[2] Als weitere Idee bietet sich an, das Minimierungsproblem (5.16) jeweils nicht „exakt“ zu lösen, sondern sich lediglich mit einer Verbesserung gegenüber der vorigen Minimierung zufrieden zu geben. Wir haben diese Idee noch nicht weiter verfolgt.

Section 12.8), oder eine approximative Lösung mittels Monte-Carlo-Simulation unter Verwendung von Importance-Sampling, siehe z.B. Müller-Gronbach u. a. (2010) und Evans u. Swartz (1995, 2000), vielversprechend zu sein. Wir planen beide Methoden in weiteren Untersuchungen zum Vergleich zu den hier vorgestellten Approximationen zu verwenden.

5.2.3. Vergleich der Approximationen an einem Beispiel

In den vergangenen beiden Abschnitten 5.2.1 und 5.2.2 haben wir eine Approximation durch Linearisierung (siehe Algorithmus 5.7) ein Verfahren basierend auf einer Laplace-Approximation (siehe (5.14)) und ein Verfahren basierend auf einer Laplace-Approximation und der Gauß-Newton-Idee (siehe (5.15)) vorgestellt.

Diese drei Verfahren wurden im Modell

$$\begin{aligned} P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathcal{N}(\eta(\vartheta), \sigma_0^2 I_N), \\ P^{\boldsymbol{\theta}} &= \mathcal{N}(1, \rho^2) \end{aligned} \tag{5.17}$$

mit

$$\eta : \mathbb{R} \to \mathbb{R}^N,\ \eta(\vartheta) = \frac{1}{2}\, \mathbf{1}_N\, \vartheta^2$$

und bekanntem Parameter $\sigma_0^2 > 0$ und unbekanntem Parameter $\rho > 0$ untersucht und mit einem deutlich aufwendigeren Verfahren unter Zuhilfenahme von deterministischer numerischer Integration für die jeweilige Berechnung der Randdichte verglichen.

Für eine Beobachtung y mit $\bar{y} = 1$ sind die entsprechenden Ergebnisse der empirischen Bayes-Schätzung für ϑ, die die vier Verfahren liefern, in Abhängigkeit von σ_0^2/N in Abbildung 5.4 aufgetragen.

Obwohl die drei Approximationen mit deutlich weniger Rechenaufwand verbunden waren, liefern sie recht ähnliche Ergebnisse und erscheinen daher alle Erfolg versprechend. Vor allem aber die Laplace-Approximation ohne die Vernachlässigung des H_η-Terms, siehe (5.14), stimmt hervorragend mit den Ergebnissen aus der numerischen Integration überein, während im Bereich zwischen $\sigma_0^2/N = 0.2$ und $\sigma_0^2/N = 0.5$ eine Abweichung in den Ergebnissen des Verfahrens basierend auf einer Linearisierung (Algorithmus 5.7) und der Laplace-Approximation in Kombination mit der Gauß-Newton-Idee, siehe (5.15), zu sehen ist. Diese beiden Ansätze verhalten sich in diesem Beispiel sehr ähnlich. Da man in praktischen Anwendungen die Bestimmung der Hesse-Matrizen von η_i, $i = 1, \ldots, N$, aufgrund des damit verbundenen Rechenaufwands oftmals vermeiden möchte, bleiben die Approximation durch die Linearisierung und die Laplace-Approximation unter Vernachlässigung von H_η trotz der schlechteren Ergebnisse praktisch relevant.

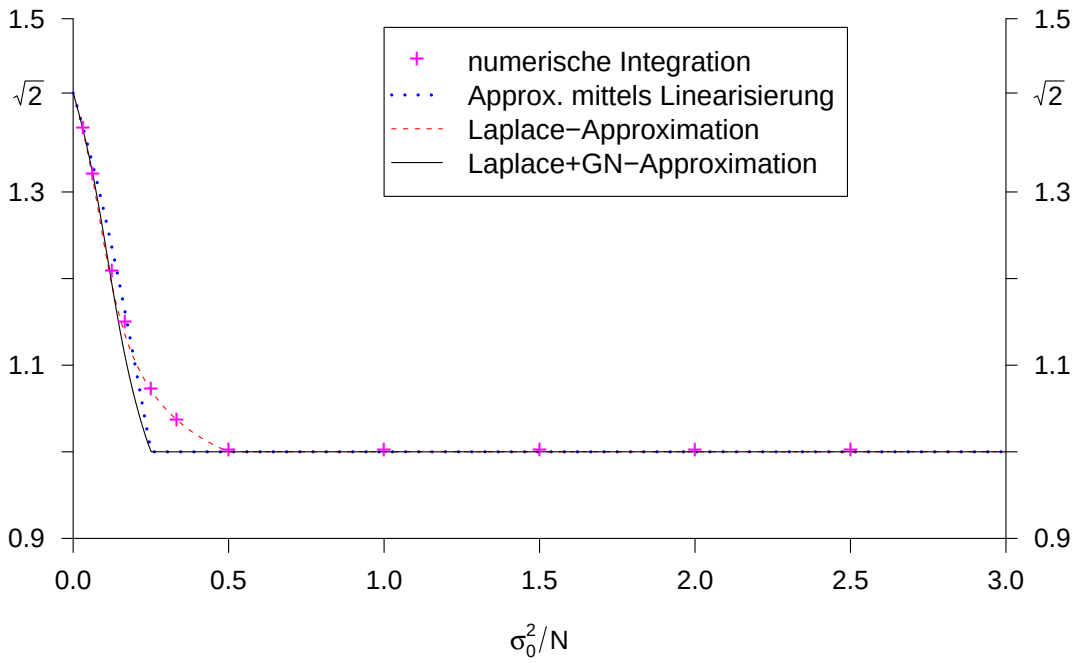


Abbildung 5.4.: Ergebnisse der empirischen Bayes-Schätzung für ϑ im Modell (5.17) zu einer Beobachtung y mit $\bar{y} = 1$ in Abhängigkeit von σ_0^2/N
a) für ein Verfahren basierend auf numerischer Integration,
b) für das Verfahren basierend auf einer Linearisierung (Alg. 5.7),
c) für den Ansatz (5.14) mittels Laplace-Approximation und
d) für den Ansatz (5.15) mittels Laplace-Approximation in Kombination mit der Gauß-Newton-Idee

6. Anwendungen

In diesem Abschnitt werden wir im Detail zwei Anwendungen aus den Ingenieurwissenschaften vorstellen. Die erste Anwendung beschreibt Kriechversuche, bei der zweiten betrachten wir strukturmechanische Probleme.

6.1. Kriechversuche

Beim einachsigen Zeitstandversuch unter Zugbeanspruchung nach EN 10291:2000 bzw. DIN 50118 wird eine Materialprobe der Länge l_0 eingespannt und auf die gewünschte (konstante) Temperatur T erhitzt, vgl. Abbildung 6.1. Ab der Zeit $t = 0$ wird sie dann einer konstanten Spannung $\boldsymbol{\sigma}$ = Kraft/Querschnitt ausgesetzt. Sofort nach dem Anlegen der Spannung ergibt sich eine Anfangsdehnung, so dass die Probe eine Länge von $l(0)$ aufweist, wobei $l(0) > l_0$. Die Länge zur Zeit t bezeichnet man mit $l(t)$. Die relative (technische) Kriechdehnung zur Zeit t ist als

$$\varepsilon_{\mathrm{rel}}(t) = \frac{l(t) - l_0}{l_0}$$

definiert. Daneben wird auch die logarithmische oder natürliche Kriechdehnung betrachtet, die durch

$$\varepsilon(t) = \ln\left(\frac{l(t)}{l(0)}\right)$$

definiert ist. Diese ist zwar im ersten Moment weniger anschaulich, hat aber angenehmere Eigenschaften, vgl. Evans u. Wilshire (1993, S. 15).

Eine Modellierung der natürlichen Kriechdehnung $\varepsilon(t)$ zur Zeit t für biegsame (duktile) Materialien basierend auf einer Theorie der Vorgänge auf Mikroebene ist gemäß Evans u. Wilshire (1993, Chapter 6 und Appendix A) gegeben durch:

$$\varepsilon(t) = \vartheta_1(1 - e^{-\vartheta_2 t}) + \vartheta_3(e^{\vartheta_4 t} - 1) \tag{6.1}$$

mit $\vartheta_i > 0$ für $i = 1, 2, 3, 4$.
Ein typischer Verlauf der Kriechdehnung bei einer hohen Temperatur (zwischen $0.4T_m$ und $0.8T_m$, wenn T_m den Schmelzpunkt in Kelvin bezeichnet) ist in Abbildung 6.2 zu finden.

Die primäre Phase (I) wird dabei vor allem durch die Parameter ϑ_1 und ϑ_2 charakterisiert, wobei ϑ_2 die Krümmung bestimmt und ϑ_1 für die Skalierung sorgt. Die tertiäre Phase (III) wird hingegen hauptsächlich von den Parametern ϑ_3 und ϑ_4 bestimmt, wobei ϑ_4 vor allem die Krümmung und ϑ_3 die Skalierung beeinflusst. Die Parameter $\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4$ können dabei mit Materialkonstanten, Fließspannungen, der ange-

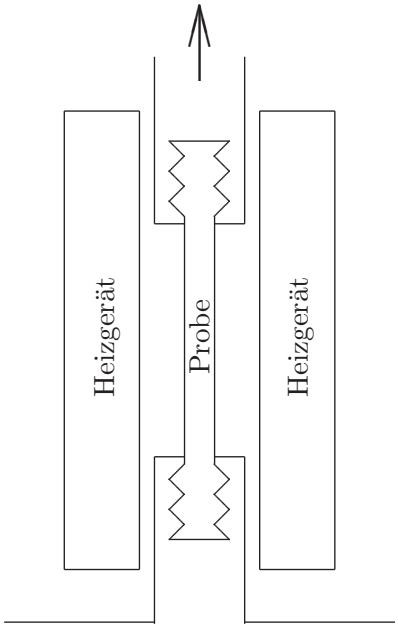


Abbildung 6.1.: schematische Darstellung eines Zugversuchs

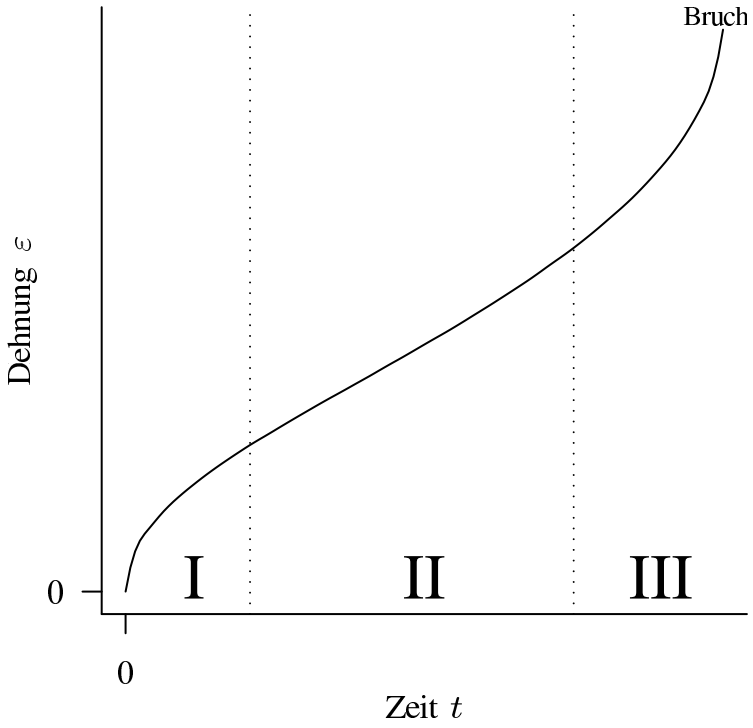


Abbildung 6.2.: typische Kriechkurve

Zeit t_i	0	60	180	300	420	540	660
Kriechdehnung $\varepsilon(t_i)$	0.000	0.227	0.413	0.531	0.628	0.711	0.788
Zeit t	780	900	1020	1140	1260	1380	1500
Kriechdehnung $\varepsilon(t_i)$	0.859	0.926	0.990	1.049	1.109	1.168	1.226
Zeit t_i	1620	1740	1860	1980	2100	2220	2340
Kriechdehnung $\varepsilon(t_i)$	1.283	1.338	1.394	1.450	1.508	1.564	1.622
Zeit t_i	2460	2580	2700	2820	2940	3060	3180
Kriechdehnung $\varepsilon(t_i)$	1.680	1.737	1.797	1.859	1.921	1.985	2.050
Zeit t_i	3300	3420	3540	3660	3780	3900	4020
Kriechdehnung $\varepsilon(t_i)$	2.118	2.190	2.260	2.336	2.417	2.503	2.591
Zeit t_i	4140	4260	4380	4500	4620	4740	4800
Kriechdehnung $\varepsilon(t_i)$	2.689	2.797	2.915	3.057	3.230	3.494	3.756

Tabelle 6.1.: Messungen der Kriechspannung für polykristallines Kupfer bei einer Spannung von $\boldsymbol{\sigma} = 76\,MN\,m^{-2}$ und einer Temperatur von 608 Kelvin nach Evans u. Wilshire (1993, Table B1)

legten Kraft, der Temperatur, der Querschnittsfläche und Aktivierungsenergien in Beziehung gestellt werden, siehe Evans u. Wilshire (1993, S. 84) Wir werden uns hier mit der Schätzung von $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4)$ zufrieden geben.

Bei spröden Materialien hingegen ist die tertiäre Phase bis zum Materialversagen recht kurz. Um daher Probleme mit der Parameteridentifizierung zu vermeiden, schlagen Evans u. Wilshire (1993, Appendix A) folgende Modellierung für spröde Materialien vor:[1]

$$\tilde{\varepsilon}(t) = \vartheta_1(1 - e^{-\vartheta_2 t}) + \vartheta_5 t. \tag{6.2}$$

Die Kriechantwort des Systems wird zu den Zeiten t_i gemessen, d.h. für biegsame Materialien ist die Modellfunktion $\eta : [0, \infty)^4 \to \mathbb{R}^N$ durch

$$\eta_i(\vartheta) = \varepsilon(t_i) = \vartheta_1(1 - \exp(-\vartheta_2 t_i)) + \vartheta_3(\exp(\vartheta_4 t_i) - 1) \quad \text{für } i = 1, \ldots, N$$

gegeben bzw. für spröde Materialien ist die Modellfunktion $\tilde{\eta} : [0, \infty)^3 \to \mathbb{R}^N$ durch

$$\tilde{\eta}_i(\vartheta) = \tilde{\varepsilon}(t_i) = \vartheta_1(1 - \exp(-\vartheta_2 t_i)) + \vartheta_5 t_i \quad \text{für } i = 1, \ldots, N$$

gegeben.

Wir untersuchen den Datensatz in Tabelle 6.1 aus Evans u. Wilshire (1993, Table B1), die Messungen bei einer Spannung von $\boldsymbol{\sigma} = 76\,MN\,m^{-2}$ und einer Temperatur von 608 Kelvin für polykristallines Kupfer auflisten.

Zuerst betrachten wir ein Modell ohne a-priori-Annahme für biegsame Materialien:

$$\boldsymbol{Y} \sim \mathcal{N}(\eta(\vartheta), \sigma I_N)$$

[1] Linearisiert man den Term $e^{\vartheta_4 t} - 1$ um $t = 0$, d.h. ersetzt ihn durch $\vartheta_4 t$ und ferner $\vartheta_3\vartheta_4$ durch einen neuen Parameter ϑ_5, so erhält man aus Gleichung (6.1) die Gleichung (6.2).

mit $\sigma > 0$ unbekannt.
In diesem Fall erhalten wir bei einem Startwert $\vartheta_0 = (2, 0.002, 0.1, 0.001)$ mit dem Gauß-Newton-Verfahren nach 19 Schritten[2] folgende Schätzung für die Daten aus Tabelle 6.1:

$$\widehat{\vartheta}(y) = (1.2198,\ 0.0013617,\ 0.14265,\ 0.00059474).$$

Betrachtet man stattdessen mit

$$\boldsymbol{Y} \sim \mathcal{N}(\tilde{\eta}(\vartheta), \sigma I_N)$$

ein Modell ohne a-priori-Annahme für spröde Materialien, so ergibt sich bei einem Startwert von $\vartheta_0 = (0.5, 0.01, 0.001)$ nach 8 Iterationen[3] des Gauß-Newton-Verfahrens folgende Schätzung

$$\widehat{\vartheta}(y) = (0.30183,\ 0.018232,\ 0.00059432).$$

Als Nächstes betrachten wir ein Modell mit a-priori-Annahme für biegsame Materialien:

$$P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} = \mathcal{N}(\eta(\vartheta), \sigma^2 I_N),$$
$$P^{\boldsymbol{\theta}} = \mathcal{N}(\mu, B),$$

wobei $\sigma^2 = 0.01$ und $\mu = (2,\ 0.002,\ 0.1,\ 0.001)$ bekannt sind und

$$B = \operatorname{diag}(1^2,\ 0.001^2,\ 0.05^2,\ 0.0005^2) = \begin{pmatrix} 1^2 & 0 & 0 & 0 \\ 0 & 0.001^2 & 0 & 0 \\ 0 & 0 & 0.05^2 & 0 \\ 0 & 0 & 0 & 0.0005^2 \end{pmatrix} \in \mathbb{R}^{4\times 4}.$$

Hier ergibt sich für die Modalwertschätzung

$$\widehat{\vartheta}(y) = (1.3042,\ 0.0012404,\ 0.10465,\ 0.00065035).$$

Setzt man stattdessen das Modell

$$P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} = \mathcal{N}(\tilde{\eta}(\vartheta), \sigma^2 I_N),$$
$$P^{\boldsymbol{\theta}} = \mathcal{N}(\mu, B),$$

für spröde Materialien mit $\sigma^2 = 0.01$ und $\mu = (1, 0.01, 0.001)$ und

$$B = \operatorname{diag}(0.5^2,\ 0.005^2,\ 0.0005^2)$$

an, so ergibt sich für die Modalwertschätzung

$$\widehat{\vartheta}(y) = (0.30969,\ 0.011890,\ 0.00059199).$$

Die Kurven sind zusammen mit den Datenpunkten in Abbildung 6.3 dargestellt. Die Kurven mit und ohne a-priori-Annahme unterscheiden sich jeweils nicht wesentlich, die Modellfunktion für spröde Materialien passt für Kupfer nicht zu den Datenpunkten.

Für weiterführende Literatur zum ingenieurwissenschaftlichen Hintergrund dieses Themas verweisen wir auf Evans u. Wilshire (1993) und Altenbach (1999).

[2] Zum Vergleich: Das Verfahren NL2SOL, siehe S. 95, benötigt bei diesem Startwert 26 Schritte.
[3] Zum Vergleich: Das Verfahren NL2SOL, siehe S. 95, benötigt bei diesem Startwert 11 Schritte.

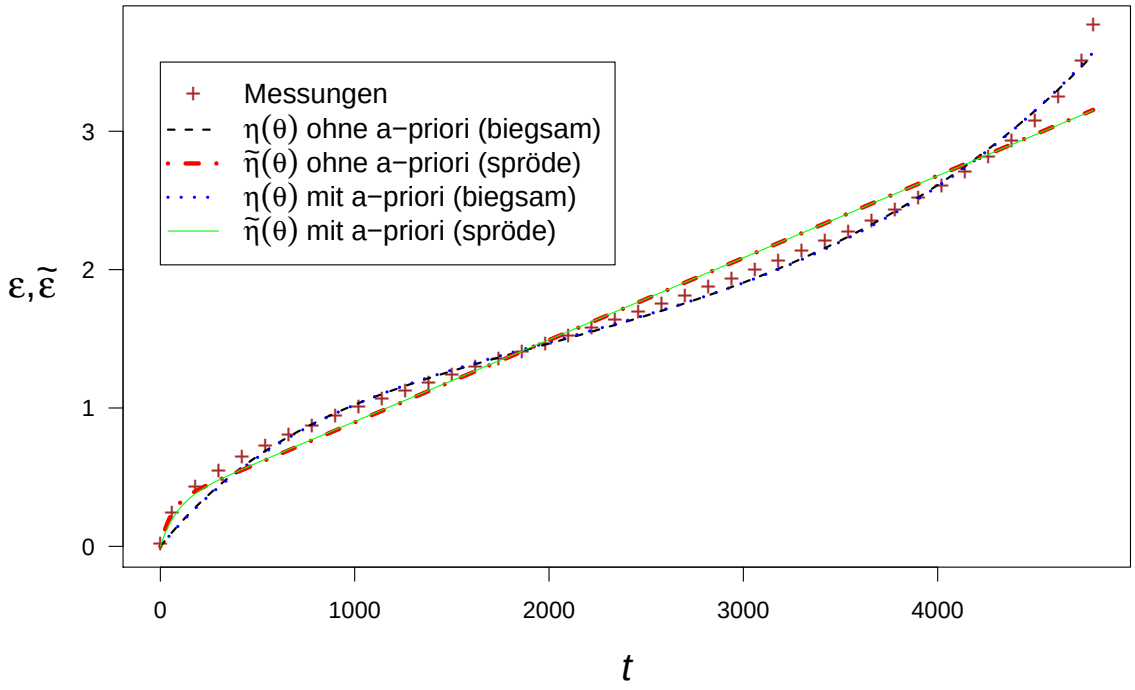


Abbildung 6.3.: Messdaten und geschätzte Kriechkurven

6.2. Finite-Elemente-Modelle für Eigenwertprobleme

6.2.1. Eigenwertprobleme bei partiellen Differentialgleichungen

Wir geben in diesem einführenden Abschnitt einen kleinen Einblick in die Finite-Elemente-Methode zur Approximation von Eigenwertproblemen bei gewöhnlichen oder partiellen Differentialgleichungen. In unserer Darstellung folgen wir hierbei Babuška u. Osborn (1991, Chapter 1), auf die wir ebenso wie auf Tveito u. Winther (2005, Section 2.4) für einen tieferen Einstieg in das Thema verweisen. Für eine allgemeine Einführung in die Finite-Elemente-Methode empfehlen wir z.B. Göring u. a. (1993).

Eigenwertprobleme bei Differentialgleichungen kommen in der Physik und den Ingenieurwissenschaften häufig vor. Bei der Überführung der Differentialgleichung in eine Variationsgleichung erhält man für einen unendlich-dimensionalen Funktionenraum V und für symmetrische und positiv definite Bilinearformen a und b auf $V \times V$ das folgende Problem:

Gesucht sind Paare $(\lambda, u) \in (0, \infty) \times V$ mit $u \not\equiv 0$, so dass

$$a(u, v) = \lambda\, b(u, v) \quad \text{für alle } v \in V.$$

Approximiert man mit der Finite-Elemente-Methode den Funktionenraum V durch einen endlich-dimensionalen Funktionenraum V_h mit einer Basis $\phi_1, \ldots, \phi_n$, so erhält man

folgendes Problem:

Gesucht sind Paare $(\lambda_h, u_h) \in (0,\infty) \times V_h$ mit $u_h \not\equiv 0$, so dass

$$a(u_h, v_h) = \lambda_h\, b(u_h, v_h) \quad \text{für alle } v_h \in V_h. \tag{6.3}$$

Setzt man

$$A = (a_{ij})_{1\le i,j\le n} \quad \text{mit } a_{ij} = a(\phi_i, \phi_j) \quad \text{und}$$
$$B = (b_{ij})_{1\le i,j\le n} \quad \text{mit } b_{ij} = b(\phi_i, \phi_j),$$

dann kann man das Problem äquivalent durch das folgende Problem ersetzen:

Gesucht sind Paare $(\lambda_h, z) \in (0,\infty) \times \mathbb{R}^n$ mit $z \neq \mathbf{0}_n$, so dass

$$A\, z = \lambda_h\, B\, z.$$

Für jedes solche Paar (λ_h, z) ist dann $(\lambda_h, \sum_{i=1}^n z_i\phi_i)$ eine Lösung des Problems (6.3) und umgekehrt.

Die Bilinearformen a und b bzw. die symmetrischen und positiv definiten Matrizen A und B hängen dabei meist noch von physikalischen Parametern $\vartheta_1, \ldots, \vartheta_p \in \mathbb{R}$ ab, d.h. mit $\vartheta = (\vartheta_1, \ldots, \vartheta_p)$ ergeben sich dann Probleme der Form

$$A(\vartheta)\, z = \lambda_h\, B(\vartheta)\, z$$

und man sucht zu gegebenem $\vartheta \in \mathbb{R}^p$ entsprechend Paare $(\lambda_h(\vartheta), z(\vartheta)) \in (0,\infty) \times \mathbb{R}^n$.

Wir werden im Folgenden exemplarisch die Finite-Elemente-Approximation eines Schwingungsproblems vorstellen, bei der man auch die Diskretisierung eigenständig interpretieren kann.

6.2.2. Lineare Schwingungen

Wir stellen in diesem Abschnitt einen eindimensionalen Massen-Feder-Schwinger mit n Massen $m_1, \ldots, m_n > 0$ und mit $n+1$ Steifigkeiten $k_1, \ldots, k_{n+1} > 0$ vor, siehe Abbildung 6.4, den man auch als Diskretisierung eines gewichtslosen linear elastischen Stabes auffassen kann, der das Hookesche Gesetz erfüllt, vergleiche Babuška u. Osborn (1991, Section 1.5).

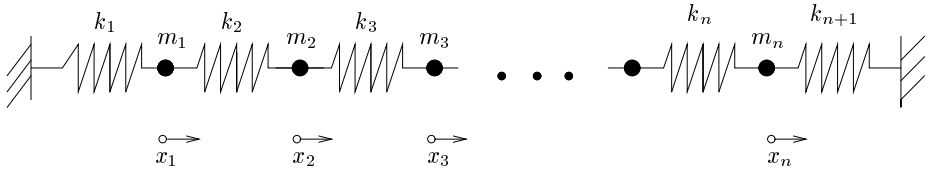


Abbildung 6.4.: gebundener n-Massen-Feder-Schwinger in Reihenschaltung

Wir nehmen an, dass man Reibung vernachlässigen kann[4] und keine externen Kräfte vorliegen und bezeichnen mit $x_i(t)$, $i = 1, \ldots, n$, die Auslenkung der i-ten Masse zur Zeit t gegenüber ihren jeweiligen zeitunabhängigen Ruhelagen und mit $x = x(t) = (x_1(t), \ldots, x_n(t))^{\mathrm{T}}$ den Ortsvektor. Dann sind die Newtonschen Bewegungsgleichungen von der Form

$$M(\vartheta)\ddot{x} + K(\vartheta)x = 0,$$

wobei $\ddot{x}$ die 2. Ableitung des Ortsvektors x nach t bezeichnet und die **Massenmatrix** sich zu

$$M(\vartheta) = \operatorname{diag}(m_1, \ldots, m_n) \tag{6.4}$$

und die **Steifigkeitsmatrix** sich zu

$$K(\vartheta) = \begin{bmatrix} k_1 + k_2 & -k_2 & & & & & \\ -k_2 & k_2 + k_3 & -k_3 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & -k_{n-1} & k_{n-1} + k_n & -k_n & \\ & & & & -k_n & k_n + k_{n+1} \end{bmatrix} \tag{6.5}$$

ergibt,[5] wobei $k_i = k_i(\vartheta)$, $i = 1, \ldots, n+1$, und $m_i = m_i(\vartheta)$, $i = 1, \ldots, n$. Zum Beispiel kann der Strukturparameter ϑ dabei Komponenten der Massenmatrix und der Steifigkeitsmatrix enthalten, soweit diese unbekannt sind.[6] Werden die Massen als bekannt vorausgesetzt und alle Steifigkeiten ohne weitere Einschränkungen als unbekannt, so ist der Strukturparameter von der Form $\vartheta = (k_1, \ldots, k_{n+1})^{\mathrm{T}} \in (0, \infty)^{n+1}$. Wie man leicht sieht, ist für $k_i > 0$, $i = 1, \ldots, n+1$, die Steifigkeitsmatrix $K(\vartheta)$ positiv definit,[7] die Massenmatrix ist für $m_i > 0$, $i = 1, \ldots, n$, trivialerweise positiv definit.

Wie bereits im Abschnitt 1.3.3 erläutert, überführt man diese Differentialgleichung durch den Ansatz $x(t) = ve^{i\omega t}$ in das strukturelle Eigenwertproblem

$$K(\vartheta)v = \lambda M(\vartheta)v \tag{6.6}$$

[4] Für die Behandlung von Systemen mit Reibung, die ohne externe Kräfte von der Form

$$M(\vartheta)\ddot{x} + C\dot{x} + K(\vartheta)x = 0$$

sind, vergleiche Huppert (1990, Section III.6).

[5] Dieses Problem weist somit die typischen Eigenschaften von Problemen auf, die aus der Finite-Elemente-Diskretisierung von Eigenwertproblemen aus Abschnitt 6.2.1 entstehen, indem die Massenmatrix und die Steifigkeitsmatrix Bandstruktur haben und symmetrisch und positiv definit sind.

[6] In den Ingenieurwissenschaften beliebt sind sogenannte Substrukturparameter. Hierbei wird z.B. für die Steifigkeitsmatrix die Form

$$K(\vartheta) = K_0 + \sum_{i=1}^{p} \vartheta_i K_i$$

angesetzt, siehe z.B. Natke (1992).

[7] Für allgemeine „Oszillationsmatrizen“ K mit gebundenen Komponenten siehe hierzu Huppert (1990, Hilfssatz 5.3), bei freien Komponenten ist die Oszillationsmatrix K nur positiv semidefinit.

und sucht nun Paare $(\lambda, v) \in (0, \infty) \times \mathbb{R}^n$ mit $v \neq \mathbf{0}_n$, die dieses verallgemeinerte Eigenwertproblem lösen. Statt des Eigenwerts $\lambda > 0$ betrachtet man in den Ingenieurwissenschaften gerne die Eigenfrequenz $\omega = \sqrt{\lambda}$.

Für eine weiterführende Vertiefung in die mathematische Theorie von linear-elastischen Schwingungen verweisen wir auf Huppert (1990, Chapter 3) und Huppert u. Willems (2006, Abschnitt 8.5) und für eine tiefergehende Betrachtung aus ingenieurwissenschaftlicher Sicht auf Hahn (1992, Kapitel 7) und Clough u. Penzien (2003).

6.2.3. Verteilungsannahmen für die Messungen der Eigenfrequenzen und der Eigenvektoren

Mit Hilfe der experimentellen Modalanalyse erhält man nun Messungen der Eigenfrequenzen aus einem bestimmten Frequenzbereich und einige Komponenten der zugehörigen Eigenvektoren. Man macht für die Messungen gerne[8] die Normalverteilungsannahme

$$\boldsymbol{Y} \sim \mathcal{N}(\eta(\vartheta), \Sigma)$$

mit einer Modellfunktion von der Form

$$\eta : \Theta \to \mathbb{R}^N,\ \eta(\vartheta) = \begin{pmatrix} \lambda_I(\vartheta) \\ v_{LI}(\vartheta) \end{pmatrix}$$

mit $\Theta \subseteq \mathbb{R}^p$ wie in Abschnitt 1.3.3 eingeführt.
Während die Normalverteilungsannahme für die Eigenwertkomponenten plausibel erscheint, ist die Normalverteilungsannahme für die Eigenvektoren bedenklich.

Da die Modalanalyse als Ergebnis normierte[9] Eigenvektoren liefert, erwartet man hierfür eine Verteilung auf der Einheitssphäre oder genauer, wenn man die Mehrdeutigkeit im Vorzeichen auch noch beachtet, sogar eine Verteilung auf dem zugehörigen projektiven Raum.[10]

Eine naheliegende Verteilung auf der Einheitssphäre ist die Fisher-Bingham-Verteilung, die man wie folgt erhält, vergleiche Mardia u. Jupp (2000, Section 9.3.3):
Sei $c \in \mathbb{R}^n$, $A \in \mathrm{PD}(n)$ und für die Zufallsvariable $\boldsymbol{X}$ gelte

$$P^{\boldsymbol{X}} \sim \mathcal{N}(c, A).$$

Dann ist die bedingte Verteilung

$$P^{\boldsymbol{X}|\,\|\boldsymbol{X}\|=1} = \mathrm{FB}(c, A)$$

eine Fisher-Bingham-Verteilung[11], wobei

$$g_{c,A} : S^{n-1} \to \mathbb{R},\ g_{c,A}(y) = \frac{1}{\int_{S^{n-1}} f_{c,A}(t)\,\mathrm{d}\mu_S(t)}\, f_{c,A}(y)$$

[8] Für eine solche Normalverteilungsannahme siehe z.B. Friswell u. Mottershead (1995) und Martin (2001).

[9] Wir nehmen für den Moment an, dass die Eigenvektoren bezüglich der Euklidschen Norm normiert werden, siehe hierzu jedoch Abschnitt 6.2.4.

[10] Zum projektivem Raum $\mathbb{R}P^n$ vergleiche Anhang B.22.

[11] Üblicherweise wird die Fisher-Bingham-Verteilung jedoch anders parametrisiert, siehe Mardia u. Jupp (2000, Section 9.3.3).

eine Dichtefunktion $g_{c,A}$ bezüglich der Gleichverteilung μ_S auf der Einheitssphäre

$$S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$$

ist. Ist $c \in S^{n-1}$, so hängt wegen der Invarianz der Gleichverteilung μ_S auf der Einheitssphäre bezüglich orthogonaler Transformationen das Integral $\kappa(A) = \int_{S^{n-1}} f_{c,A}(t)\,\mathrm{d}\mu_S(t)$ nur von A ab.

Der Einfachheit halber betrachten wir im folgenden nur die Messung der Eigenvektoren. Gegeben sei nun das Modell

$$\boldsymbol{Y} \sim \mathrm{FB}(\eta(\vartheta), \Sigma)$$

mit $\eta : \Theta \to S^{N-1} \subset \mathbb{R}^N$ bei bekannter Matrix $\Sigma \in \mathrm{PD}(N)$. Die Maximum-Likelihood-Schätzung für $\vartheta \in \Theta$ zu einer Beobachtung $y \in S^{N-1}$ führt dann zu dem Kleinste-Quadrate-Problem

$$\arg\min_{\vartheta\in\Theta} \|y - \eta(\vartheta)\|_\Sigma^2 . \tag{6.7}$$

Analog erhalten wir für $\Theta = \mathbb{R}^p$, $\mu \in \mathbb{R}^p$, $B \in \mathrm{PD}(p)$ im Modell

$$\begin{aligned} P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} &= \mathrm{FB}(\eta(\vartheta), \Sigma), \\ P^{\boldsymbol{\theta}} &= \mathcal{N}(\mu, B) \end{aligned}$$

mit Vorinformation wieder als Modalwertschätzung[12] zu einer Beobachtung y das Minimierungsproblem

$$\arg\min_{\vartheta\in\mathbb{R}^p} \|y - \eta(\vartheta)\|_\Sigma^2 + \|\vartheta - \mu\|_B^2 .$$

Beachtet man noch, dass die Eigenvektoren nur bis auf das Vorzeichen bestimmt sind, so wäre

$$\tilde{g}_{\eta(\vartheta),\Sigma} : \mathbb{R}P^n = S^{n-1}/\{\pm\mathrm{id}\} \to \mathbb{R},\ \tilde{g}_{\eta(\vartheta),\Sigma}(\pm x) = \frac{1}{2}\, g_{\eta(\vartheta),\Sigma}(x) + \frac{1}{2}\, g_{\eta(\vartheta),\Sigma}(-x) \tag{6.8}$$

eine Dichtefunktion bezüglich einer Gleichverteilung auf $S^{n-1}/\{\pm\mathrm{id}\}$. Ist $\|\Sigma\|$ hinreichend klein und somit die Normal-Verteilung hinreichend konzentriert, so kann man in (6.8) einen Summanden vernachlässigen und man erhält z.B. im Modell ohne Vorinformation bei der Maximum-Likelihood-Schätzung zu einer Beobachtung $\pm y$ approximativ wieder das Minimierungsproblem

$$\arg\min_{\vartheta\in\Theta} \|y - \eta(\vartheta)\|_\Sigma^2 ,$$

falls y und $\eta(\vartheta)$ in der von Σ induzierten Norm einen Winkel kleiner als $\pi/2$ bilden, d.h. falls $< y, \eta(\vartheta) >_\Sigma > 0$

[12] Wir haben im Abschnitt 4.1 zwar angenommen, dass man Lebesgue-Dichten f_ϑ von P_ϑ hat, aber die Resultate übertragen sich entsprechend auf die Situation $P_\vartheta = f_\vartheta \mu$ für ein σ-endliches Maß μ auf $(\mathbb{R}^N, \mathcal{B}^N)$.

Bemerkung 6.1
Da die Messungen der Eigenvektoren durch Normierung entstehen (und sieht man von der Vorzeichenwahl ab), wäre ein anderes naheliegendes Modell die Messung der Eigenvektoren als Zufallsvariable $\boldsymbol{Y} = \boldsymbol{X}/\|\boldsymbol{X}\|$ mit einer Zufallsvariablen $\boldsymbol{X} \sim \mathcal{N}(\eta(\vartheta), \Sigma)$ aufzufassen. Die so entstehende projizierte Normalverteilung, auch Winkelnormalverteilung genannt, vgl. Watson (1983, Section 3.6), ist jedoch deutlich schwerer zu handhaben; eine geschlossene Form für die Dichten wird hierbei von Pukkila u. Rao (1988) angegeben.

6.2.4. Ableitungen der Eigenwert- und Eigenvektorfunktionen

Betrachtet man zu unserer Modellfunktion

$$\eta : \Theta \to \mathbb{R}^N,\ \eta(\vartheta) = \begin{pmatrix} \lambda_I(\vartheta) \\ v_{LI}(\vartheta) \end{pmatrix}$$

mit $\Theta \subseteq \mathbb{R}^p$ aus Abschnitt 1.3.3 Minimierungsprobleme der Form

$$\arg\min_{\vartheta \in \Theta} \|y - \eta(\vartheta)\|_\Sigma^2$$

und

$$\arg\min_{\vartheta \in \mathbb{R}^p} \|y - \eta(\vartheta)\|_\Sigma^2 + \|\vartheta - \mu\|_B^2\,,$$

so benötigt man hierbei für die von uns vorgestellten Algorithmen die Jacobi-Matrix der Modellfunktion η. Wir widmen uns daher zunächst dem Problem, ob die Eigenwerte und Eigenvektoren eines strukturellen Eigenwertproblems

$$K(\vartheta)\, v = \lambda M(\vartheta)\, v \tag{6.9}$$

in Abhängigkeit von ϑ differenzierbar sind, ehe wir Berechnungsformeln vorstellen, die in der Praxis verwendet werden.

Wir haben bisher stillschweigend angenommen, dass die Normierung der Eigenvektoren bezüglich der Euklidschen Norm geschieht. Üblicher ist in der Praxis jedoch die Normierung bezüglich der Massematrix M. Zu beachten ist jedoch: Nur wenn M nicht vom Strukturparameter ϑ abhängt, übertragen sich durch eine Transformation mit einer Wurzel von M die entsprechenden Überlegungen aus Abschnitt 6.2.3 im Allgemeinen.

Der Beweis zum folgendem Satz folgt dabei dem Vorgehen in Harville (1997, Lemma 21.15.1) für den Fall, dass die Eigenvektoren bezüglich der Euklidschen Norm normiert sind. Eine kurze Vorbemerkung:
Sind für gegebenes ϑ^0 alle Eigenwerte $\lambda_1(\vartheta^0) \leq \ldots \leq \lambda_n(\vartheta^0)$ des strukturellen Eigenwertproblems (6.9) einfach und bezeichnen $v_i(\vartheta)$ die zugehörigen und bezüglich $M(\vartheta)$ normierten Eigenvektoren von (6.9), d.h. gilt

$$v_i^{\mathrm{T}}(\vartheta) M(\vartheta) v_i(\vartheta) = 1 \quad \text{für} \ \ i = 1, \ldots, n\,, \tag{6.10}$$

dann sind dabei bis auf das Vorzeichen in den Eigenvektoren alle Größen eindeutig definiert. Diese Mehrdeutigkeit in den Eigenvektoren kann man auch noch umgehen,

indem man die entsprechenden Äquivalenzklassen betrachtet. Wir werden auf dieses technische Detail in den Ausführungen jedoch verzichten.

Satz 6.2
Seien $K : \mathbb{R}^p \to \mathrm{PD}(n)$ und $M : \mathbb{R}^p \to \mathrm{PD}(n)$ jeweils C^k-differenzierbar nach ϑ in einer Umgebung von ϑ^0 und für gegebenes ϑ^0 seien alle Eigenwerte einfach.
Dann sind in einer Umgebung von ϑ^0 die Eigenwerte $\lambda_1(\vartheta) \leq \ldots \leq \lambda_n(\vartheta)$ und die zugehörigen nach $M(\vartheta)$ normierten Eigenvektoren $v_i(\vartheta)$, $i = 1, \ldots, n$, C^k-differenzierbare Funktionen von ϑ.

Beweis: Betrachte

$$F : \mathbb{R}^p \times (\mathbb{R}^n \times \mathbb{R})^n \to \left(\mathbb{R}^{n+1}\right)^n, (\vartheta, (v_i, \lambda_i)_{i=1}^n) \mapsto \begin{pmatrix} K(\vartheta)v_i - \lambda_i M(\vartheta)v_i \\ v_i^{\mathrm{T}} M(\vartheta) v_i - 1 \end{pmatrix}_{i=1}^n.$$

Dann ist

$$\frac{\partial F\left(\vartheta, (v_i, \lambda_i)_{i=1}^n\right)}{\partial \left(v_i, \lambda_i\right)_{i=1}^n}$$

eine Blockdiagonalmatrix, deren i-ter Block, $i = 1, \ldots, n$, die folgende Form hat:

$$\frac{\partial F}{\partial (v_i, \lambda_i)} \left(\vartheta, (v_i, \lambda_i)_{i=1}^n\right) = \begin{bmatrix} K(\vartheta) - \lambda_i M(\vartheta) & -M(\vartheta)v_i \\ 2v_i^{\mathrm{T}} M(\vartheta) & 0 \end{bmatrix}.$$

Wir zeigen nun, dass der i-te Block, $i = 1, \ldots, n$, (und damit $\frac{\partial F\left(\vartheta, (v_i, \lambda_i)_{i=1}^n\right)}{\partial (v_i, \lambda_i)_{i=1}^n}$) invertierbar ist, falls $(v_i, \lambda_i)_{i=1}^n$ einfache Eigenwerte und normierte Eigenvektoren des strukturellen Eigenwertproblems zu ϑ sind, d.h. falls neben der Einfachheit $F(\vartheta, (v_i, \lambda_i)_{i=1}^n) = \mathbf{0}_{n(n+1)}$ gilt.
Seien $z \in \mathbb{R}^n, \alpha \in \mathbb{R}$, dann gilt:

$$\begin{aligned}
\left(\frac{\partial F}{\partial (v_i, \lambda_i)} \left(\vartheta, (v_i, \lambda_i)_{i=1}^n\right)\right) \begin{bmatrix} z \\ \alpha \end{bmatrix} = 0 \Longrightarrow v_i^{\mathrm{T}} M(\vartheta) z = 0 \quad \text{und} \\
M^{1/2}(\vartheta) v_i \alpha = M^{-1/2}(\vartheta) \left[K(\vartheta) - \lambda_i M(\vartheta)\right] z \\
\in \mathrm{Bild}\left(M^{-1/2}(\vartheta) \left[K(\vartheta) - \lambda_i M(\vartheta)\right]\right) \cap \mathrm{Ker}\left(\left[K(\vartheta) - \lambda_i M(\vartheta)\right] M^{-1/2}\right) \\
= \mathrm{Bild} \cap \mathrm{Ker}\left(M^{-1/2} \left[K(\vartheta) - \lambda_i M(\vartheta)\right] M^{-1/2}\right) = \{\mathbf{0}_n\},
\end{aligned}$$

da $M^{-1/2} \left[K(\vartheta) - \lambda_i M(\vartheta)\right] M^{-1/2}$ symmetrisch ist. Also folgt $\alpha = 0$, d.h. z ist wie v_i Eigenvektor zum einfachen Eigenwert λ_i. Da aber $v_i^{\mathrm{T}} M(\vartheta) z = 0$ gilt, folgt $z = 0$.

Aufgrund der Invertierbarkeit sind also die Voraussetzungen des impliziten Funktionentheorems erfüllt, so dass man in einer Umgebung von $(v_i, \lambda_i)_{i=1}^n$ die Eigenwerte und die Eigenvektoren eindeutig als stetig differenzierbare Funktionen von ϑ erhält, wobei

$$\begin{pmatrix} \frac{\partial v_i(\vartheta)}{\partial \vartheta_s} \\ \frac{\partial \lambda_i(\vartheta)}{\partial \vartheta_s} \end{pmatrix} = -\begin{bmatrix} K(\vartheta) - \lambda_i(\vartheta) M(\vartheta) & -M(\vartheta) v_i(\vartheta) \\ 2v_i^{\mathrm{T}}(\vartheta) M(\vartheta) & 0 \end{bmatrix}^{-1} \cdot \begin{pmatrix} \frac{\partial K(\vartheta)}{\partial \vartheta_s} v_i(\vartheta) - \lambda_i(\vartheta) \frac{\partial M(\vartheta)}{\partial \vartheta_s} v_i(\vartheta) \\ v_i^{\mathrm{T}}(\vartheta) \frac{\partial M(\vartheta)}{\partial \vartheta_s} v_i(\vartheta) \end{pmatrix}$$

gilt. □

Bemerkung 6.3

a) Für das gewöhnliche Eigenwertproblem

$$K(\vartheta)v = \lambda v,$$

d.h. im Fall $M = I_n$, geben Lancaster (1964) und Andrew u. Tan (1999) einige Resultate zur Differenzierbarkeit für mehrfache Eigenwerte an. Eine Übertragung auf das verallgemeinerte Eigenwertproblem oder eine praktische Umsetzung der theoretischen Resultate erscheint uns z.B. deswegen problematisch, da die Matrizen $K(\vartheta)$ dabei nur von einem Parameter $\vartheta \in \mathbb{R}$ analytisch abhängen dürfen.

b) Für Formeln zu zweiten Ableitungen der Eigenwerte nach ϑ im Fall $M = I_n$ vergleiche zum Beispiel Magnus u. Neudecker (1995, Section 8.11).

Zur praktischen Berechnung der Ableitungen der Eigenvektoren verwendet man meist nicht die im Satz hergeleitete Formel, sondern andere Methoden, die wir im Folgenden erläutern.

Differenzieren der Eigenwertgleichung (6.9) nach der s-ten Komponente des strukturellen Parameters $\vartheta = (\vartheta_1, \ldots, \vartheta_p)$ ergibt:

$$[K(\vartheta) - \lambda_i(\vartheta)\, M(\vartheta)] \frac{\partial v_i(\vartheta)}{\partial \vartheta_s} = f_i \tag{6.11}$$

mit

$$f_i = \left[-\frac{\partial K(\vartheta)}{\partial \vartheta_s} + \lambda_i(\vartheta) \frac{\partial M(\vartheta)}{\partial \vartheta_s} + \frac{\partial \lambda_i(\vartheta)}{\partial \vartheta_s} M(\vartheta) \right] v_i(\vartheta).$$

Aus (6.11) erhalten wir mittels Multiplikation mit $v_i^{\mathrm{T}}(\vartheta)$ von links:

$$\frac{\partial \lambda_i(\vartheta)}{\partial \vartheta_s} = v_i^{\mathrm{T}}(\vartheta) \left[\frac{\partial K(\vartheta)}{\partial \vartheta_s} - \lambda_i(\vartheta) \frac{\partial M(\vartheta)}{\partial \vartheta_s} \right] v_i(\vartheta)$$

Die Gleichung (6.11) bedeutet, dass $x = \frac{\partial v_i(\vartheta)}{\partial \vartheta_s}$ Lösung von

$$[K(\vartheta) - \lambda_i(\vartheta)\, M(\vartheta)]\, x = f_i.$$

Ist $\lambda_i(\vartheta)$ ein einfacher Eigenwert, so hat die Koeffizientenmatrix dieses inhomogenen linearen Gleichungssystems Rangdefekt 1. Ist w_i eine spezielle Lösung, so ist die allgemeine Lösung von der Form $w_i + cv_i(\vartheta)$ für $c \in \mathbb{R}$. Die allgemeine Methode besteht nun darin, sich eine verallgemeinerte Inverse[13] G von $[K(\vartheta) - \lambda_i(\vartheta)\, M(\vartheta)]$ zu verschaffen. Eine spezielle Lösung ist dann $w_i = Gf_i$, vergleiche Rao u. Mitra (1971, S. 20 f.).

[13] Zu einer Matrix $A \in \mathbb{R}^{n \times l}$ ist die Menge A^- aller verallgemeinerten Inversen definiert durch:

$$A^- := \left\{ G \in \mathbb{R}^{l \times n} | AGA = A \right\}.$$

Schließlich ergibt sich c im Ansatz

$$\frac{\partial v_i(\vartheta)}{\partial \vartheta_s} = Gf_i + cv_i(\vartheta) \tag{6.12}$$

aus der Normierungsforderung (6.10): Differenziert man diese, so hat man

$$2v_i^{\mathrm{T}}(\vartheta)M(\vartheta)\frac{\partial v_i(\vartheta)}{\partial \vartheta_s} + v_i^{\mathrm{T}}(\vartheta)\frac{\partial M(\vartheta)}{\partial \vartheta_s}v_i(\vartheta) = 0\,.$$

Setzt man jetzt den Ansatz (6.12) für $\frac{\partial v_i(\vartheta)}{\partial \vartheta_s}$ ein, so erhält man:

$$c = -v_i^{\mathrm{T}}(\vartheta)M(\vartheta)Gf_i - \frac{1}{2}v_i^{\mathrm{T}}(\vartheta)\frac{\partial M(\vartheta)}{\partial \vartheta_s}v_i(\vartheta)\,. \tag{6.13}$$

Die beiden Methoden für die Bestimmung der Ableitungen der Eigenvektoren, die wir nun vorstellen werden, unterscheiden sich dabei in der Wahl der verallgemeinerten Inversen G. Wir stellen zuerst die Methode von Friswell (1989) vor.

Satz 6.4
Es gilt:

$$G = \sum_{k\neq i} \frac{1}{\lambda_k(\vartheta) - \lambda_i(\vartheta)}\, v_k(\vartheta)v_k^{\mathrm{T}}(\vartheta)$$

ist eine verallgemeinerte Inverse von $[K(\vartheta) - \lambda_i(\vartheta)\, M(\vartheta)]$.

Beweis: Setzt man $z_i = M^{1/2}(\vartheta)v_i(\vartheta)$ für $i = 1, \ldots, n$, so gilt

$$M^{-1/2}(\vartheta)K(\vartheta)M^{-1/2}(\vartheta)z_i = \lambda_i(\vartheta)z_i$$

und $I_n = \sum_{k=1}^n z_k z_k^{\mathrm{T}}$. Daher lautet die Spektralzerlegung von $M^{-1/2}(\vartheta)K(\vartheta)M^{-1/2}(\vartheta)$:

$$M^{-1/2}(\vartheta)K(\vartheta)M^{-1/2}(\vartheta) = \sum_k^n \lambda_k(\vartheta)z_k z_k^{\mathrm{T}}\,.$$

Folglich ist

$$\begin{aligned}
[K(\vartheta) - \lambda_i(\vartheta)M(\vartheta)]^- &= \left[M^{1/2}(\vartheta)\left(M^{-1/2}(\vartheta)K(\vartheta)M^{-1/2}(\vartheta) - \lambda_i(\vartheta)I_n\right)M^{1/2}(\vartheta)\right]^- \\
&= M^{-1/2}(\vartheta)\left[\left(M^{-1/2}(\vartheta)K(\vartheta)M^{-1/2}(\vartheta) - \lambda_i(\vartheta)I_n\right)\right]^- M^{-1/2}(\vartheta) \\
&= M^{-1/2}(\vartheta)\left[\sum_{k\neq i}\left(\lambda_k(\vartheta) - \lambda_i(\vartheta)\right)z_k z_k^{\mathrm{T}}\right]^- M^{-1/2}(\vartheta).
\end{aligned}$$

Aufgrund der Orthogonalität der z_k, $k = 1, \ldots, n$, ist dann

$$M^{-1/2}(\vartheta)\sum_{k\neq i}\frac{1}{\lambda_k(\vartheta) - \lambda_i(\vartheta)}z_k z_k^{\mathrm{T}}M^{-1/2}(\vartheta) = \sum_{k\neq i}\frac{1}{\lambda_k(\vartheta) - \lambda_i(\vartheta)}v_k(\vartheta)v_k^{\mathrm{T}}(\vartheta) = G$$

eine verallgemeinerte Inverse. □

Wendet man nun diesen Satz in der oben beschriebenen allgemeinen Methode an, so erhält man die folgende Formel für die Ableitung der Eigenvektoren:

$$\begin{aligned}\frac{\partial v_i(\vartheta)}{\partial \vartheta_s} &= Gf_i + cv_i(\vartheta) \\ &= -\frac{1}{2}\left(v_i^{\mathrm{T}}(\vartheta)\frac{\partial M(\vartheta)}{\partial \vartheta_s}v_i(\vartheta)\right)v_i(\vartheta) \\ &\quad + \sum_{k\neq i}\frac{1}{\lambda_k(\vartheta)-\lambda_i(\vartheta)}\,v_k(\vartheta)v_k^{\mathrm{T}}(\vartheta)\left[\lambda_i(\vartheta)\frac{\partial M(\vartheta)}{\partial \vartheta_s} - \frac{\partial K(\vartheta)}{\partial \vartheta_s}\right]v_i(\vartheta).\end{aligned}$$

Als Zweites stellen wir die Methode von Nelson (1976) vor und erläutern, welche verallgemeinerte Inverse G dabei verwendet wird.

Betrachte die Matrix A_l, die durch Streichen der l-ten Zeile und der l-ten Spalte von $[K(\vartheta) - \lambda_i(\vartheta)\, M(\vartheta)]$ entsteht.

Satz 6.5
Falls $v_i(\vartheta)$ mit $(v_i(\vartheta))_l \neq 0$ Eigenvektor zum einfachen Eigenwert $\lambda_i(\vartheta)$, so gilt:
A_l ist invertierbar.

Beweis: Bezeichnet L_i die i-te Spalte von $L = [K(\vartheta) - \lambda_i(\vartheta)\, M(\vartheta)]$, so folgt wegen $Lv_i(\vartheta) = 0$, dass

$$L_l = -\frac{1}{(v_i(\vartheta))_l}\sum_{j\neq l} L_j \cdot (v_i(\vartheta))_j$$

gilt, d.h. die l-te Zeile ist Linearkombination der anderen Zeilen, ebenso ist wegen der Symmetrie von L die l-te Spalte Linearkombination der anderen Spalten. Da aber L nach Voraussetzung Rangdefekt 1 hat, hat A_l vollen Rang, ist also invertierbar. □

Damit ist die Matrix $G \in \mathbb{R}^{n\times n}$, die durch Einfügen eines Nullvektors als l-te Zeile und als l-te Spalte von $A_l^{-1} \in \mathbb{R}^{(n-1)\times n-1}$ entsteht, eine verallgemeinerte Inverse von $[K(\vartheta) - \lambda_i(\vartheta)\, M(\vartheta)]$, vergleiche Rao u. Mitra (1971, S. 208). Mit dieser verallgemeinerten Inversen G erhält man mittels (6.12) und (6.13) die Ableitung $\frac{\partial v_i(\vartheta)}{\partial \vartheta_s}$.

Der Vorteil dieser zweiten Methode liegt darin, dass zur Berechnung der Ableitungen nicht alle Eigenwerte und Eigenvektoren benötigt werden.

6.2.5. Räumliche Modelle für a-priori-Verteilungen

Wir skizzieren in diesem Abschnitt, wie räumliche Modelle für die Wahl von a-priori-Verteilungen zum Einsatz kommen können.

In unserem Federbeispiel mit $K(\vartheta)$ und $M(\vartheta)$ gemäß (6.4) und (6.5) und $\vartheta = (k_1, \ldots, k_{n+1})$ ist ein einfacher Ansatz für eine a-priori-Verteilung des Strukturparameters

$$P^{\boldsymbol{\theta}} \sim \mathcal{N}(\vartheta_0 \mathbf{1}_{n+1}, \tau^2 I_{n+1}).$$

Dieser Ansatz beschreibt eine zufällige und unabhängige Abweichung der Komponenten des Strukturparameters um einen gemeinsamen Mittelwert $\vartheta_0 \in \mathbb{R}$. Hat man dagegen die Vorstellung, dass die Steifigkeiten räumlich korreliert sind, so wäre

$$P^{\boldsymbol{\theta}} \sim \mathcal{N}(\vartheta_0 \mathbf{1}_{n+1}, \tau^2 B(\rho)) \tag{6.14}$$

mit unbekanntem $\rho \in (-1, 1)$ und $\tau > 0$ und

$$B(\rho) = (b_{ij}(\rho))_{1 \le i,j \le n+1} \quad \text{mit } b_{ij} = \rho^{|i-j|}$$

ein einfacher Ansatz, der beschreibt, dass die räumliche Korrelation zwischen den Komponenten mit zunehmender Entfernung exponentiell abklingt.

Wir erläutern exemplarisch, wie man diese Überlegungen in der Praxis auf ein komplizierteres Modell überträgt und verweisen für die nötigen Begriffe und weitere Einblicke zur räumlichen Statistik hierbei auf Cressie (1993).

Zehn u. a. (2000, S. 697 ff.) vom Institut für Mechanik der Otto-von-Guericke-Universität haben ein strukturmechanisches Modell einer Platte beschrieben, das als Parameterkomponenten die Dicken einer Platte an verschiedenen Stellen beinhaltet. Um einen Eindruck zu gewinnen, welche Modelle für die räumliche Korrelation der Dickenparameter in Frage kommen, wurden hierzu mehrere Platten ausgemessen. Die Dicke einer dieser Platten kann man aus Abbildung 6.5 ersehen, die Messungen fanden dabei in den Gitterpunkten statt.

Diese Datenwerte zu dieser Platte wurden von uns untersucht und wir erstellten folgendes Modell: Für gegebene Gitterpunkte $u_i \in \mathbb{R}^2$, $i = 1, \ldots, n$, beschreibt die Zufallsvariable $\boldsymbol{\theta}_i$ die Dicke der Platte an der Stelle u_i. Um die gemeinsame Verteilung der Dicken zu beschreiben, wurde dabei ein isotropes Modell für die Korrelation gewählt, d.h. es wurde angenommen, dass die Kovarianz

$$\operatorname{Cov}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = C(\|u_i - u_j\|) \tag{6.15}$$

für eine geeignete Funktion $C : [0, \infty) \to \mathbb{R}$ nur vom Abstand der zugehörigen Gitterpunkte abhängt.

Wir haben nun verschiedene parametrische Modelle untersucht, um das Semivariogramm

$$\gamma(h) = C(0) - C(h) : [0, \infty) \to \mathbb{R} \tag{6.16}$$

anzupassen.[14] Von den üblichen parametrischen Ansätzen für γ, siehe Cressie (1993, S. 61 f.) erwies sich dabei das Exponentialmodell[15] am geeignetsten, d.h. der Ansatz

$$\gamma_{a,c}(h) = c(1 - \exp(-h/a)). \tag{6.17}$$

[14] Da $2\gamma(\|v_i - v_j\|) = \operatorname{Var}(\vartheta_i - \vartheta_j)$ liegt es nahe, das Semivariogramm aus den quadrierten Dickendifferenzen zu schätzen. Zur Schätzung wurde dabei als Zielfunktion ein Vorschlag aus Cressie (1993, (2.6.12)) verwendet. Dieser Ansatz liefert typischerweise besonders für kleine Distanzen eine sehr gute Näherung.

[15] Das Modell (6.14) ist der Spezialfall eines Exponentialmodells im $\mathbb{R}^1$.

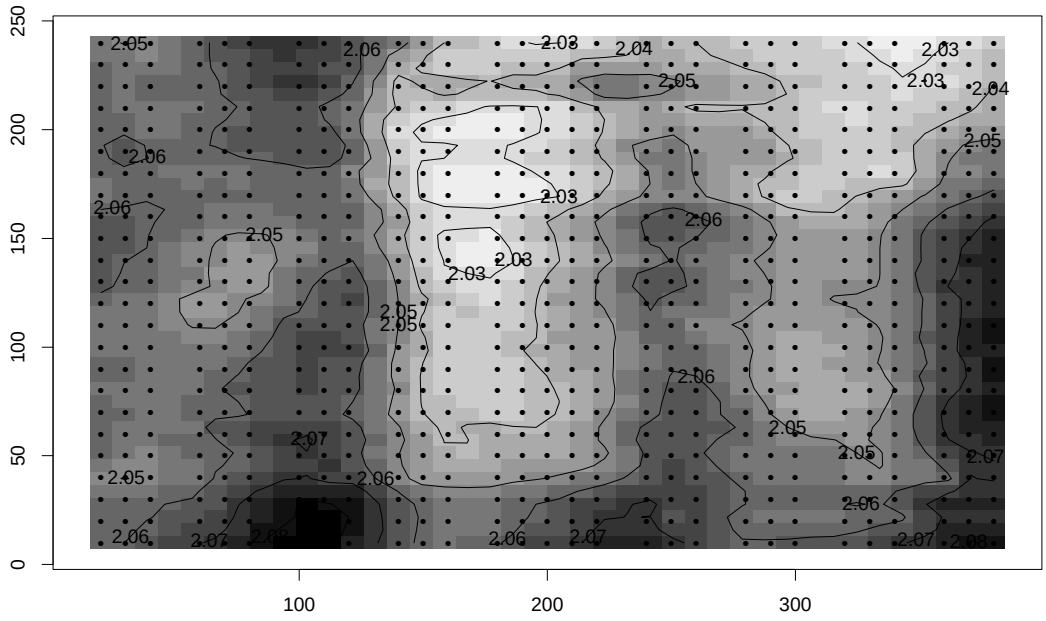


Abbildung 6.5.: Eine Platte mit nichtkonstanter Dicke
(an den dunkleren Stellen ist die Plattendicke größer)

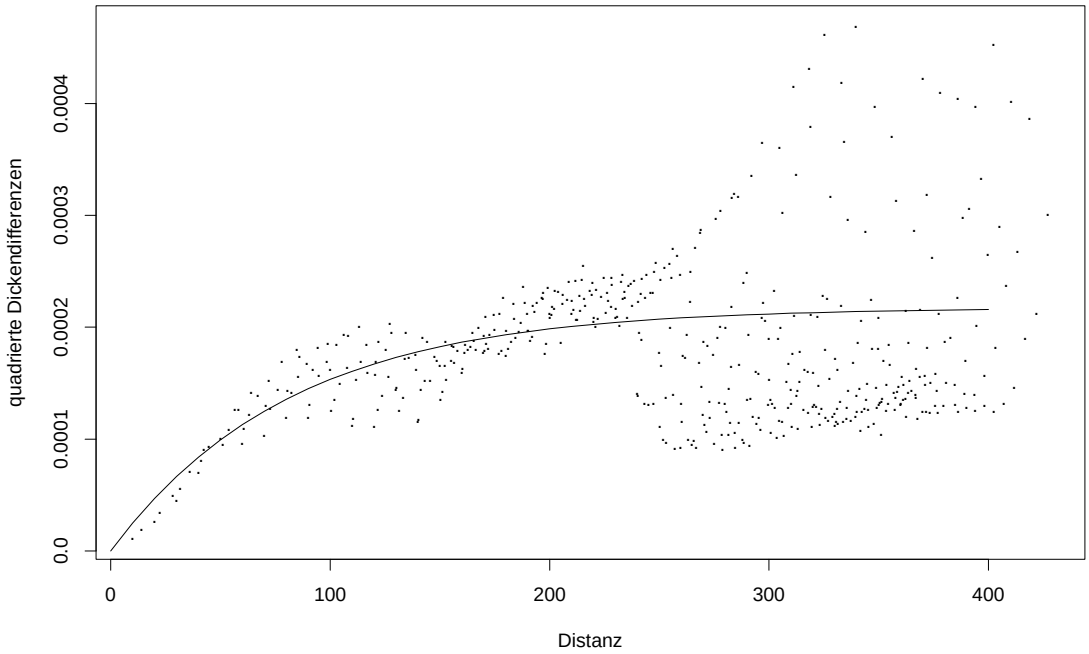


Abbildung 6.6.: Variogrammschätzung für die Platte aus Abbildung 6.5

Der Mittelwert der quadrierten Dickendifferenzen in Abhängigkeit vom Abstand der Gitterpunkte ist dabei in Abbildung 6.6 zusammen mit der Kurve für das Variogramm 2γ dargestellt.

Für ein praxisrelevantes Modell wird man im Allgemeinen die Dicke einer Platte an den verschiedenen Stellen nicht durch Versuche ermitteln, sondern bei der Anpassung eines Finite-Elemente-Modells versuchen, durch Messungen in einem repräsentativen Ausschnitt ein geeignetes Modell für die Kovarianzfunktion zu bestimmen. Man wählt dann also z.B. das Exponentialmodell (6.17). Falls man zudem keine systematischen Abweichungen von einer bekannten mittleren Dicke $\vartheta_0 \in (0, \infty)$ erwartet, wäre dann

$$P^{\boldsymbol{\theta}} = \mathcal{N}(\mathbf{1}_p \vartheta_0, B(C(0), a, c)) \tag{6.18}$$

ein geeignetes Modell für eine a-priori-Verteilung. Hierbei ergibt sich die Kovarianzmatrix $B(C(0), a, c)$ gemäß (6.15), (6.16) und (6.17). Die Parameter a, c und $C(0)$ werden für das weitere Vorgehen als unbekannt angesetzt und dann aus einer Messung y in dem Modell

$$P^{\boldsymbol{Y}|\boldsymbol{\theta}=\vartheta} = \mathcal{N}(\eta(\vartheta), \Sigma)$$

zusammen mit der a-priori-Annahme (6.18) eine empirische Bayes-Schätzung für ϑ bestimmt.

Dieses von uns in Offinger (2000) initiierte Vorgehen, Ansätze aus der räumlichen Statistik zur Spezifikation einer a-priori-Verteilung für den Strukturparameter zu verwenden, wurde von Zehn u. Saitov (2003) und von Zehn u. Machina (2005) fortgeführt.

Wir berichten abschließend noch von numerischen Experimenten zum Algorithmus 5.7, der auf einer iterativen Linearisierung der Modellfunktion η basierte, zur Berechnung einer empirischen Bayes-Schätzung.

Wir haben dazu in unserem Federbeispiel mit $K(\vartheta)$ und $M(\vartheta)$ gemäß (6.4), (6.5) und der a-priori-Verteilung (6.14) für die Steifigkeiten bei unbekanntem ρ und τ die empirische Bayes-Schätzung für ϑ untersucht. Dabei zeigte sich, dass beim Minimierungsproblem in Schritt 2 von Algorithmus 5.7 die verwendeten Minimierungsroutinen oft verschiedene lokale Minima fanden und die Schätzung, die das Gesamtverfahren lieferte, von der verwendeten Unterroutine abhängig war. Dieses numerische Problem ist momentan nicht hinreichend gelöst.

Zehn u. Saitov (2003), die unseren Algorithmus 5.7 in einem Modell aus der Praxis verwendeten, berichten ebenfalls von numerischen Schwierigkeiten. Praktische Erfahrungen mit den Approximationen aus Abschnitt 5.2.2 stehen noch aus und sind in Planung. Weiterhin wollen wir in unserem Federbeispiel auch den Einsatz anderer numerischer Methoden für das Integrationsproblem zur Bestimmung der Randverteilung im empirischen Bayes-Modell untersuchen und Vergleiche mit unseren Approximationen anstellen.

A. Beweise

A.1. Beispiel 2.5

a) Wir werden an dieser Stelle lediglich zeigen, dass es im Modell $\boldsymbol{Y} = \eta(\vartheta) + \epsilon$ mit $\Theta = [0, \pi]$, $\eta : \Theta \to \mathbb{R}^2$, $\eta(\vartheta) = (\cos(\vartheta), \sin(\vartheta))^{\mathrm{T}}$ und $\epsilon \sim \mathcal{N}(0, I_2)$ keinen erwartungstreuen Schätzer für ϑ gibt, der lediglich auf der ersten Komponente beruht, d.h. es gibt keine Borel-messbare Funktion $g : \mathbb{R} \to \mathbb{R}$, so dass $\mathrm{E}_\vartheta(g(\boldsymbol{Y}_1)) = \vartheta$ für alle $\vartheta \in \Theta$.
Man kann dieses Problem auch so formulieren, dass man in dem Modell $\boldsymbol{Z} \sim \mathcal{N}(\mu, 1)$ mit $\mu \in [-1, 1]$ einen erwartungstreuen Schätzer $\phi(\boldsymbol{Z})$ für $h(\mu) = \arccos(\mu)$ sucht, wobei $\phi : \mathbb{R} \to \mathbb{R}$ integrierbar bezüglich des Wahrscheinlichkeitsmaßes $\mathcal{N}(\mu, 1)$ ist. Wir definieren zu einer solchen Funktion ϕ die Funktion

$$\beta : [-1, 1] \to \mathbb{R},\ \beta(\mu) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \phi(z) \exp\left(-\frac{1}{2}(z - \mu)^2\right) \mathrm{d}z$$

und betrachten die entsprechende Gleichung für die Erwartungstreue

$$\arccos(\mu) = \beta(\mu) \quad \text{für alle } \mu \in [-1, 1]. \tag{A.1}$$

Analog zu Witting (1985, Hilfssatz 1.162 und Korollar 1.163) kann man zeigen, dass die Funktion β analytisch ist, insbesondere existiert die Ableitung in den Randpunkten $\mu = -1$ und $\mu = 1$. Da jedoch die Funktion $\arccos : [-1, 1] \to \mathbb{R}$ nicht in $\mu = -1$ und $\mu = 1$ differenzierbar ist, kann die Gleichung (A.1) nicht für alle $\mu \in [-1, 1]$ gelten, und es gibt keinen erwartungstreuen Schätzer.

b) Wenn es im Modell $\boldsymbol{Y} \sim \mathcal{N}(1/\vartheta \cdot \mathbf{1}_N, I_N)$ mit $\vartheta \in \mathbb{R} \setminus \{0\}$ einen erwartungstreuen Schätzer für ϑ gäbe, so hinge er von der vollständigen und suffizienten Statistik $\boldsymbol{Z} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{Y}_i$ ab, d.h. wir können stattdessen die Frage stellen, ob es im Modell $\boldsymbol{Z} \sim \mathcal{N}(\mu, \sigma_0^2)$ mit $\mu \in \mathbb{R} \setminus \{0\}$ und bekanntem $\sigma_0 > 0$ (hier: $\sigma_0 = 1/N$) einen erwartungstreuen Schätzer für $h(\mu) = 1/\mu$ gibt. Gesucht ist also eine Funktion $g : \mathbb{R} \to \mathbb{R}$, die $\mathcal{N}(\mu, \sigma_0^2)$-integrierbar für $\mu \neq 0$ ist und

$$\frac{1}{\mu} = \frac{1}{\sigma_0 \sqrt{2\pi}} \int_{\mathbb{R}} g(z) \exp\left(-\frac{(z - \mu)^2}{2\sigma_0^2}\right) \mathrm{d}z \tag{A.2}$$

für $\mu \neq 0$ erfüllt.
Da

$$(z - \mu)^2 \geq (z - 1)^2 - 1 \text{ für } z > 0 \text{ und } \mu \in [-1, 1]$$

und

$$(z-\mu)^2 \geq (z+1)^2 - 1 \text{ für } z < 0 \text{ und } \mu \in [-1,1],$$

wäre für eine solche Funktion

$$\begin{aligned} g^+(z)\exp\left(-\frac{(z-\mu)^2}{2\sigma_0^2}\right) \leq & \exp\left(\frac{1}{2\sigma_0^2}\right) g^+(z)\exp\left(-\frac{(z+1)^2}{2\sigma_0^2}\right)\mathbf{1}_{(-\infty,0)} \\ & + \exp\left(\frac{1}{2\sigma_0^2}\right) g^+(z)\exp\left(-\frac{(z-1)^2}{2\sigma_0^2}\right)\mathbf{1}_{(0,\infty)} \end{aligned}$$

für $\mu \in [-1,1]$. Mit der analogen Aussage für g^- erhält man eine integrierbare Majorante und damit existiert nach dem Satz von der majorisierten Konvergenz der Grenzwert

$$\lim_{\mu\to 0}\int_{\mathbb{R}} g(z)\exp\left(-\frac{(z-\mu)^2}{2\sigma_0^2}\right)\mathrm{d}z = \int_{\mathbb{R}} g(z)\exp\left(-\frac{z^2}{2\sigma_0^2}\right)\mathrm{d}z$$

im Widerspruch zur Gleichung (A.2).

Betrachtet man hingegen das entsprechende Modell mit dem eingeschränkten Parameterbereich $\Theta = (0,\infty)$ anstelle von $\mathbb{R}\setminus\{0\}$, so gibt es einen erwartungstreuen Schätzer, vgl. Voinov (1985).

A.2. Beispiel 2.32

Wir untersuchen, wie viele Nullstellen das Polynom

$$p_y(t) = t^4 - y_1 t^3 + y_2 t - 1$$

in Abhängigkeit von y_1 und y_2 auf $\mathbb{R}^+$ hat. Es gilt

$$\begin{aligned} p_y'(t) &= 4t^3 - 3y_1t^2 + y_2 \\ p_y''(t) &= 12t^2 - 6y_1t. \end{aligned}$$

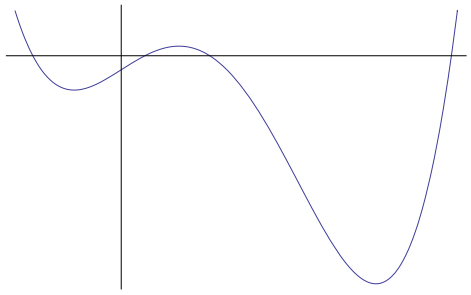
1.Fall: $y_1 \leq 0$
Dann ist $p''(t) > 0$ für $t > 0$, d.h. p' ist streng monoton steigend auf $\mathbb{R}^+$. Ist $y_2 \geq 0$, dann ist $p_y'(t) > 0$ für $t > 0$ d.h. p_y ist streng monoton steigend und wegen $p_y(0) = -1$ gibt es folglich eine eindeutige Nullstelle von p_y auf $\mathbb{R}^+$. Ist $y_2 < 0$, dann gibt es eine eindeutige Nullstelle t_0 von p_y' und $p_y'(t) < 0$ für $t < t_0$ und $p_y'(t) > 0$ für $t > t_0$, d.h. p hat ein eindeutiges Minimum auf $\mathbb{R}^+$ in t_0 und wegen $p_y(0) = -1$ gibt es ebenfalls eine eindeutige Nullstelle von p_y auf $\mathbb{R}^+$.

2.Fall: $y_1 > 0$
Dann hat p_y' auf $\mathbb{R}^+$ ein eindeutiges Minimum in $t = y_1/2$ mit Wert $p_y'(y_1/2) = -t_1^3/4 + y_2$. Wenn $y_2 \leq 0$, dann ist wegen $p_y'(0) = y_2$ das Polynom p_y erst streng monoton fallend und

dann streng monoton steigend auf $\mathbb{R}^+$, d.h. es hat dort ein eindeutiges Minimum und wegen $p_y(0) = -1$ gibt es wieder eine eindeutige Nullstelle von p_y auf $\mathbb{R}^+$. Im Fall $y_2 > 0$ kann man erst noch weiter unterteilen: Im Fall von $y_2 \geq y_1^3/4$ ist $p_y'(t) \geq 0$ auf $\mathbb{R}^+$, da für den Wert im Minimum $p_y'(y_1/2) > 0$ gilt. Bleibt also noch der Fall $y_2 > y_1^3/4$. Dann hat p_y' zwei Nullstellen auf $\mathbb{R}^+$, eine Nullstelle auf $\mathbb{R}^-$, folglich ein eindeutiges Minimum auf $\mathbb{R}^-$ und wegen $p_y(0) = -1$ daher genau eine Nullstelle auf $\mathbb{R}^-$. Es bleibt also die Frage, wann in diesem Fall das Polynom p_y drei Nullstellen auf $\mathbb{R}^+$ hat bzw. äquivalent wann das Polynom p_y vier Nullstellen auf $\mathbb{R}^+$ hat. Einsetzen in die Bedingungen von Abramowitz u. Stegun (1972, S. 17 f.) liefert dann insgesamt

$$U = \{y \in \mathbb{R}^2 : y_1 > 0,\, y_2 > 0,\, 4(-y_1 y_2 + 4)^3 + 27(y_1^2 - y_2^2)^2 < 0\}$$

als die Menge der y-Werte, für die das Polynom p_y drei Nullstellen auf $\mathbb{R}^+$ hat und dann von folgender Gestalt ist:



A.3. Beispiel 2.38

Sei $X_1 \in \mathbb{R}^{N\times u}$ und $X_2 \in \mathbb{R}^{N\times v}$. Wir zeigen

$$\text{Bild}\begin{pmatrix} I_u \\ \mathbf{0}_{v\times u} \end{pmatrix} \subseteq \text{Bild}\begin{pmatrix} X_1^{\mathrm{T}} \\ X_2^{\mathrm{T}} \end{pmatrix} \iff \text{rg}((I_N - \mathcal{P}_{X_2})X_1) = u.$$

Beweis:
$\Rightarrow$: Gilt

$$\text{Bild}\begin{pmatrix} I_u \\ \mathbf{0}_{v\times u} \end{pmatrix} \subseteq \text{Bild}\begin{pmatrix} X_1^{\mathrm{T}} \\ X_2^{\mathrm{T}} \end{pmatrix},$$

so existiert $A \in \mathbb{R}^{N\times u}$ mit $I_u = X_1^{\mathrm{T}} A$ und $\mathbf{0}_{v\times u} = X_2^{\mathrm{T}} A$. Damit folgt

$$A^{\mathrm{T}}((I_N - \mathcal{P}_{X_2})X_1) = A^{\mathrm{T}}X_1 - A^{\mathrm{T}}X_2(X_2^{\mathrm{T}}X_2)^+X_2^{\mathrm{T}}X_1 = I_u,$$

d.h. $\text{rg}((I_N - \mathcal{P}_{X_2})X_1) = u$.
$\Leftarrow$: Gilt hingegen $\text{rg}((I_N - \mathcal{P}_{X_2})X_1) = u$, dann gibt es wegen $\text{rg}(X_1^{\mathrm{T}}(I_N - \mathcal{P}_{X_2})) = u$ ein $B \in \mathbb{R}^{N\times u}$ mit

$$X_1^{\mathrm{T}}(I_N - \mathcal{P}_{X_2})B = I_u.$$

Da

$$X_2^{\mathrm{T}}(I_N - \mathcal{P}_{X_2})B = (X_2^{\mathrm{T}} - X_2^{\mathrm{T}})B = \mathbf{0}_{v\times u}$$

gilt, folgt mit $A = (I_N - \mathcal{P}_{X_2})B$, dass $I_u = X_1^{\mathrm{T}} A$ und $\mathbf{0}_{v\times u} = X_2^{\mathrm{T}} A$ und daher

$$\text{Bild}\begin{pmatrix} I_u \\ \mathbf{0}_{v\times u}\end{pmatrix} \subseteq \text{Bild}\begin{pmatrix} X_1^{\mathrm{T}} \\ X_2^{\mathrm{T}}\end{pmatrix}.$$

□

A.4. Rechenoperationen bei der QR-Zerlegung

Zur Untersuchung der Komplexität des Verfahrens erinnern wir an die Grundzüge der QR-Zerlegung mittels Householder-Matrizen[1] und starten hierbei mit einigen Bezeichnungen.

Eine Matrix $H \in \mathbb{R}^{m\times m}$ heißt **Householder-Matrix**, falls sie von der Form

$$H = H_v = I_m - \frac{2}{v^{\mathrm{T}} v} v v^{\mathrm{T}}$$

mit $v \in \mathbb{R}^m$ ist. H ist also die Spiegelungsmatrix zu der Hyperebene mit Normalenvektor $v \in \mathbb{R}^m$, und insbesondere gilt $H = H^{\mathrm{T}}$ und $H^{\mathrm{T}} H = H^2 = I_m$. Sucht man ferner zu gegebenen Vektoren $x, y \in \mathbb{R}^m$, $x \neq y$ mit $\|x\| = \|y\|$, eine Householder-Matrix H_v, so dass $H_v x = y$, so wird diese Aufgabenstellung durch $v = x - y$ gelöst.
Mit dem Anwendungsziel bei der Linksmultiplikation einer Matrix mit einer Householder-Matrix die Einträge einer Spalte unter der Hauptdiagonalen zu 0 werden zu lassen, interessiert man sich insbesondere für den Fall[2] $x \neq 0$ und $y = -\mathrm{sgn}(x_1)\,\|x\|\,e_1$, wobei

$$\mathrm{sgn}(a) = \begin{cases} +1 & \text{für } a \geq 0, \\ -1 & \text{für } a < 0. \end{cases}$$

Mit

$$s(x) = \mathrm{sgn}(x_1)\,\|x\|\,, \tag{A.3}$$

$$v(x) = x + \mathrm{sgn}(x_1)\,\|x\|\,e_1 = x + s(x)e_1, \tag{A.4}$$

$$\beta(x) = \frac{2}{v^{\mathrm{T}}(x)v(x)} = \frac{1}{s(x)(x_1 + s(x))} \tag{A.5}$$

ergibt sich dann die zugehörige Matrix

$$H_{v(x)} = I_m - \beta(x) v(x) v^{\mathrm{T}}(x).$$

[1] Für eine QR-Zerlegung mittels Givens-Rotationen, vgl. Higham (2002, Section 19.6). Für vollbesetzte Matrizen fallen ca. 50 Prozent mehr Rechenoperationen als bei der QR-Zerlegung mittels Householder-Matrizen an. Das Haupteinsatzgebiet dieses Verfahrens liegt bei der Zerlegung von schwach besetzten Matrizen, z.B. von Tridiagonal-Matrizen.

[2] Die Wahl des Vorzeichens ist vor allem historisch bedingt; auch für den Fall $y = +\mathrm{sgn}(x_1)\,\|x\|\,e_1$ kann man durch geeignete Berechnung das Problem der numerischen Auslöschung vermeiden, vgl. Higham (2002, Section 19.1).

Ist $C = [c_1, \ldots, c_p] \in \mathbb{R}^{N\times p}$ mit $\operatorname{rg}(C) = p$, $B = [b_1, \ldots, b_p] \in \mathbb{R}^{N\times p}$, so ist $c_1 \neq \mathbf{0}_N$ und man erhält durch Linksmultiplikation mit $H_1 = H_{v(b_1)}$ eine Matrix der Form

$$H_1 B = \begin{pmatrix} * & * & \ldots & * \\ 0 & * & \ldots & * \\ \vdots & \vdots & & \vdots \\ 0 & * & \ldots & * \end{pmatrix}.$$

Nun zur induktiven Berechnung der Rechenoperationen bei der QR-Zerlegung einer Matrix $A \in \mathbb{R}^{N\times p}$ mittels Householder-Matrizen. Allgemein hat man vor dem k-ten Schritt eine Matrix B_k der Form

$$B_k = H_{k-1} \cdots H_1 B = \begin{pmatrix} R_k & z_k & C_k \\ \mathbf{0}_{(N-k)\times k} & x_k & D_k \end{pmatrix} \tag{A.6}$$

mit einer oberen Dreiecksmatrix $R_k \in \mathbb{R}^{(k-1)\times(k-1)}$, dem Vektor $z_k \in \mathbb{R}^{k-1}$, einer Matrix $C_k \in \mathbb{R}^{(k-1)\times(p-k)}$, dem Vektor $x_k \in \mathbb{R}^{N-k+1}$ und einer Matrix $D_k \in \mathbb{R}^{(N-k+1)\times(p-k)}$.
Ist $N = p$, so ist man schon nach Schritt $k = p-1$ mit der QR-Zerlegung fertig, ansonsten nach Schritt $k = p$.
Zu Beginn von Schritt k bietet sich eine Pivot-Strategie an, indem man unter dem Vektor x_k und den Spalten von D_k den betragsgrößten Vektor sucht und durch Permutation an die Stelle von x_k bringt, d.h. mit einer Permutationsmatrix P_{k+1} von rechts multipliziert. Man hat dann wieder eine Matrix der Form (A.6).[3] Wir werden im Folgenden die Bezeichnungen nach einem Vorgehen ohne Pivotstrategien ausrichten.
Mit $v(x_k) = x_k + \operatorname{sgn}((x_k)_1)\,\|x_k\|\,e_1$ und

$$\begin{aligned} H_{v(x_k)} &= I_{N-k+1} - \frac{2}{v^{\mathrm{T}}(x_k)v(x_k)} v(x_k)v^{\mathrm{T}}(x_k) \\ &= I_{N-k+1} - \beta(x_k)\, v(x_k)v^{\mathrm{T}}(x_k) \in \mathbb{R}^{(N-k+1)\times(N-k+1)} \end{aligned}$$

setzt man

$$H_k = \begin{pmatrix} I_{k-1} & \mathbf{0}_{(k-1)\times(N-k+1)} \\ \mathbf{0}_{(N-k+1)\times(k-1)} & H_{v(x_k)} \end{pmatrix}$$

und multipliziert B_k von links mit H_k, wobei

$$H_k B_k = H_k \begin{pmatrix} R_k & z_k & C_k \\ \mathbf{0}_{(N-k)\times k} & x_k & D_k \end{pmatrix} = \begin{pmatrix} R_k & z_k & C_k \\ \mathbf{0}_{(N-k)\times k} & -s(x_k)e_1 & H_{v(x_k)}D_k \end{pmatrix}$$

und

$$H_{v(x_k)} D_k = D_k - v(x_k)\left(\beta(x_k)(v^{\mathrm{T}}(x_k)D_k)\right). \tag{A.7}$$

[3] Zusätzlich kann man zu Beginn des Algorithmus noch eine Zeilen-Pivotstrategie anwenden, indem man die Zeilen nach der Maximumsnorm anordnet, d.h. zu Beginn eine Multiplikation von links mit einer Permutationsmatrix P_0 durchführt. Am Ende ist dann $Q = P_0 H_1 \cdots H_{\min\{N,p\}}$. Zu Pivotstrategien bei der QR-Faktorisierung siehe auch Higham (2002, Section 19.4).

Ohne Pivotisierungsstrategie ergibt sich schließlich für $N > p$

$$H_p \cdots H_1 B = \begin{pmatrix} R \\ \mathbf{0}_{(N-p)\times p} \end{pmatrix}$$

mit einer oberen Dreiecksmatrix $R \in \mathbb{R}^{p\times p}$ und im Fall $N = p$

$$H_{p-1} \cdots H_1 B = R$$

mit einer oberen Dreiecksmatrix $R \in \mathbb{R}^{p\times p}$. Mit $Q = (H_p \cdots H_1)^{\mathrm{T}} = H_1 \cdots H_p$ im Fall $N > p$ bzw. $Q = (H_{p-1} \cdots H_1)^{\mathrm{T}} = H_1 \cdots H_{p-1}$ im Fall $N = p$ gilt also

$$B = Q \begin{pmatrix} R \\ \mathbf{0}_{(N-p)\times p} \end{pmatrix}.$$

Wendet man noch die Spalten-Pivotstrategie an, so ergibt sich stattdessen

$$BP = Q \begin{pmatrix} R \\ \mathbf{0}_{(N-p)\times p} \end{pmatrix}$$

mit einer Permutationsmatrix $P = P_1 \cdots P_{p-1}$.

Nutzt man die Formeln (A.3), (A.4), (A.5) und (A.7), so ergeben sich im k-ten Schritt ohne Pivotstrategie folgende Rechenoperationen:
Die Bestimmung von $s(x_k)$ erfordert $N-k+1$ Quadrierungen (Multiplikationen), $N-k$ Additionen, eine Quadratwurzelziehung, eine Vorzeichenermittlung und eventuell eine Vorzeichenänderung. Zur Bestimmung von $\beta(x_k)$ wird dann eine Addition, eine Multiplikation und eine Division benötigt, wobei damit auch die erste Komponente von $v(x_k)$ berechnet ist. Um $v^{\mathrm{T}}(x_k)D_k$ zu ermitteln, werden $(N-k+1)(p-k)$ Multiplikationen und $(N-k)(p-k)$ Additionen benötigt, die anschließende Multiplikation mit dem Skalar $\beta(x_k)$ benötigt $p-k$ Multiplikationen. Hiernach fällt das äußere Produkt eines $N-k+1$-Vektors mit einem $p-k$-Vektor an, was $(N-k+1)(p-k)$ Multiplikationen erfordert. Die abschließende Differenz der beiden Matrizen erfodert $(N-k+1)(p-k)$ Subtraktionen. Insgesamt fallen also im k-ten Schritt $N-k+1+(N-k)(p-k)+(N-k+1)(p-k)$ Additionen/Subtraktionen, $(N-k+1)+1+(N-k+1)(p-k)+(p-k)+(N-k+1)(p-k)$ Multiplikationen, 1 Division, 1 Quadratwurzelziehung und 1 Vorzeichenermittlung und eventuell eine Vorzeichenänderung an. Im Fall $N = p$ fallen daher über alle Schritte neben den $p-1$ Divisionen, $p-1$ Quadratwurzelziehungen und den $p-1$ Vorzeichenermittlungen bzw. eventuellen Vorzeichenänderungen

$$\sum_{k=1}^{p-1} N-k+1+(N-k)(p-k)+(N-k+1)(p-k) = \frac{-3+p+2p^3}{3}$$

Additionen/Subtraktionen und

$$\sum_{k=1}^{p-1} (N-k+1)+1+(N-k+1)(p-k)+(p-k)+(N-k+1)(p-k) = \frac{-6+p+3p^2+2p^3}{3}$$

Multiplikationen an. Betrachtet man also Additionen, Subtraktionen, Multiplikationen, Divisionen und Quadratwurzelziehungen zusammen, so ergeben sich

$$\frac{-3+p+2p^3}{3}+\frac{-6+p+3p^2+2p^3}{3}+2(p-1)=\frac{-15+8p+3p^2+4p^3}{3}$$

Rechenoperationen.[4]
Im Fall $N > p$ fallen neben p Divisionen, p Quadratwurzelziehungen, p Vorzeichenermittlungen bzw. eventuellen Vorzeichenänderungen

$$\sum_{k=1}^{p} N-k+1+(N-k)(p-k)+(N-k+1)(p-k)=\frac{p-p^3+3Np^2}{3} \tag{A.8}$$

Additionen/Subtraktionen und

$$\sum_{k=1}^{p}(N-k+1)+1+(N-k+1)(p-k)+(p-k)+(N-k+1)(p-k)=\frac{p+3p^2-p^3+3Np^2}{3} \tag{A.9}$$

Multiplikationen an, also insgesamt

$$\frac{p-p^3+3Np^2}{3}+\frac{p+3p^2-p^3+3Np^2}{3}+2p=\frac{8p+3p^2-2p^3+6Np^2}{3}$$

Rechenoperationen.
Wendet man noch obige Pivotstrategie an, dann muss man im k-ten Schritt noch das Quadrat der Normen der Spalten von D_k bestimmen, d.h. von $p-k$ Vektoren im $\mathbb{R}^{N-k+1}$. Dies benötigt $(N-k+1)(p-k)$ Quadrierungen und $(N-k)(p-k)$ Additionen. Über alle Schritte fallen also im Fall $N \geq p$ zusätzlich

$$\sum_{k=1}^{p}(N-k)(p-k)=\frac{1}{6}(3N-p-1)(p-1)p=\frac{1}{2}Np^2-\frac{1}{6}p^3-\frac{1}{2}Np+\frac{1}{6}p$$

Additionen und

$$\sum_{k=1}^{p}(N-k+1)(p-k)=\frac{1}{6}(p-1)p(-2-3N+p)=\frac{1}{2}Np^2-\frac{1}{6}p^3-\frac{1}{2}Np+\frac{1}{2}p^2-\frac{1}{3}p$$

Multiplikationen an.
Allerdings steht in beiden Fällen hierbei Q nicht explizit zur Verfügung, sondern lediglich die Matrizen H_k bzw. die Vektoren $v(x_k)$.[5] Zur Lösung des Kleinste-Quadrate-Problems, d.h. insbesondere für Schritt 3, genügt dies aber. Wir untersuchen, wie viele Rechenoperationen im Fall $N > p$ benötigt werden, um $c = Q_1^{\mathrm{T}}U(y-y_0)$ zu ermitteln. Im Fall

[4] Durch die etwas geschicktere Berechnung von H_{k+1} sind dies weniger Rechenoperationen als in Hämmerlin u. Hoffmann (1991, Abschnitt 2.3.4).

[5] Will man Q bzw. Q_1 explizit bestimmen, so erfordert dies $4(N^2p-Np^2+p^3/3)$ bzw. $2p^2(N-p/3)$ Rechenoperationen, wenn man hierbei die Glieder kleinerer Ordnung weglässt, siehe Higham (2002, Section 19.2).

$N = p$ ändern sich nur Glieder kleinerer Ordnung. Für $w = U(y - y_0)$ werden dabei $p(N-1)$ Additionen, N Subtraktionen und pN Multiplikationen benötigt. Bezeichnet man $w_1 = w$, so kann man rekursiv $w_{k+1} = H_k w_k$ berechnen, die ersten p Einträge von w_{p+1} sind dann der gesuchte Vektor c. Zerlegt man

$$w_k = \begin{pmatrix} w_{k,1} \\ w_{k,2} \end{pmatrix}$$

mit $w_{k,1} \in \mathbb{R}^{k-1}$ und $w_{k,2} \in \mathbb{R}^{N-k+1}$, so ist

$$H_k w_k = \begin{pmatrix} w_{k,1} \\ H_{v(x_k)} w_{k,2}, \end{pmatrix}$$

wobei

$$H_{v(x_k)} w_{k,2} = w_{k,2} - v(x_k)\beta(x_k)v^{\mathrm{T}}(x_k)w_{k,2}.$$

Die Bestimmung von $v^{\mathrm{T}}(x_k)w_{k,2}$ benötigt $N-k+1$ Multiplikationen und $N-k$ Additionen, die anschließende Multiplikation mit $\beta(x_k)$ benötigt eine Multiplikation, und die Multiplikation mit dem Vektor $v(x_k)$ benötigt $N-k+1$ Multiplikationen und die Subtraktion vom Vektor $w_{k,2}$ benötigt $N-k+1$ Subtraktionen. Über alle $k = 1, \ldots, p$ und mit der Berechnung von w fallen daher

$$p(N-1) + N + \sum_{k=1}^{p}(2(N-k)+1) = 3Np + N - p^2 - p$$

Additionen/Subtraktionen und

$$pN + \sum_{k=1}^{p}(2(N-k+1)+1) = 3Np - p^2 + 2p$$

Multiplikationen in Schritt 3 für die Bestimmung von c an.

A.5. Rechenoperationen bei der Singulärwertzerlegung

Für die Rechenoperationen in Schritt 1 des Verfahrens basierend auf der Singulärwertzerlegung vergleiche die Ausführungen zur QR-Zerlegung und zur Cholesky-Zerlegung. Wir werden im Folgenden untersuchen, wie viele Rechenoperationen die Singulärwertzerlegung einer Matrix $C \in \mathbb{R}^{N\times p}$ benötigt. Wir werden dabei $N > p$ annehmen, da man für $N < p$ ansonsten die Zerlegung für C^{T} betrachtet und daher in den Formeln entsprechend N und p zu vertauschen sind bzw. für $N = p$ sich nur Koeffizienten zu Gliedern der Ordnung 2 oder kleiner in N und p ändern.

Der klassische Algorithmus von Golub u. Kahan (1965) bzw. Golub u. Reinsch (1970) besteht aus zwei Phasen.

In der ersten Phase wird analog zur QR-Zerlegung mittels Householder-Matrizen $P_k = P_{v(x_k)}$, $k = 1, \ldots, p$, und $Q_k = Q_{v(y_k)}$, $k = 1, \ldots, p-2$, erreicht, dass

$$\tilde{C} = P_p \cdots P_1 C Q_1 \cdots Q_{p-2} = \begin{pmatrix} q_1 & s_2 & 0 & \ldots & 0 \\ & q_2 & s_3 & \ddots & \vdots \\ & & \ddots & \ddots & 0 \\ & \mathbf{0} & & q_{p-1} & s_p \\ & & & & q_p \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ & & \mathbf{0}_{(N-p)\times p} & & \end{pmatrix}$$

eine bidiagonale Matrix ist.
Die erste Phase benötigt zur Bestimmung von $v(x_k)$, $k = 1, \ldots, p$, und $v(y_k)$, $k = 1, \ldots, p-2$, und $\tilde{C}$ für die Linksmultiplikationen

$$\sum_{k=1}^{p} N - k + 1 + (N-k)(p-k) + (N-k+1)(p-k) = Np^2 - \frac{1}{3}p^3 + \frac{1}{3}p$$

Additionen/Subtraktionen, vgl. Formel (A.8), und

$$\sum_{k=1}^{p} (N-k+1) + 1 + (N-k+1)(p-k) + (p-k) + (N-k+1)(p-k) = Np^2 - \frac{1}{3}p^3 + p^2 + \frac{1}{3}p$$

Multiplikationen, vgl. Formel (A.9), und für die Rechtsmultiplikationen

$$\sum_{k=1}^{p-2} (p-k-1) + 1 + (p-k-1)(N-k) + (p-k)(N-k) = Np^2 - \frac{1}{3}p^3 - 2Np + p^2 + \frac{1}{3}p - 2$$

Additionen/Subtraktionen und

$$\sum_{k=1}^{p-2} (p-k) + 1 + (p-k)(N-k) + (N-k) + (p-k)(N-k) = Np^2 - \frac{1}{3}p^3 - 4N + \frac{13}{3}p - 6$$

Multiplikationen, ferner insgesamt $2p-2$ Divisionen, $2p-2$ Quadratwurzeln und $2p-2$ Vorzeichenermittlungen und eventuelle bis zu $2p-2$ Vorzeichenänderungen. Insgesamt fallen also in Phase 1

$$2Np^2 - \frac{2}{3}p^3 - 2Np + p^2 + \frac{2}{3}p - 2$$

Additionen/Subtraktionen und

$$2Np^2 - \frac{2}{3}p^3 + p^2 - 4N + \frac{20}{3}p - 8$$

Multiplikationen/Divisionen und $2p-2$ Quadratwurzeln an.

In der zweiten Phase wird mit Hilfe von Givens-Rotationen aus der bidiagonalen Matrix eine diagonale Matrix gebildet.

Definition A.1
Für $n \in \mathbb{N}$, $i \neq j \in \{1, \ldots, n\}$ und $\vartheta \in [0, 2\pi[$ heißt hierbei die Drehmatrix $G(i, j, \phi) = (g_{ij})_{i,j \in \{1,\ldots,n\}} \in \mathbb{R}^{n \times n}$ mit

$$g_{kl} = \begin{cases} 1, & k = l, k \neq i, k \neq j, \\ \cos(\phi), & k = l = i, \\ \sin(\phi), & k = i, l = j, \\ -\sin(\phi), & k = j, l = i, \\ \cos(\phi), & k = j, l = j, \\ 0, & \text{sonst}, \end{cases}$$

Givens-Rotation in der Ebene (i, j) mit dem Drehwinkel ϕ.

Um klarer zum Ausdruck zu bringen, dass man für die Matrix $G(i, j, \phi)$ nicht explizit den Winkel ϕ, sondern nur $\cos(\phi)$ und $\sin(\phi)$ benötigt, führen wir die Bezeichnung $\tilde{G}(i, j, c, s) = G(i, j, \phi(c, s))$ ein, wobei $\phi(c, s)$ den Winkel $\phi(c, s) \in [0, 2\pi[$ bezeichne mit $\cos(\phi(c, s)) = c$ und $\sin(\phi(c, s)) = s$, d.h. $\tilde{G}(i, j, c, s) = (\tilde{g}_{ij})_{i,j \in \{1,\ldots,n\}}$ mit

$$\tilde{g}_{kl} = \begin{cases} 1, & k = l, k \neq i, k \neq j, \\ c, & k = l = i, \\ s, & k = i, l = j, \\ -s, & k = j, l = i, \\ c, & k = j, l = j, \\ 0, & \text{sonst}. \end{cases}$$

Für nähere Erläuterungen und Motivationen zu den folgenden drei Algorithmenmodulen für die zweite Phase, d.h. die Singulärwertzerlegung einer bidiagonalen Matrix, verweisen wir auf Golub u. Van Loan (1996).

Algorithmus A.2 (Entkoppelung)
Input: Bidiagonal-Matrix

$$C = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \ldots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & \vdots \\ \vdots & \ddots & \alpha_3 & \ddots & 0 \\ \vdots & & \ddots & \ddots & \beta_{n-1} \\ 0 & \ldots & \ldots & 0 & \alpha_n \end{pmatrix}$$

mit $\alpha_i = 0$, für ein $i = 1, \ldots, n$.

Output:

$$C_{neu} = G_1 \cdots G_{i-1} C \check{G}_i \cdots \check{G}_{n-1} = \begin{pmatrix} C_1 & & \\ & 0 & \\ & & C_2 \end{pmatrix}$$

mit einer Matrix $C_1 \in \mathbb{R}^{(i-1)\times(i-1)}$ und einer Matrix $C_2 \in \mathbb{R}^{(n-i)\times(n-i)}$.

for $k = i+1$ to n **do**
$c \leftarrow -\frac{c_{k,k}}{\sqrt{c_{k,k}^2+c_{i,k}^2}}$; $s \leftarrow \frac{c_{i,k}}{\sqrt{c_{k,k}^2+c_{i,k}^2}}$
$C \leftarrow G(i,k,c,s)C$
end for{nun Zeile i Nullzeile, jetzt noch Spalte i zu Nullzeile machen}
for $k = i-1$ to 1 **do**
$c \leftarrow -\frac{c_{k,k}}{\sqrt{c_{k,k}^2+c_{i,k}^2}}$; $s \leftarrow \frac{c_{i,k}}{\sqrt{c_{k,k}^2+c_{i,k}^2}}$
$C \leftarrow CG(k,i,c,s)$
end for

Algorithmus A.3 (Golub-Kahan-Schritt)
Input: Bidiagonal-Matrix

$$C = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & \vdots \\ \vdots & \ddots & \alpha_3 & \ddots & 0 \\ \vdots & & \ddots & \ddots & \beta_{n-1} \\ 0 & \dots & \dots & 0 & \alpha_n \end{pmatrix}$$

mit $\alpha_i \neq 0$, $i = 1,\dots,n$ und $\beta_i \neq 0$, $i = 1,\dots,n-1$.
Output: $C_{\text{neu}} = G_1 \cdots G_{n-1} C \check{G}_1 \cdots \check{G}_{n-1}$

$a \leftarrow \alpha_{n-1}^2 + \beta_{n-2}^2$; $b \leftarrow \alpha_{n-1}\beta_{n-1}$; $c \leftarrow \alpha_n^2 + \beta_n^2$
if $a \leq c$ **then**
$\lambda = \sqrt{b^2 + \left(\frac{a-c}{2}\right)^2} + \frac{a+c}{2}$
else
$\lambda = -\sqrt{b^2 + \left(\frac{a-c}{2}\right)^2} + \frac{a+c}{2}$
end if{λ ist der Eigenwert der untersten rechtesten 2×2 Untermatrix von $T = C^{\mathrm{T}}C$, der näher an t_{nn} ist}
$y \leftarrow \alpha_1^2 - \lambda$; $z \leftarrow \alpha_1\beta_1$
for $k = 1$ to $n-1$ **do**
$c \leftarrow -\frac{y}{\sqrt{z^2+y^2}}$; $s \leftarrow \frac{z}{\sqrt{z^2+y^2}}$; $G_k \leftarrow G(k,k+1,c,s)$
$C \leftarrow CG_k$
$y \leftarrow c_{kk}$; $z \leftarrow c_{k+1,k}$
$c \leftarrow \frac{y}{\sqrt{z^2+y^2}}$; $s \leftarrow \frac{z}{\sqrt{z^2+y^2}}$; $\check{G}_k \leftarrow G(k,k+1,c,s)^{\mathrm{T}}$
$C \leftarrow \check{G}_k C$
if $k < n-1$ **then**
$y \leftarrow c_{k,k+1}$; $z \leftarrow c_{k,k+2}$
end if
end for

Für die Darstellung $C = \check{W} C_{\text{neu}} \check{V}^{\mathrm{T}}$ benötigt die Bestimmung von C_{neu} lediglich $O(n)$ Rechenoperationen, die Bestimmung von $\check{W} = G_{n-1}^{\mathrm{T}} \cdots G_1^{\mathrm{T}}$, und die Bestimmung von $\check{V} = \check{G}_1 \cdots \check{G}_{n-1}$ jeweils $6n^2 + O(n)$ Rechenoperationen. Algorithmus A.2 braucht offensichtlich weniger Rechenoperationen.

Algorithmus A.4
Input: Bidiagonal-Matrix

$$C = \begin{pmatrix} \alpha_1 & \beta_1 & 0 & \dots & 0 \\ 0 & \alpha_2 & \beta_2 & \ddots & \vdots \\ \vdots & \ddots & \alpha_3 & \ddots & 0 \\ \vdots & & \ddots & \ddots & \beta_{n-1} \\ 0 & \dots & \dots & 0 & \alpha_n \end{pmatrix}.$$

$\delta = c\epsilon_{\text{Mach}}$ für „c klein“, z.B. $c \in (1, 8]$
Output: $C_{\text{neu}} = U_1 \cdots U_k C \check{V}_1 \cdots \check{V}_k$, so dass C_{neu} diagonal
repeat
  for $i = 1$ to $n - 1$ **do**
    if $|c_{i,i+1}| \leq \delta(|c_{ii}| + |c_{i+1,i+1}|)$ **then**
      $c_{i,i+1} \leftarrow 0$
    end if{Event. auch „kleine“ Diagonaleinträge auf 0 setzen}
  end for
  Finde größtes q und s, so dass

$$C = \begin{pmatrix} C_{11} & & \\ & C_{22} & \\ & & C_{33} \end{pmatrix}$$

  mit einer Diagonalmatrix $C_{33} \in \mathbb{R}^{q \times q}$ und einer Bidiagonalmatrix $C_{22} \in \mathbb{R}^{s \times s}$, deren Einträge in der oberen Nebendiagonalen alle ungleich 0 sind.
  if $q < n$ **then**
    if Ein Diagonaleintrag von C_{22} ist 0 **then**
      Wende Algorithmus A.2 auf C_{22} an
    else
      Wende Algorithmus A.3 auf C_{22} an
    end if
  end if
until $q = n$

Wilkinson (1969) zeigt die Konvergenz dieses Algorithmus. Golub u. Reinsch (1970) berichten, dass in ihren numerischen Experimenten die Zahl der Durchläufe der Repeat-Schleife in Algorithmus A.4 im Mittel kleiner als $2n$ war, geben aber keine weiteren Einzelheiten zu diesen numerischen Experimenten. Es sind keine systematische numerische Untersuchungen der Schleifenanzahl mittels einer Simulation oder gar theoretische

Ergebnisse ausgehend von einer Wahrscheinlichkeitsverteilung für die Start-Matrizen C bekannt. Wir gehen bei den folgenden Überlegungen trotzdem von $O(n)$ Schleifendurchläufen im Erwartungswert aus. Da die Anzahl der Rechenoperationen in den Algorithmen A.2 und A.3 jeweils $O(n)$ ist, fallen im Erwartungswert für die Bestimmung der singulären Werte, d.h. der Diagonalmatrix D in der Singulärwertzerlegung $C = WDV^{\mathrm{T}}$, $O(n^2)$ Rechenoperationen an, für die Bestimmung von V im Erwartungswert $\frac{28}{3}n^3 + O(n^2)$ Rechnenoperationen und für die Bestimmung von $W^{\mathrm{T}}c$ wiederum im Erwartungswert $O(n^2)$ Rechenoperationen. Ingesamt führt dies im Erwartungswert zu $2pN^2 + 4Np^2 + 8p^3 + O(p^2 + pN)$ Rechenoperationen.

Von Lawson u. Hanson (1974) und Chan (1982) wurde ein Verfahren zur Singulärwertzerlegung vorgeschlagen, das darauf beruht, dass man eine QR-Zerlegung vorschaltet, also lediglich ein anderes Vorgehen in Phase 1 hat. Geht man von der gleichen Annahme der erwarteten Schleifendurchläufe in Phase 2 aus, dann führt dies zu insgesamt $2pN^2 + 2Np^2 + 11p^3 + O(p^2 + pN)$ Rechenoperationen, vgl. auch Golub u. Van Loan (1996, Section 5.4.5). Da $2Np^2 + 11p^3 \geq 4Np^2 + 8p^3$ genau dann gilt, wenn $N \geq \frac{3}{2}p$ gilt, ist für große N und p in diesem Fall das Verfahren von Lawson-Hanson-Chan effizienter. Ein weiteres Verfahren wurde von Barlow u. a. (2005) vorgeschlagen, das dann zu insgesamt $2pN^2 + 3Np^2 + \frac{28}{3}p^3 + O(p^2 + pN)$ Rechenoperationen führt, also für große N und p bei $\frac{4}{3}p \leq N \leq \frac{5}{3}p$ effizienter als die anderen beiden Verfahren ist.

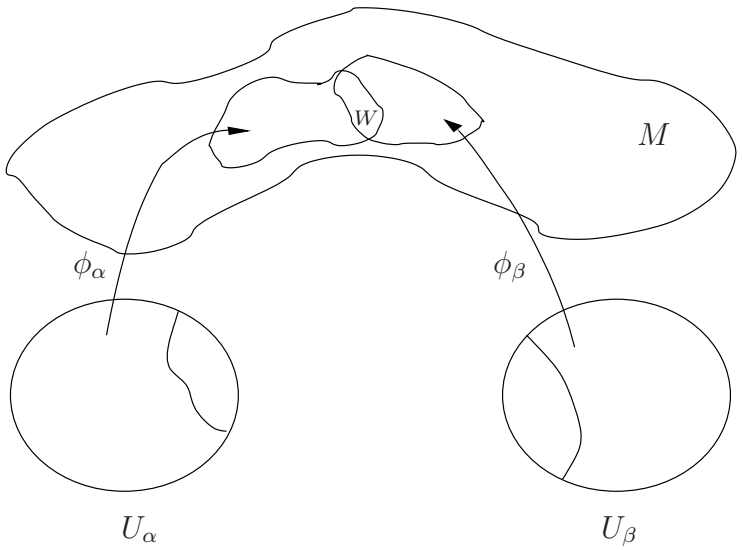
Resultate zur Genauigkeit der Berechnung der Singulärwertzerlegung in Phase 1, also der Berechnung der bidiagonalen Matrix, bei der Verwendung von Gleitkommazahlen endlicher Mantisse finden sich in Wilkinson (1965).

B. Grundlagen aus der Differentialgeometrie

Wir stellen in diesem Abschnitt einige Grundlagen aus der Differentialgeometrie vor und verweisen für Beweise und weitergehende Literatur auf Spivak (1965, 1979), do Carmo (1983), Postnikov (1989) und Barden u. Thomas (2003).

Definition B.1 (Abstrakte Definition von Mannigfaltigkeiten mittels Karten)
Eine n-dimensionale differenzierbare (oder C^k-) **Mannigfaltigkeit** besteht aus einer Menge M und einer Familie $(\phi_\alpha)_{\alpha\in A}$ von injektiven Abbildungen, den sogenannten **Karten** oder **lokalen Parametrisierungen**, $\phi_\alpha : U_\alpha \to M$, mit $U_\alpha \subseteq \mathbb{R}^n$ offen für $\alpha \in A$, so dass

1. die Karten M überdecken („bilden einen **Atlas**“):
$$\bigcup_{\alpha\in A} \phi_\alpha(U_\alpha) = M,$$
2. die Karten verträglich sind („differenzierbarer **Kartenwechsel**“):
Für alle $\alpha, \beta \in A$ mit $W := \phi_\alpha(U_\alpha) \cap \phi_\beta(U_\beta) \neq \emptyset$ ist $\phi_\alpha^{-1}(W)$ offen und der Kartenwechsel $\phi_\beta^{-1} \circ \phi_\alpha : \phi_\alpha^{-1}(W) \to \phi_\beta^{-1}(W)$ ist differenzierbar (oder C^k).



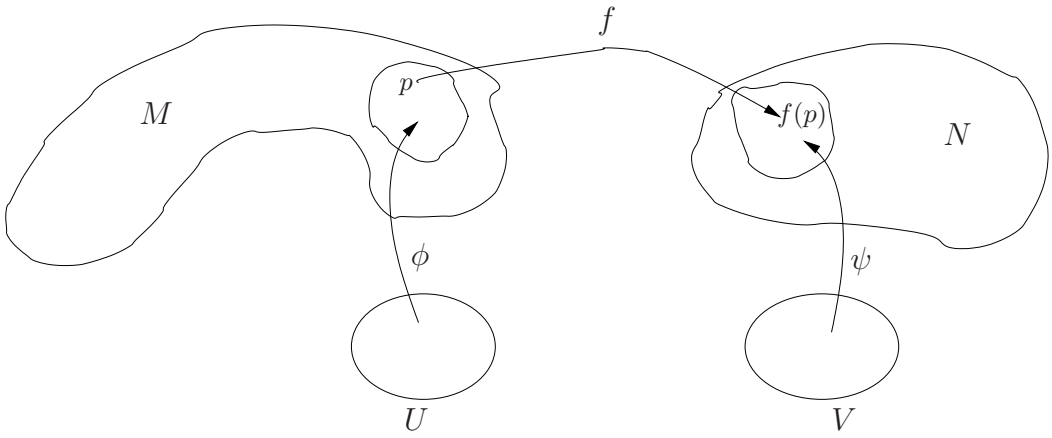
Bemerkung B.2
Das Mengensystem

$$\mathcal{O} = \{O \subseteq M : \phi_\alpha^{-1}(\phi_\alpha(U_\alpha) \cap O) \text{ offen in } U_\alpha \text{ für alle } \alpha \in A\}$$

definiert eine Topologie von offenen Mengen auf M. Diese wird fortan verwendet.

Definition B.3 (Differenzierbarkeit von Abbildungen zwischen Mannigfaltigkeiten)
Seien M und N Mannigfaltigkeiten, U_M eine in M offene Teilmenge, V_M eine in N offene Teilmenge und $f : U_M \to U_N$. Dann heißt f in $p \in U_M$ **differenzierbar**, wenn es Karten $\phi : U \to M$ und $\psi : V \to N$ mit $p \in \phi(U)$ und $f(p) \in \psi(V)$ gibt, so dass $\psi^{-1} \circ f \circ \phi : U \to V$ differenzierbar in $\phi^{-1}(p)$ ist.

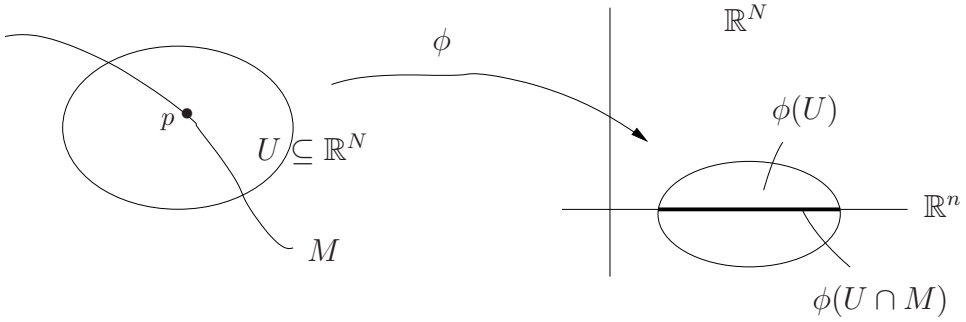
Hierbei betrachtet man mit $\phi_\alpha : U_\alpha \to M$ für $\alpha \in A$ auch alle Einschränkungen von ϕ_α der Form $\tilde{\phi}_\alpha : V_\alpha \to M$ mit $V_\alpha \subset U_\alpha$ offen als Karten, so dass man auf diese Weise gegebenfalls $\phi(U) \subseteq U_M$ und $f(\phi(U)) \subseteq \psi(V)$ erreichen kann.



Seien $U, V \subseteq \mathbb{R}^n$ offen. Eine Abbildung $\phi : U \to V$ heißt C^s**-Diffeomorphismus**, falls ϕ bijektiv ist und ϕ und die Umkehrabbildung ϕ^{-1} jeweils s-mal stetig differenzierbar sind.
Seien M und N Mannigfaltigkeiten. Entsprechend heißt für eine in M offene Teilmenge U und eine in N offene Teilmenge V eine Abbildung $f : U \to V$ ein **Diffeomorphismus**, falls f bijektiv ist und f und f^{-1} differenzierbar sind.

Definition B.4 (Untermannigfaltigkeiten des $\mathbb{R}^N$)
$M \subseteq \mathbb{R}^N$ ist n-dimensionale C^k-**Untermannigfaltigkeit** des $\mathbb{R}^N$, falls es für alle $p \in M$ eine offene Umgebung $U \subseteq \mathbb{R}^N$ mit $p \in U$ und einen C^k-Diffeomorphismus $\phi : U \to \phi(U) \subseteq \mathbb{R}^N$ gibt mit

$$\phi(U \cap M) = \phi(U) \cap (\mathbb{R}^n \times \{\mathbf{0}_{N-n}\}).$$



Bemerkung B.5

- Damit ist mit $\tilde{\phi}_p : \mathrm{pr}^{\mathbb{R}^N}_{\mathbb{R}^n}(\phi(U)) \to M, x \mapsto \phi^{-1}(x, 0)$ für $p \in M$ eine Familie von Karten gegeben und jede Untermannigfaltigkeit des $\mathbb{R}^N$ ist somit eine Mannigfaltigkeit.
- Definition B.1 und B.4 sind noch enger verwandt: Ist nämlich M nach Definition B.1 (mit der vererbten Topologie aus Bemerkung B.2) Hausdorffsch und erfüllt das zweite Abzählbarkeitsaxiom (also abz. Basis der Topologie), so gilt der Satz von Whitney:
  M ist diffeomorph zu einer Untermannigfaltigkeit des $\mathbb{R}^{2n}$.

Lemma B.6 (Untermannigfaltigkeiten durch Gleichungen)
Sei $N > n$ und $U \subseteq \mathbb{R}^N$ offen, $f : U \to \mathbb{R}^{N-n}$ eine differenzierbare C^k-Abbildung und $a \in \mathbb{R}^{N-n}$ ein regulärer Wert von f, d.h. $J_f(u)$ ist surjektiv für alle $u \in U$ mit $f(u) = a$. Dann ist $M = f^{-1}(a)$ eine n-dimensionale C^k-Untermannigfaltigkeit des $\mathbb{R}^N$.

Lemma B.7 (Charakterisierungen einer n-dimensionalen C^k-Untermannigfaltigkeit)
Sei $M \subseteq \mathbb{R}^N$. Folgende Bedingungen sind äquivalent:

a) M ist eine C^k-Untermannigfaltigkeit von $\mathbb{R}^N$ der Dimension n.

b) Es gibt zu jedem Punkt $a \in M$ eine (relativ zu M) offene Umgebung $V_a \subseteq M$, eine offene Teilmenge $U_a \subseteq \mathbb{R}^n$ und eine C^k-Abbildung $\phi_a : U_a \to V_a$ mit $\mathrm{rg}(J_{\phi_a}(u)) = n$ für alle $u \in U_a$ (eine C^k-Immersion), die U_a homöomorph auf V_a abbildet, d.h. die Umkehrfunktion ϕ_a^{-1} ist auch stetig.

c) M ist lokal Urbild eines regulären Wertes $a \in \mathbb{R}^{N-n}$ einer C^k-Funktion, d.h. für alle $p \in M$ gibt es eine offene Umgebung V, so dass $M \cap V = f^{-1}(a)$ für eine Funktion f und einen regulären Wert $a \in \mathbb{R}^{N-n}$ von f wie in Lemma B.6.

d) M ist lokal der Graph einer C^k-Funktion über der n-dimensionalen Koordinatenebene, d.h. es gibt für alle $p \in M$ nach eventueller Umnummerierung der Koordinaten offene Umgebungen $U \in \mathbb{R}^n$ von $(p_1, \ldots, p_d)$ und $V \in \mathbb{R}^{N-n}$ von $(p_{d+1}, \ldots, p_n)$ sowie eine C^k-differenzierbare Abbildung $F : U \to V$, so dass

$$M \cap (U \times V) = \{(x, F(x)) : x \in U\}.$$

Bemerkung B.8 (Quotientenmannigfaltigkeiten)
Sei M eine Mannigfaltigkeit und $(\Gamma, \circ, \mathrm{id})$ eine Gruppe von Diffeomorphismen, die auf M **operieren**, d.h.

1. $g \circ h(p) = g(h(p))$ für alle $g, h \in \Gamma$, $p \in M$,
2. $\mathrm{id}(p) = p$ für alle $p \in M$.

Man sagt, Γ operiert **eigentlich diskontinuierlich**[1] und **fixpunktfrei**[2], falls für alle $p \in M$ eine Umgebung $U(p)$ existiert, so dass für alle $g \in \Gamma \setminus \{\mathrm{id}\}$ gilt

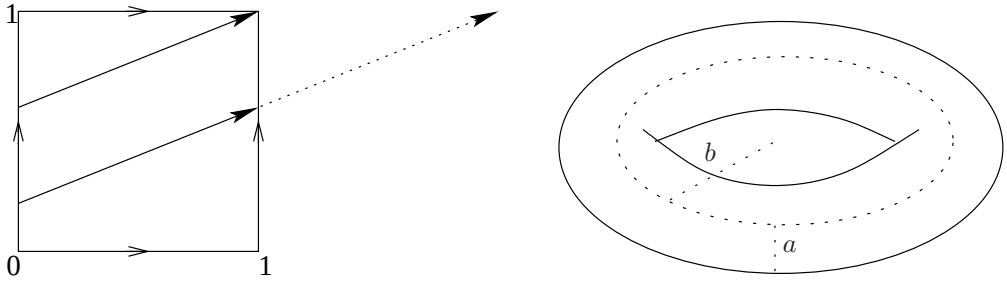
$$g(U(p)) \cap U(p) = \emptyset. \tag{B.1}$$

Dann bilden die Äquivalenzklassen[3] $M/\Gamma := \{\Gamma(x) | \, x \in M\}$, der sogenannte **Orbitraum**, in kanonischer Weise eine Mannigfaltigkeit der gleichen Dimension wie M. Genauer: Die kanonische Einbettung $\pi : M \to M/G$ ist ein lokaler Diffeomorphismus.[4]

Beispiel B.9 (Torus)
Sei $T = \mathbb{R}^2/\mathbb{Z}^2$, wobei $\mathbb{Z}^2$ mittels Translation auf $\mathbb{R}^2$ operiert, d.h. die Äquivalenzrelation $\sim$ ist für $x, y \in \mathbb{R}^2$ wie folgt definiert:

$$x \sim y :\Longleftrightarrow x - y \in \mathbb{Z}^2.$$

Dann ist der Orbitraum T, der sogenannte flache Torus, diffeomorph zum zweidimensionalen Rotationstorus $T^2_{a,b} \subseteq \mathbb{R}^3$, $b > a$ („Autoschlauch“).



Der Diffeomorphismus vom flachen Torus auf den Rotationstorus ist gegeben durch:

$$g : \mathbb{R}^2/\mathbb{Z}^2 \to T^2_{a,b},\ (x, y) + \mathbb{Z}^2 \mapsto \begin{pmatrix} (b + a\cos(2\pi y))\cos(2\pi x) \\ (b + a\cos(2\pi y))\sin(2\pi x) \\ a\sin(2\pi y) \end{pmatrix}.$$

Man kann zeigen, dass $\mathbb{R}^2/\mathbb{Z}^2$ zudem diffeomorph zu $S^1 \times S^1 \subset \mathbb{R}^4$. Dies zeigt, dass es keinen kanonischen umgebenden Euklidischen Raum (im Satz von Whitney) gibt, in dem eine Mannigfaltigkeit liegt.

[1] Γ operiert eigentlich diskontinuierlich, wenn der Schnitt in Gleichung (B.1) für $g \in \Gamma \setminus \{\mathrm{id}\}$ endlich ist.
[2] Γ operiert fixpunktfrei, falls alle $g \in \Gamma \setminus \{\mathrm{id}\}$ keinen Fixpunkt haben.
[3] Die Äquivalenzrelation $\sim$ ist also wie folgt definiert: $x \sim y :\Longleftrightarrow x = g(y)$ für ein $g \in \Gamma$, d.h. x und y sind in gleicher Bahn unter Γ.
[4] Daher ist $\pi \circ \phi_\alpha$ nach Einschränkung Karte für M/Γ, wenn ϕ_α Karte für M.

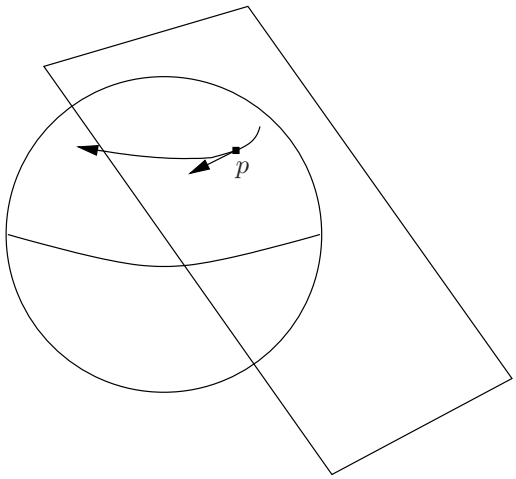
Definition B.10 (Tangentialraum)
Sei M eine n-dimensionale C^k-Untermannigfaltigkeit des $\mathbb{R}^N$.
Dann ist der **Tangentialraum** im Punkt $p \in M$ gegeben durch

$$T_pM := \left\{ \frac{\partial}{\partial t}\Big|_{t=0} \gamma \mid \gamma : (-\epsilon, \epsilon) \to M \text{ ist } C^k\text{-differenzierbar}, \gamma(0) = p, \epsilon > 0 \right\} .$$

Dieser wird meist mit (p, T_pM) identifiziert. Die damit disjunkte Vereinigung

$$TM := \bigcup_{p \in M}^{\bullet} T_pM$$

der Tangentialräume heißt **Tangentialbündel**.



Falls M keine Untermannigfaltigkeit des $\mathbb{R}^N$ ist, sondern eine abstrakte Mannigfaltigkeit, so ist diese Definition nicht möglich und wir müssen den Tangentialraum abstrakter definieren und können ihn dann auch nicht mehr wie oben als Unterraum des $\mathbb{R}^N$ auffassen:
Sei $C^\infty(M,p) := \{f : M \to \mathbb{R} \mid f \text{ ist } C^\infty \text{ in Umgebung von } p\}$. Dann definiert man:

$$T_pM := \big\{ v : C^\infty(M,p) \to \mathbb{R} \mid \text{ es gibt } \epsilon > 0, \gamma : (-\epsilon, \epsilon) \to M \text{ mit } \gamma(0) = p \text{ und } v(f) = \frac{\partial}{\partial t}\Big|_{t=0} f \circ \gamma \text{ für alle } f \in C^\infty(M,p) \big\}.$$

Bemerkung B.11
T_pM ist ein n-dimensionaler linearer Raum. Falls ϕ Karte für $p = \phi(x)$, so ist $T_pM = \text{Bild}(J_\phi(x))$. Weiterhin ist das Tangentialbündel TM eine Mannigfaltigkeit der Dimension $2n$, da

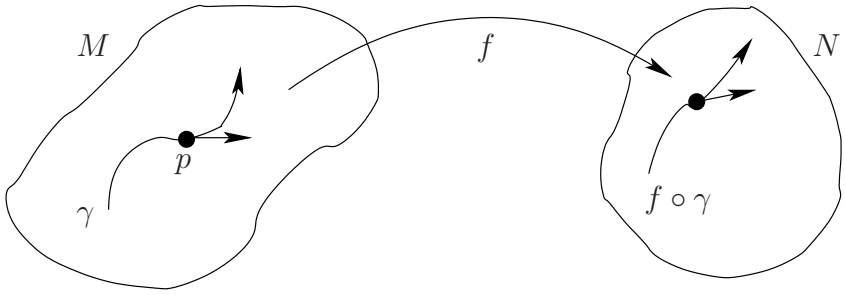
$$\tilde{\phi}_\alpha : U_\alpha \times \mathbb{R}^n \to TM,\ (x,a) \mapsto \left(\phi_\alpha(x), \sum_{i=1}^n a_i \frac{d}{dt}\Big|_{t=0} \phi_\alpha(x + te_i) \right)$$

Karte für $\alpha \in A$ ist.

Definition B.12 (Ableitung von Funktionen)
Sei $f : M \to N$ eine differenzierbare Abbildung (nach Definition B.3). Die Abbildung $f_{*p} : T_pM \to T_{f(p)}N$ mit

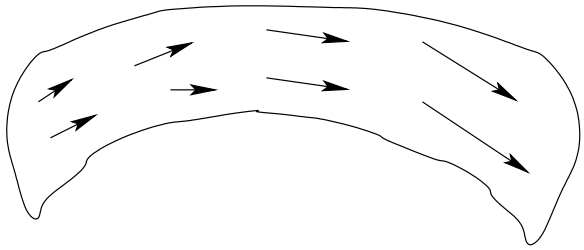
$$f_{*p}\left(\frac{\partial}{\partial t}\Big|_{t=0}\gamma\right) := \frac{\partial}{\partial t}\Big|_{t=0} f \circ \gamma$$

für differenzierbare Kurven $\gamma : (-\epsilon, \epsilon) \to M$ mit $\gamma(0) = p$ und $\epsilon > 0$, heißt **Ableitung** von f in p.



Bemerkung B.13
$f_{*p} : T_pM \to T_{f(p)}M$ ist eine lineare Abbildung. Weiterhin gelten die üblichen Gesetze für Ableitungen wie Kettenregel und implizites Funktionentheorem.

Definition B.14 (Vektorfelder)
Eine differenzierbare Abbildung $X : M \to TM$ mit $X(p) \in T_pM$ für $p \in M$ heißt **differenzierbares Vektorfeld**.

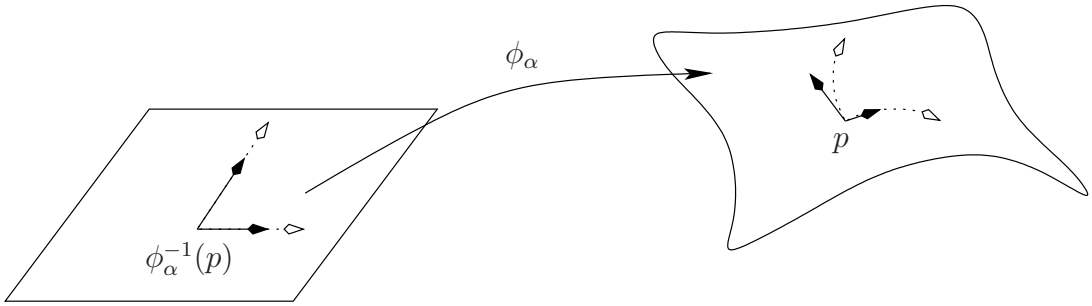


Bemerkung B.15

- Aufgrund der Definition B.3 und der Bemerkung B.11 ist erklärt, was in diesem Zusammenhang mit differenzierbar gemeint ist.
- Die Vektorfelder $\frac{\partial}{\partial x_i} : \phi_\alpha(U_\alpha) \subseteq M \to TM$ mit

$$\frac{\partial}{\partial x_i}(p) := \frac{\partial}{\partial x_{ip}} := \left(p, \frac{d}{dt}\Big|_{t=0}\phi_\alpha(\phi_\alpha^{-1}(p) + te_i)\right)$$

für $i = 1, \ldots, n$ heißen **Basisvektorfelder** (der Karte ϕ_α).



Definition B.16 (Riemannsche Mannigfaltigkeiten)
Sei M eine Mannigfaltigkeit. Dann heißt eine Familie $g = (g_p)_{p \in M}$ von (pos. def.) Skalarprodukten $g_p : T_pM \times T_pM \to \mathbb{R}$ für $p \in M$ **Riemannsche Metrik**, wenn die Skalarprodukte differenzierbar von $p \in M$ abhängen, d.h. ist $U \subseteq M$ offen und sind X, Y differenzierbare Vektorfelder auf U, so ist $p \mapsto g_p(X(p), Y(p))$ differenzierbar.
Eine Mannigfaltigkeit M mit Riemannscher Metrik g heißt **Riemannsche Mannigfaltigkeit** (M, g).

Bemerkung B.17 (Messen von Kurvenlängen)
Mittels der Skalarprodukte g_p für $p \in M$ definiert man für stückweise differenzierbare Kurven $c : [a, b] \to M$ die Länge der Kurve durch

$$L(c) := \int_a^b \| \dot{c}(t) \|_{g_{c(t)}} \, dt = \int_a^b \left(g_{c(t)} \left(\dot{c}(t), \dot{c}(t) \right) \right)^{1/2} dt \, .$$

Ist (M, g) eine (weg-)zusammenhängende[5] Riemannsche Mannigfaltigkeit, so definiert man die **Riemannsche Abstandsfunktion** $d : M \times M \to \mathbb{R}$:

$$d(p, q) := \inf\{ L(c) \,|\, c : [a, b] \to M \text{ stückw. diffb.}, \, c(a) = p, \, c(b) = q \} \, .$$

Mit dieser Abstandsfunktion ist (M, d) ein metrischer Raum, wobei die durch d gegebene Topologie mit der ursprünglichen übereinstimmt.

Beispiel B.18

- $\mathbb{R}^n$ *als Riemannsche Mannigfaltigkeit*
  $\mathbb{R}^n$ zusammen mit dem üblichen Skalarprodukt auf $T_p\mathbb{R}^n \cong \mathbb{R}^n$ für $p \in \mathbb{R}^n$ bildet eine Riemannsche Mannigfaltigkeit.

[5] Für differenzierbare Mannigfaltigkeiten sind die beiden Zusammenhangsbegriffe äquivalent, da es zu jedem Punkt eine wegzusammenhängende Umgebung gibt.

- *Induzierte Metrik*
  Ist N Untermannigfaltigkeit einer Riemannschen Mannigfaltigkeit M mit Metrik g, so kann man für $p \in N$ den Tangentialraum T_pN in natürlicher Weise als Unterraum von T_pM auffassen, genauer:
  Bezeichnet $i : N \to M$ die Inklusion, so identifiziere (für $p \in N$) $v \in T_pM$ mit $i_{*p}v \in T_pN$.
  Dann ist N mit der Einschränkung der Skalarprodukte von TM auf TN eine Riemannsche Mannigfaltigkeit.
  Beispiele: Die Einheitskugel $S^{n-1} \subseteq \mathbb{R}^n$, Rotationstorus $T^2_{a,b} \subseteq \mathbb{R}^3$.

Definition B.19 (Isometrien)
Seien (M, g) und (N, h) zwei Riemannsche Mannigfaltigkeiten. Dann heißt ein Diffeomorphismus $f : M \to N$ eine **Isometrie**, wenn $f_{*p} : T_pM \to T_{f(p)}N$ für $p \in M$ eine Isometrie ist, d.h.

$$h_{f(p)}(f_{*p}v, f_{*p}w) = g_p(v, w) \quad \text{für alle } p \in M,\ v, w \in T_pM.$$

Bemerkung B.20 (Satz von Nash)
Es gilt auch für Riemannsche Mannigfaltigkeiten das Analogon zum Satz von Whitney (vergleiche Bemerkung B.5), nämlich der Satz von Nash:
Jede Riemannsche Mannigfaltigkeit M der Dimension n ist isometrisch zu einer Untermannigfaltigkeit des $\mathbb{R}^N$ mit induzierter Metrik. Hierbei kann man $N \leq n^2 + 5n + 3$ wählen.

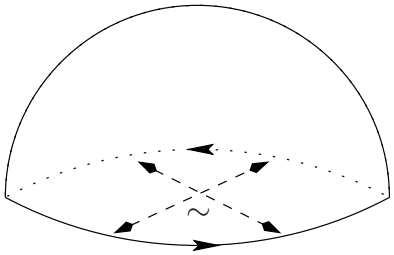
Bemerkung B.21 (Riemannsche Quotientenmannigfaltigkeiten)
Ist (M, g) eine Riemannsche Mannigfaltigkeit und Γ eine Gruppe von Isometrien, die eigentlich diskontinuierlich und fixpunktfrei auf M operieren, dann besitzt der Orbitraum M/Γ genau eine Riemannsche Metrik $\tilde{g}$, so dass die kanonische Einbettung $\pi : M \to M/\Gamma$ eine lokale Isometrie ist, also so dass $\pi_{*\pi(p)} : T_pM \to T_{\pi(p)}M/\Gamma$ eine Isometrie von Vektorräumen ist.

Beispiel B.22 (Projektiver Raum)
Der (reelle) **projektive Raum** $\mathbb{R}P^n := S^n/\{\pm\mathrm{id}\}$ entsteht aus der Sphäre S^n im $\mathbb{R}^{n+1}$ durch Identifikation der Antipoden,[6] d.h. für $x, y \in S^n$ gilt

$$x \sim y :\Longleftrightarrow y = -x.$$

Er besitzt also eine Metrik, die lokal isometrisch zur Sphäre S^n ist.

[6] Ein anderes Modell hierfür wäre $(\mathbb{R}^n \setminus \{0\})/(\mathbb{R} \setminus \{0\})$ (Raum der Geraden), wobei $\mathbb{R} \setminus \{0\}$ mittels Skalarmultiplikation auf $\mathbb{R}^n \setminus \{0\}$ operiert, d.h. $x \sim y \Longleftrightarrow x = \lambda y$ für ein $\lambda \neq 0$.



Der (flache) Torus $\mathbb{R}^2/\mathbb{Z}^2$ besitzt eine Metrik, so dass er lokal isometrisch zum $\mathbb{R}^2$ ist. Er ist dann allerdings nicht isometrisch zum Rotationstorus $T^2_{a,b}$, wenn dieser mit der induzierten Metrik des $\mathbb{R}^3$ ausgestattet ist.

Definition B.23 (Integration auf Mannigfaltigkeiten)
Sei (M, g) eine Riemannsche Mannigfaltigkeit, etwa eine Untermannigfaltigkeit des $\mathbb{R}^N$ mit induzierter Metrik. Sei $f : U \to \mathbb{R}$ mit $U \subseteq M$ offen, und $\phi : \mathbb{R}^n \supseteq V \to M$ Karte von M mit $\phi(V) \supseteq U$, d.h. ϕ ist globale Karte[7] von U. Dann definiert man

$$\int_U f \, dM := \int_{\phi^{-1}(U)} f \circ \phi(x) \sqrt{\det(g_{ij}(x))} \, dx \, ,$$

wobei $g_{ij}(x) := g_{\phi(x)}(\phi_{*x}(e_i), \phi_{*x}(e_j))$ für $i, j \in \{1, \ldots, n\}$. Für Untermannigfaltigkeiten des $\mathbb{R}^N$ mit induzierter Metrik gilt $g_{ij}(x) = <\frac{\partial \phi}{\partial x_i}(x), \frac{\partial \phi}{\partial x_j}(x)>$. Insbesondere setzt man

$$\mathrm{vol}_M(U) := \int_{\phi^{-1}(U)} \sqrt{\det(g_{ij}(x))} \, dx \, .$$

Ist $\mathrm{vol}_M(M) < \infty$, so würde man $P : \mathcal{B}_M \to [0, 1]$, $P(B) = \mathrm{vol}_M(B)/\mathrm{vol}_M(M)$ als Gleichverteilung auf der Borelschen σ-Algebra $\mathcal{B}_M$ (erzeugt durch die offenen Mengen von M) auffassen.

Bemerkung B.24
Diese Definition ist kompatibel mit dem Messen von Kurvenlängen. Ist $c : [a, b] \to M$ eine reguläre (d.h. $\dot{c}(t) \neq 0$) und injektive Kurve, so gilt, da $c : [a, b] \to \mathrm{Bild}(c)$ Karte von $\mathrm{Bild}(c) = c[a, b]$ ist,

$$\mathrm{vol}_{c[a,b]}(c[a, b]) = \int_{[a,b]} \sqrt{\det(g_{c(t)}(\dot{c}(t), \dot{c}(t)))} \, dt = \int_{[a,b]} \| \dot{c}(t) \|_{g_{c(t)}} \, dt = L(c).$$

Weiterhin sind die Formeln konsistent, wenn man $M = \mathbb{R}^n$ als Untermannigfaltigkeit des $\mathbb{R}^N$ auffasst.

[7] Falls es keine globale Karte gibt, so behilft man sich, indem man das Integral mit Hilfe einer (wenn nötig differenzierbaren) Zerlegung der Eins in Gebiete aufspaltet, für die es globale Karten gibt.

Literaturverzeichnis

Abramowitz u. Stegun 1972
ABRAMOWITZ, Milton ; STEGUN, Irene A.: *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. 10th printing. Washington, D.C.: For sale by the Superintendent of Documents, U.S. Government Printing Office, 1972 [154]

Al-Baali 1985
AL-BAALI, M.: Descent property and global convergence of the Fletcher-Reeves method with inexact line search. In: *IMA J. Numer. Anal.* 5 (1985), Nr. 1, S. 121–124 [66]

Altenbach 1999
ALTENBACH, Holm (Hrsg.) ; CISM Advanced School on „Modelling of Creep and Damage Processes in Materials and Structures“ (1998, Udine) (Veranst.): *Creep and damage in materials and structures*. Wien et al.: Springer, 1999 [138]

Andrew u. Tan 1999
ANDREW, Alan L. ; TAN, Roger C. E.: Computation of derivatives of repeated eigenvalues and the corresponding eigenvectors of symmetric matrix pencils. In: *SIAM Journal on Matrix Analysis and Applications* 20 (1999), Nr. 1, S. 78–100 [146]

Babuška u. Osborn 1991
BABUŠKA, I. ; OSBORN, J.: Eigenvalue Problems. In: CIARLET, Philippe G. (Hrsg.): *Finite element methods (Part 1)* Bd. 2. Amsterdam: North-Holland, 1991, S. 641–787 [139], [140]

Barden u. Thomas 2003
BARDEN, Dennis ; THOMAS, Charles: *An introduction to differential manifolds*. London: Imperial College Press, 2003 [165]

Barlow u. a. 2005
BARLOW, Jesse L. ; BOSNER, Nela ; DRMAČ, Zlatko: A new stable bidiagonal reduction algorithm. In: *Linear Algebra and its Applications* 397 (2005), S. 35–84 [164]

Bates u. Watts 1981
BATES, Douglas M. ; WATTS, Donald G.: A relative offset orthogonality convergence criterion for nonlinear least squares. In: *Technometrics* 23 (1981), S. 179–183 [98]

Bates u. Watts 1988
Bates, Douglas M. ; Watts, Donald G.: *Nonlinear Regression and Its Application.* New York et al.: John Wiley & Sons, 1988 [4], [36], [98], [123]

Bauer 1966
Bauer, F. L.: Genauigkeitsfragen bei der Lösung linearer Gleichungssysteme. In: *Zeitschrift für Angewandte Mathematik und Mechanik* 46 (1966), S. 409–421 [49]

Bazaraa u. a. 2006
Bazaraa, Mokhtar S. ; Sherali, Hanif D. ; Shetty, C. M.: *Nonlinear Programming: Theory and Algorithms.* 3rd edition. New York et al.: John Wiley & Sons, 2006 [61]

Beale 1960
Beale, E.M.L.: Confidence Regions in Non-linear Estimation. In: *Journal of the Royal Statistical Society B* 22 (1960), S. 41–88 [36]

Ben-Israel u. Greville 2003
Ben-Israel, Adi ; Greville, Thomas N. E.: *Generalized inverses.* 2nd edition. New York: Springer-Verlag, 2003 [46]

Berger 1985
Berger, James O.: *Statistical decision theory and Bayesian analysis.* 2nd edition. New York: Springer-Verlag, 1985 [110], [112]

Bertsekas 1999
Bertsekas, Dimitri P.: *Nonlinear programming.* 2nd. Belmont, MA: Athena Scientific, 1999 [61], [68], [92], [93], [98]

Bickel u. Li 2006
Bickel, Peter J. ; Li, Bo: Regularization in statistics (with discussion). In: *Test* 15 (2006), Nr. 2, S. 271–344 [109]

Björck 1990
Björck, Åke: Least squares methods. In: *Handbook of numerical analysis, Vol. I.* Amsterdam: North-Holland, 1990 (Handb. Numer. Anal., I), S. 465–652 [55], [58]

Björck 1991
Björck, Åke: Component-wise perturbation analysis and error bounds for linear least squares solutions. In: *BIT. Numerical Mathematics* 31 (1991), Nr. 2, S. 238–244 [49]

Björck 1996
Björck, Åke: *Numerical Methods for Least Squares Problems.* Philadelphia: SIAM, 1996 [61], [88]

Björck 2004
Björck, Åke: The calculation of linear least squares problems. In: *Acta Numer.* 13 (2004), S. 1–53 [61]

Box u. Draper 1987
Box, George E. ; Draper, Norman R.: *Empirical model-building and response surfaces*. New York et al.: John Wiley & Sons, 1987 [4]

Broyden 1970
Broyden, C. G.: The convergence of a class of double-rank minimization algorithms. II. The new algorithm. In: *Journal of the Institute of Mathematics and its Applications* 6 (1970), S. 76–90 und 222–231 [89]

Carlin u. Louis 1996
Carlin, Bradley P. ; Louis, Thomas A.: *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall, 1996 [112]

do Carmo 1983
Carmo, Manfredo P.: *Differentialgeometrie von Kurven und Flächen*. Braunschweig/Wiesbaden: Friedr. Vieweg & Sohn, 1983 [165]

EN 10291:2000
CEN: *Metallic materials - Uniaxial creep testing in tension - Method of test* [135]

Chan 1982
Chan, Tony F.: An improved algorithm for computing the singular value decomposition. In: *ACM Trans. Math. Softw.* 8 (1982), Nr. 1, S. 72–83 [164]

Christensen 2002
Christensen, Ronald: *Plane answers to complex questions*. 3rd edition. New York: Springer-Verlag, 2002 [35]

Clough u. Penzien 2003
Clough, Ray W. ; Penzien, Joseph: *Dynamics of Structures*. 3rd edition. Berkeley: Computers & Structures, 2003 [142]

Collins u. a. 1974
Collins, J. D. ; Hart, G. C. ; Hasselman, T. K. ; Kennedy, B.: Statistical Identification of Structures. In: *AIAA Journal* 12 (1974), Nr. 2, S. 185–190 [114]

Cressie 1993
Cressie, Noel A. C.: *Statistics for spatial data*. New York et al.: John Wiley & Sons, 1993 [149]

Dahlquist u. Björck 2008
Dahlquist, Germund ; Björck, Åke: *Numerical methods in scientific computing. Vol. I*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2008 [108]

Dai 2002
Dai, Yu-Hong: Convergence properties of the BFGS algorithm. In: *SIAM Journal on Optimization* 13 (2002), Nr. 3, S. 693–701 [94]

Davidon 1959
DAVIDON, William C.: *Variable metric method for minimization.* AEC Research Development Report ANL-5990, 1959 - auch: SIAM Journal on Optimization 1 (1991), Nr. 1, S. 1–17 [91]

Demidenko 1989
DEMIDENKO, Eugene: *Optimization and Regresia (in Russian).* Moskau: Nauka, 1989 [14], [22]

Demidenko 2000
DEMIDENKO, Eugene: Is this the least squares estimate? In: *Biometrika* 87 (2000), Nr. 2, S. 437–452 [22]

Demidenko 2006
DEMIDENKO, Eugene: Criteria for global minimum of sum of squares in nonlinear regression. In: *Computational Statistics & Data Analysis* 51 (2006), Nr. 3, S. 1739–1753 [14]

Demidenko 2008
DEMIDENKO, Eugene: Criteria for Unconstrained Global Optimization. In: *Journal of Optimization Theory and Applications* 136 (2008), Nr. 3, S. 375–395 [14]

Dennis 1977
DENNIS, J. E.: Non-linear least squares and equations. In: *The state of the art in numerical analysis (Proc. Conf., Univ. York, Heslington, 1976).* London: Academic Press, 1977, S. 269–312 [97], [97], [123]

Dennis u. Schnabel 1983
DENNIS, John E. ; SCHNABEL, Robert B.: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Englewood Cliffs: Prentice-Hall, 1983 [62], [65], [77], [83], [92], [92], [95], [97]

Dennis u. a. 1981
DENNIS, John E. ; GAY, David M. ; WELSCH, Roy E.: An adaptive nonlinear least-squares algorithm. In: *Association for Computing Machinery. Transactions on Mathematical Software* 7 (1981), Nr. 3, S. 348–368 [95], [98], [122]

DIN 50118
Zeitstandversuch unter Zugbeanspruchung [135]

Draper u. Smith 1998
DRAPER, Norman R. ; SMITH, Harry: *Applied regression analysis.* 3rd edition. New York et al.: John Wiley & Sons, 1998 [4]

Eckart u. Young 1936
ECKART, C. ; YOUNG, G.: The approximation of one matrix by another of lower rank. In: *Psychometrika, Chicago,* 1 (1936), S. 211–218 [46]

Engl u. a. 1996
ENGL, Heinz W. ; HANKE, Martin ; NEUBAUER, Andreas: *Mathematics and its Applications*. Bd. 375: *Regularization of inverse problems*. Dordrecht: Kluwer Academic Publishers Group, 1996 [109]

Eriksson u. a. 2005
ERIKSSON, J. ; WEDIN, P. A. ; GULLIKSSON, M. E. ; SÖDERKVIST, I.: Regularization methods for uniformly rank-deficient nonlinear least-squares problems. In: *Journal of Optimization Theory and Applications* 127 (2005), Nr. 1, S. 1–26 [109]

Evans u. Swartz 1995
EVANS, Michael ; SWARTZ, Tim: Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. In: *Statistical Science* 10 (1995), Nr. 3, S. 254–272 [133]

Evans u. Swartz 2000
EVANS, Michael ; SWARTZ, Tim: *Approximating integrals via Monte Carlo and deterministic methods*. Oxford: Oxford University Press, 2000 [133]

Evans u. Wilshire 1993
EVANS, Russell W. ; WILSHIRE, Brian: *Introduction to Creep*. London: Institute of Materials, 1993 [135], [137], [138]

Fahrmeir u. a. 2007
FAHRMEIR, Ludwig ; KNEIB, Thomas ; LANG, Stefan: *Regression*. New York et al.: Springer, 2007 [4]

Fang u. O'Leary 2008
FANG, Haw-ren ; O'LEARY, Dianne P.: Modified Cholesky algorithms: a catalog with new approaches. In: *Mathematical Programming. A Publication of the Mathematical Programming Society* 115 (2008), Nr. 2, Ser. A, S. 319–349 [54]

Fletcher 1970
FLETCHER, R.: A new approach to variable metric algorithm. In: *The Computer Journal* 13 (1970), Nr. 3, S. 317–322 [89]

Fletcher u. Powell 1963
FLETCHER, R. ; POWELL, M.J.D.: A rapidly convergent descent method for minimization. In: *The Computer Journal* 6 (1963), S. 163–168 [91]

Fletcher u. Xu 1987
FLETCHER, R. ; XU, C.: Hybrid methods for nonlinear least squares. In: *IMA J. Numer. Anal.* 7 (1987), Nr. 3, S. 371–389 [96]

Fletcher 1987
FLETCHER, Roger: *Practical Methods of Optimization*. Chichester et al.: John Wiley & Sons, 1987 [61], [66]

Forster 1984
FORSTER, Otto: *Analysis 3.* 3. Auflage. Braunschweig: Vieweg, 1984 [11], [104]

Foster 1997
FOSTER, Leslie V.: The growth factor and efficiency of Gaussian elimination with rook pivoting. In: *Journal of Computational and Applied Mathematics* 86 (1997), Nr. 1, S. 177–194 [55]

Foster 1998
FOSTER, Leslie V.: Corrigendum: "The growth factor and efficiency of Gaussian elimination with rook pivoting". In: *Journal of Computational and Applied Mathematics* 98 (1998), Nr. 1, S. 177 [55]

Fraley 1989
FRALEY, C.: Computational behavior of Gauss-Newton methods. In: *SIAM J. Sci. Statist. Comput.* 10 (1989), Nr. 3, S. 515–532 [89]

Friswell 1989
FRISWELL, Mike I.: The Adjustment of Structural Parameters using a Minimum Variance Estimator. In: *Mechanical Systems and Signal Processing* 3 (1989), Nr. 2, S. 143–155 [117]

Friswell u. Mottershead 1995
FRISWELL, Mike I. ; MOTTERSHEAD, John E.: *Finite Element Model Updating in Structural Dynamics.* Dordrecht et al.: Kluwer Academic Publishers, 1995 [142]

Gallant 1987
GALLANT, A.Ronald: *Nonlinear Statistical Models.* New York et al.: John Wiley & Sons, 1987 [28]

Gay 1990
GAY, David M.: *Usage Summary For Selected Optimization Routines.* Murray Hill: AT& T Bell Laboratories, 1990 [98]

Gill u. a. 1989
GILL, Philip E. ; MURRAY, Walter ; WRIGHT, Margaret H.: *Practical Optimization.* London: Academic Press, 1989 [61], [62], [98]

Goldfarb 1970
GOLDFARB, Donald: A family of variable-metric methods derived by variational means. In: *Mathematics of Computation* 24 (1970), S. 23–26 [89]

Golub 1965
GOLUB, G.: Numerical methods for solving linear least squares problems. In: *Numerische Mathematik* 7 (1965), S. 206–216 [55]

Golub u. Kahan 1965
GOLUB, G. ; KAHAN, W.: Calculating the singular values and pseudo-inverse of a

matrix. In: *J. Soc. Ind. Appl. Math., Ser. B, Numer. Anal.* 2 (1965), S. 205–224 [159]

Golub u. Van Loan 1996
GOLUB, Gene H. ; VAN LOAN, Charles F.: *Matrix computations*. Third. Baltimore, MD: Johns Hopkins University Press, 1996 [52], [58], [161], [164]

Golub u. Reinsch 1970
GOLUB, G.H. ; REINSCH, C.: Singular value decomposition and least squares solutions. In: *Numerische Mathematik* 14 (1970), Nr. 5, S. 403–420 [159]

Göring u. a. 1993
GÖRING, Herbert ; ROOS, Hans-Görg ; TOBISKA, Lutz: *Finite-Element-Methode*. 3. Auflage. Berlin: Akademie Verlag, 1993 [139]

Grötschel 1991
GRÖTSCHEL, Martin: *Optimierungsmethoden II*. Vorlesungsskript, 1991 [92]

Gurland u. Tripathi 1971
GURLAND, John ; TRIPATHI, Ram C.: A simple approximation for unbiased estimation of the standard deviation. In: *The American Statistician* 25 (1971), Nr. 4, S. 30–32 [30]

Hahn 1992
HAHN, Hans G.: *Technische Mechanik fester Körper*. 2. Auflage. München: Hanser, 1992 [142]

Hämmerlin u. Hoffmann 1991
HÄMMERLIN, Günther ; HOFFMANN, Karl-Heinz: *Numerische Mathematik*. 2. Auflage. Berlin et al.: Springer-Verlag, 1991 [42], [52], [54], [71], [158]

Harville 1997
HARVILLE, David A.: *Matrix Algebra From a Statistician's Perspective*. New York et al.: Springer, 1997 [81], [86], [89], [144]

Hendriks u. Landsman 1998
HENDRIKS, Harrie ; LANDSMAN, Zinoviy: Mean location and sample mean location on manifolds: asymptotics, tests, confidence regions. In: *J. Multivariate Anal.* 67 (1998), Nr. 2, S. 227–243 [103]

Higham 1990
HIGHAM, Nicholas J.: Analysis of the Cholesky decomposition of a semi-definite matrix. In: *Reliable numerical computation*. New York: Oxford Univ. Press, 1990 (Oxford Sci. Publ.), S. 161–185 [53]

Higham 1994
HIGHAM, Nicholas J.: A survey of componentwise perturbation theory in numerical linear algebra. In: *Mathematics of Computation 1943–1993: a half-century of*

computational mathematics (Vancouver, BC, 1993) Bd. 48. Providence, RI: Amer. Math. Soc., 1994, S. 49–77 [49]

Higham 2002
HIGHAM, Nicholas J.: *Accuracy and stability of numerical algorithms*. 2nd edition. Philadelphia: SIAM, 2002 [42], [51], [55], [55], [56], [61], [106], [155], [156], [158]

Hubbard u. a. 2001
HUBBARD, John ; SCHLEICHER, Dierk ; SUTHERLAND, Scott: How to find all roots of complex polynomials by Newton's method. In: *Inventiones Mathematicae* 146 (2001), Nr. 1, S. 1–33 [72]

Huppert 1990
HUPPERT, Bertram: *Angewandte Lineare Algebra*. Berlin: Walter de Gruyter, 1990 [117], [141]

Huppert u. Willems 2006
HUPPERT, Bertram ; WILLEMS, Wolfgang: *Linear algebra*. Wiesbaden: Teubner, 2006 [142]

IEEE 754-2008
IEEE Standards Commitee 754: *IEEE Standard for Floating-Point Arithmetic, ANSI/IEEE Standard 754-2008* [49]

Isenberg 1979
ISENBERG, J.: Progressing from Least Squares to Bayesian Estimation. In: *Proceedings of the 1979 ASME Design Engineering Technical Conference*. New York: ASME: Dynamic Systems & Control Division, 1979, S. 1–11 [109]

Kahan 1966
KAHAN, W.: Numerical linear algebra. In: *Canadian Mathematical Bulletin* 9 (1966), S. 757–801 [42]

Koutková 1992
KOUTKOVÁ, Helena: On estimable and locally-estimable functions in the non-linear regression model. In: *Kybernetika* 28 (1992), Nr. 2, S. 120–128 [20]

Lancaster 1964
LANCASTER, P.: On eigenvalues of matrices dependent on a parameter. In: *Numerische Mathematik* 6 (1964), S. 377–387 [146]

Läuchli 1961
LÄUCHLI, Peter: Jordan-Elimination und Ausgleichung nach kleinsten Quadraten. In: *Numerische Mathematik* 3 (1961), S. 226–240 [49]

Lawson u. Hanson 1974
LAWSON, Charles L. ; HANSON, Richard J.: *Solving least squares problems*. Englewood Cliffs, NJ: Prentice-Hall, 1974 [164]

Lehmann u. Casella 1998
LEHMANN, Erich L. ; CASELLA, George: *Theory of Point Estimation*. 2nd edition. New York et al.: Springer, 1998 [107], [112]

Levin u. Ben-Israel 2001
LEVIN, Yuri ; BEN-ISRAEL, Adi: A Newton method for systems of m equations in n variables. In: *Proceedings of the Third World Congress of Nonlinear Analysts, Part 3 (Catania, 2000)* Bd. 47, 2001, S. 1961–1971 [88]

Li u. Fukushima 2001
LI, Dong-Hui ; FUKUSHIMA, Masao: On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. In: *SIAM J. Optim.* 11 (2001), Nr. 4, S. 1054–1064 (electronic) [94]

Lindley u. Smith 1972
LINDLEY, D. V. ; SMITH, A. F. M.: Bayes estimates for the linear model. In: *Journal of the Royal Statistical Society, Series B* 34 (1972), S. 1–41 [105], [107]

Lindstrom u. Bates 1990
LINDSTROM, Mary J. ; BATES, Douglas M.: Nonlinear mixed effects models for repeated measures data. In: *Biometrics* 46 (1990), Nr. 3, S. 673–687 [130]

Magnus u. Neudecker 1995
MAGNUS, Jan R. ; NEUDECKER, Heinz: *Matrix differential calculus with applications in statistics and econometrics*. Chichester: John Wiley & Sons, 1995 [146]

Mäkeläinen u. a. 1981
MÄKELÄINEN, Timo ; SCHMIDT, Klaus ; STYAN, George P. H.: On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. In: *The Annals of Statistics* 9 (1981), Nr. 4, S. 758–767 [15]

Mardia u. Jupp 2000
MARDIA, Kanti V. ; JUPP, Peter E.: *Directional Statistics*. New York et al.: John Wiley & Sons, 2000 [142]

Martin 2001
MARTIN, Oliver: *Verbesserung von Finite-Element-Modellen komplexer Bauteile mit Hilfe gemessener modaler Größen*. Düsseldorf: VDI Verlag, 2001 [142]

Mascarenhas 2004
MASCARENHAS, Walter F.: The BFGS method with exact line searches fails for non-convex objective functions. In: *Mathematical Programming. A Publication of the Mathematical Programming Society* 99 (2004), Nr. 1, Ser. A, S. 49–61 [94]

Mirsky 1960
MIRSKY, L.: Symmetric gauge functions and unitarily invariant norms. In: *Q. J. Math., Oxf. II. Ser.* 11 (1960), S. 50–59 [46]

Monahan u. Genz 1997
MONAHAN, John ; GENZ, Alan: Spherical-radial integration rules for Bayesian computation. In: *Journal of the American Statistical Association* 92 (1997), Nr. 438, S. 664–674 [132]

Monahan 2001
MONAHAN, John F.: *Numerical methods of statistics*. Cambridge: Cambridge University Press, 2001 [132]

Moré u. Thuente 1994
MORÉ, Jorge J. ; THUENTE, David J.: Line search algorithms with guaranteed sufficient decrease. In: *Association for Computing Machinery. Transactions on Mathematical Software* 20 (1994), Nr. 3, S. 286–307 [62]

Müller-Gronbach u. a. 2010
MÜLLER-GRONBACH, Thomas ; NOVAK, Erich ; RITTER, Klaus: *Monte-Carlo-Methoden*. Berlin: Springer, 2010 [108], [133]

Mysovskikh 1980
MYSOVSKIKH, I. P.: The approximation of multiple integrals by using interpolatory cubature formulae. In: DEVORE, Ronald A. (Hrsg.) ; SCHERER, Karl (Hrsg.): *Quantitative approximation (Proc. Internat. Sympos., Bonn, 1979)*. New York: Academic Press, 1980, S. 217–243 [132]

Natke 1992
NATKE, Hans G.: *Einführung in die Theorie und Praxis der Zeitreihen- und Modalanalyse*. Braunschweig: Vieweg, 1992 [3], [141]

Nelson 1976
NELSON, R. B.: Simplified Calculation of Eigenvector Derivatives. In: *AIAA Journal* 14 (1976), Nr. 9, S. 1201–1205 [148]

Neumaier 2004
NEUMAIER, Arnold: Complete search in continuous global optimization and constraint satisfaction. In: *Acta Numerica* 13 (2004), S. 271–369 [98]

Nijenhuis 1974
NIJENHUIS, Albert: Strong derivatives and inverse mappings. In: *The American Mathematical Monthly* 81 (1974), S. 969–980 [12]

Nocedal u. Wright 1999
NOCEDAL, Jorge ; WRIGHT, Stephen: *Numerical Optimization*. New York et al.: Springer, 1999 [61], [71], [74], [88], [93], [95], [98]

Oettli u. Prager 1964
OETTLI, W. ; PRAGER, W.: Compatibility of approximate solution of linear equati-

ons with given error bounds for coefficients and right-hand sides. In: *Numer. Math.* 6 (1964), S. 405–409 [49]

Offinger 2000
Offinger, Robert: Empirische Bayes-Schätzung in nichtlinearen statistischen Modellen. In: *Technische Mechanik* 20 (2000), Nr. 3, S. 237–242 [130], [151]

Ortega u. Rheinboldt 1970
Ortega, J. M. ; Rheinboldt, W. C.: *Iterative solution of nonlinear equations in several variables*. New York: Academic Press, 1970 [84], [85], [86], [89]

Osborne 1985
Osborne, M. R.: *Finite algorithms in optimization and data analysis*. Chichester: John Wiley & Sons, 1985 (Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics) [61]

Paige 1979a
Paige, C. C.: Computer solution and perturbation analysis of generalized linear least squares problems. In: *Mathematics of Computation* 33 (1979), Nr. 145, S. 171–183 [52]

Paige 1979b
Paige, C. C.: Fast numerically stable computations for generalized linear least squares problems. In: *SIAM J. Numer. Anal.* 16 (1979), Nr. 1, S. 165–171 [52]

Pázman 1984
Pázman, Andrej: Nonlinear least squares - uniqueness versus ambiguity. In: *Mathematische Operationsforschung und Statistik / Series Statistics* 15 (1984), Nr. 3, S. 323–336 [20]

Pázman 1992
Pázman, Andrej: Geometry of the Nonlinear Regression with Prior. In: *Acta Math. Univ. Comenianae* LXI (1992), Nr. 2, S. 263–276 [120]

Pázman 1993
Pázman, Andrej: *Nonlinear Statistical Models*. Dordrecht et al.: Kluwer Academic Publishers, 1993 [4], [19], [24], [28], [36], [38], [120]

Pinheiro u. Bates 1980
Pinheiro, José C. ; Bates, Douglas M.: Approximations to the Log-likelihood Function in the Nonlinear Mixed Effects Model. In: *Journal of Computational and Graphical Statistics* 4 (1980), Nr. 1, S. 12–35 [130], [132]

Poole u. Neal 2000
Poole, George ; Neal, Larry: The rook's pivoting strategy. In: *Journal of Computational and Applied Mathematics* 123 (2000), Nr. 1-2, S. 353–369 [55]

Postnikov 1989
Postnikov, M.: *Smooth manifolds: Lectures in geometry, Semester III.* Moskau: Mir, 1989 [165]

Powell 1971
Powell, M. J. D.: On the convergence of the variable metric algorithm. In: *Journal of the Institute of Mathematics and its Applications* 7 (1971), S. 21–36 [94]

Powell 1972
Powell, M. J. D.: Some properties of the variable metric algorithm. In: *Numerical methods for non-linear optimization (Conf., Univ. Dundee, Dundee, 1971).* London: Academic Press, 1972, S. 1–17 [94]

Powell 1976
Powell, M. J. D.: Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In: *Nonlinear programming (Proc. Sympos., New York, 1975).* Providence, R. I.: Amer. Math. Soc., 1976, S. 53–72. SIAM–AMS Proc., Vol. IX [66], [93]

Pukelsheim 1993
Pukelsheim, Friedrich: *Optimal Design of Experiments.* New York et al.: John Wiley & Sons, 1993 [107]

Pukkila u. Rao 1988
Pukkila, Tarmo M. ; Rao, C. R.: Pattern recognition based on scale invariant discriminant functions. In: *Information Sciences* 45 (1988), Nr. 3, S. 379–389 [144]

R Development Core Team 2009
R Development Core Team: *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2009
`http://www.R-project.org` [94], [95], [98]

Ramsin u. Wedin 1977
Ramsin, H. ; Wedin, P.-Å.: A Comparison of Some Algorithms for the Nonlinear Least Squares Problem. In: *Nordisk Tidskr. Informationsbehandling (BIT)* 17 (1977), Nr. 1, S. 72–90 [88]

Rao u. Kleffe 1988
Rao, Calyampudi R. ; Kleffe, Jürgen: *Estimation of variance components and applications.* Amsterdam et al.: Elsevier Science Publishers, 1988 [112]

Rao u. Mitra 1971
Rao, Calyampudi R. ; Mitra, Sujit K.: *Generalized inverse of matrices and its applications.* New York et al.: John Wiley & Sons, 1971 [146], [148]

Rao u. Rao 1998
Rao, Calyampudi R. ; Rao, M. B.: *Matrix Algebra and Its Applications to Statistics and Econometrics.* Singapore et al.: World Scientific, 1998 [4]

Ratkowsky 1983
RATKOWSKY, David A.: *Nonlinear regression modeling: a unified practical approach*. New York/Basel: Marcel Dekker, 1983 [98]

Reid 1971
REID, J.K.: A note on the stability of Gaussian elimination. In: *Journal of the Institute of Mathematics and its Applications* 8 (1971), S. 374–375 [55]

Rencher u. Schaalje 2008
RENCHER, Alvin C. ; SCHAALJE, G. B.: *Linear models in statistics*. 2nd edition. Hoboken, NJ: Wiley-Interscience [John Wiley & Sons], 2008 [107]

Robert 2007
ROBERT, Christian P.: *Bayesian Choice*. 2nd edition. New York: Springer, 2007 [107], [108], [110], [132]

Rudin 1987
RUDIN, Walter: *Real and Complex Analysis*. New York et al.: McGraw-Hill, 1987 [103]

Rudin 1998
RUDIN, Walter: *Analysis*. München: Oldenbourg, 1998 [12]

Rump 2003a
RUMP, Siegfried M.: Structured perturbations. I. Normwise distances. In: *SIAM J. Matrix Anal. Appl.* 25 (2003), Nr. 1, S. 1–30 [49]

Rump 2003b
RUMP, Siegfried M.: Structured perturbations. II. Componentwise distances. In: *SIAM J. Matrix Anal. Appl.* 25 (2003), Nr. 1, S. 31–56 [49]

Ryan 1997
RYAN, Thomas P.: *Modern Regression Methods*. New York et al.: John Wiley & Sons, 1997 [4]

SAS Institute 2000
SAS INSTITUTE, Inc.: *SAS: Version 8*. Cary, NC: SAS Institute, Inc., 2000 [97]

Schmidt 1907
SCHMIDT, Erhard: Zur Theorie der linearen und nichtlinearen Integralgleichungen. In: *Mathematische Annalen* 63 (1907), Nr. 4, S. 433–476 [46]

Searle 1982
SEARLE, Shayle R.: *Matrix Algebra Useful for Statistics*. New York et al.: John Wiley & Sons, 1982 [4], [33]

Searle u. a. 1992
SEARLE, Shayle R. ; CASELLA, George ; MCCULLOCH, Charles E.: *Variance Components*. New York et al.: John Wiley & Sons, 1992 [112]

Seber u. Wild 1989
SEBER, G. A. F. ; WILD, C. J.: *Nonlinear regression*. New York: John Wiley & Sons Inc., 1989 [4], [36], [96], [98]

Seber 1977
SEBER, G.A.F.: *Linear Regression Analysis*. New York et al.: John Wiley & Sons, 1977 [4], [35]

Shanno 1970
SHANNO, D. F.: Conditioning of quasi-Newton methods for function minimization. In: *Mathematics of Computation* 24 (1970), S. 647–656 [89]

Shun u. McCullagh 1995
SHUN, Zhenming ; MCCULLAGH, Peter: Laplace approximation of high dimensional integrals. In: *Journal of the Royal Statistical Society B* 57 (1995), Nr. 4, S. 749–760 [132]

Skiena 1998
SKIENA, Steve: *The algorithm design manual*. Corrected 2nd printing. Berlin: Springer, 1998 [51]

van der Sluis 1974
SLUIS, A. van d.: Stability of the solutions of linear least squares problems. In: *Numer. Math.* 23 (1974/75), S. 241–254 [48]

Spivak 1965
SPIVAK, Michael: *Calculus on manifolds. A modern approach to classical theorems of advanced calculus*. W. A. Benjamin, Inc., New York-Amsterdam, 1965 [165]

Spivak 1979
SPIVAK, Michael: *A comprehensive introduction to differential geometry. Vol. I-V*. 2nd Edition. Wilmington, Del.: Publish or Perish Inc., 1979 [165]

Strassen 1969
STRASSEN, Volker: Gaussian elimination is not optimal. In: *Numer. Math.* 13 (1969), S. 354–356 [50]

Tierney u. Kadane 1986
TIERNEY, L. ; KADANE, J.B.: Accurate Approximation for Posterior Moments and Marginal Densities. In: *Journal of the American Statistical Association* 81 (1986), Nr. 393, S. 82–86 [132]

Törn u. Žilinskas 1989
TÖRN, Aimo ; ŽILINSKAS, Antanas: *Lecture Notes in Computer Science*. Bd. 350: *Global optimization*. Berlin: Springer-Verlag, 1989 [98]

Tveito u. Winther 2005
TVEITO, Aslak ; WINTHER, Ragnar: *Introduction to Partial Differential Equations*. 2nd edition. Berlin et al.: Springer, 2005 [139]

Voinov 1985
VOINOV, V. G.: Unbiased estimation of powers of the inverse of mean and related problems. In: *Sankhyā (Statistics). The Indian Journal of Statistics. Series B* 47 (1985), Nr. 3, S. 354–364 [153]

Watson 1983
WATSON, Geoffrey S.: *Statistics on spheres.* New York: John Wiley & Sons, 1983 [144]

Wedin 1973
WEDIN, Per-Åke: Perturbation theory for pseudo-inverses. In: *Nordisk Tidskr. Informationsbehandling (BIT)* 13 (1973), S. 217–232 [47]

Wedin 1974
WEDIN, Per-Åke: On the Gauss-Newton Method for the Nonlinear Least Squares Problems. Stockholm, Sweden: Institute for Applied Mathematics, 1974 (24) Working Paper [88]

Wilkinson 1965
WILKINSON, J. H.: Error analysis of transformations based on the use of matrices of the form $I - 2ww^H$. In: *Error in Digital Computation, Vol. 2 (Proc. Sympos. Math. Res. Center, U . S. Army, Univ. Wisconsin, Madison, Wis., 1965).* New York: Wiley, 1965, S. 77–101 [164]

Wilkinson 1968
WILKINSON, J. H.: A priori error analysis of algebraic processes. In: *Proc. Internat. Congr. Math. (Moscow, 1966).* Izdat. "Mir", Moscow, 1968, S. 629–640 [53]

Wilkinson 1969
WILKINSON, James H.: Global convergence of QR algorithm. (With discussion). In: *Information Processing 68 (Proc. IFIP Congress, Edinburgh, 1968), Vol. 1: Mathematics, Software.* Amsterdam: North-Holland, 1969, S. 130–133 [163]

Winterfors u. Curtis 2008
WINTERFORS, Emanuel ; CURTIS, Andrew: Numerical detection and reduction of non-uniqueness in nonlinear inverse problems. In: *Inverse Problems* 24 (2008), Nr. 2, S. 025016, 14 [13]

Witting 1985
WITTING, Hermann: *Mathematische Statistik I: Parametrische Verfahren bei festem Stichprobenumfang.* Stuttgart: Teubner, 1985 [30], [31], [32], [34], [99], [100], [107]

Witting u. Müller-Funk 1995
WITTING, Hermann ; MÜLLER-FUNK, Ulrich: *Mathematische Statistik II: Parametrische Modelle und nichtparametrische Funktionale.* Stuttgart: Teubner, 1995 [130]

Wolfe 1969
WOLFE, Philip: Convergence conditions for ascent methods. In: *SIAM Review* 11 (1969), S. 226–235 [63]

Wolfe 1971
WOLFE, Philip: Convergence conditions for ascent methods. II. Some corrections. In: *SIAM Review* 13 (1971), S. 185–188 [63]

Wolfram Research 2007
WOLFRAM RESEARCH, Inc.: *Mathematica Edition: Version 6.0*. Champaign, Illinois: Wolfram Research, Inc., 2007 [94]

Zehn u. Machina 2005
ZEHN, Manfred W. ; MACHINA, Gangadhar: Inclusion of Manufacturing Induced Uncertainties into Linear Vibration of Structures. In: *Proceedings in Applied Mathematics and Mechanics* 5 (2005), Nr. 1, S. 109–110 [151]

Zehn u. a. 1999
ZEHN, Manfred W. ; MARTIN, Oliver ; OFFINGER, Robert: Influence of Parameter Estimation Procedures on the Updating Process of Large Finite Element Models. In: FRISWELL, M.I. (Hrsg.) ; MOTTERSHEAD, J.E. (Hrsg.) ; LEES, A.W. (Hrsg.): *Identification in Engineering Systems, Proceedings of the Second International Conference held in Swansea, March 1999*. Swansea: University of Wales, 1999, S. 240–250 [117]

Zehn u. Saitov 2003
ZEHN, Manfred W. ; SAITOV, A.: How can spatially distributed uncertainties be included in FEA and in parameter estimation for model updating? In: *Shock and Vibration* 10 (2003), Nr. 1, S. 15–25 [151]

Zehn u. a. 2000
ZEHN, Manfred W. ; SAITOV, A. ; G., Schmidt: Berücksichtigung statistischer Modellunsicherheiten in den Dicken von CFK-Strukturen bei der indirekten Parameteridentifikation. In: *VDI Berichte* 1550 (2000), S. 691–705 [149]

Zhigljavsky u. Žilinskas 2008
ZHIGLJAVSKY, Anatoly ; ŽILINSKAS, Antanas: *Stochastic global optimization*. Bd. 9. New York: Springer, 2008 [98]