Technische Universität Dortmund

Fakultät Statistik

Hierarchische Bayes-Regression bei Einbettung großer Datensätze

Jonathan Rathjens

Masterarbeit

Betreuung:
Prof. Dr. Katja Ickstadt
Dipl.-Stat. Leo Geppert

2014 / 2015

Dank

Ich danke Professorin Katja Ickstadt, dass sie es mir ermöglicht hat, zu einem höchst interessanten Forschungsgebiet einen kleinen Beitrag zu leisten. Ich habe mich immer sehr gut aufgehoben und beraten gefühlt. Ihr und Dr. Uwe Ligges danke ich für das Engagement als Prüfer.

Ebenfalls für die Betreuung dieser Arbeit bedanke ich mich bei Leo Geppert. Er hat mich mit dem Forschungsprojekt bekannt gemacht und mir in allen großen und kleinen Fragen immer hilfsbereit zur Seite gestanden.

Alexander Munteanu von der Fakultät für Informatik danke ich für die geduldige Unterstützung auf seinem Gebiet. Seine Hilfe hat mir diese fachübergreifende Arbeit wesentlich erleichtert.

Ein besonderer Dank gilt Professor Roland Fried, an dessen Lehrstuhl ich seit langer Zeit als Studentische Hilfskraft tätig sein darf. Ich habe dabei sehr viel über wissenschaftliche Arbeitsweisen lernen können und zahlreiche Einblicke erhalten. Ihm und all seinen derzeitigen und ehemaligen Mitarbeitern, namentlich Dr. Anita Thieler, danke ich daher für diese wertvolle Ergänzung meiner Ausbildung.

Natürlich danke ich auch meiner gesamten Familie und allen anderen Menschen, die mich bis hierhin unterstützt haben.

Dazu gehören schließlich meine weiteren fleißigen Korrekturleser: Meiner Freundin Larisa Kaplinskaya danke ich für die Prüfung auf Verständlichkeit, meiner Mutter Renate Rathjens für das Finden jedes einzelnen Tippfehlers und meinem Schwager Jochen Behrens für die kritische Durchsicht aus Sicht eines Mathematikers.

Wetter (Ruhr), im Februar 2015

Jonathan Rathjens

Abstract

Subspace embedding is a well-known method of data reduction preserving the essential information. Its applicability in a Bayesian linear regression framework has already been proven. The posterior distribution of the coefficients estimated from the original data is approximated by that from the so-called sketch up to a small, controlled error.

As a generalization, some hierarchical regression models are analyzed in this thesis. Apart from the regression coefficients, hyperparameters are estimated, each with comparison between original and sketched data.

Simulation studies suggest a good approximation of both the regression coefficients and hyperparameters for all linear regression models with normal likelihood and various priors. For a generalized model with a logit-based link function, the approximation seems worse.

For a normal likelihood, a bounded distance between the two posterior distributions from the original and the sketched data can be derived from the non-hierarchical results, although without exact quantification. For the marginal distribution of the regression coefficients with normal priors, the previous result is reproduced exactly. If a location hyperparameter is the expected value of a regression coefficient, the same bound holds for it. A non-linear link function raises certain problems if a linear embedding is applied.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Das Ausmaß an verfügbarer Information nimmt in vielen Lebensbereichen zu. Hochdimensionale Daten werden gewonnen und sollen verarbeitet werden. Der Fortschritt der Informationstechnik erhöht zwar die Rechengeschwindigkeit und Speichermöglichkeit, was aber wiederum die Entwicklung datenintensiver Verfahren motiviert. Die technischen Ressourcen sind jedoch nicht unbegrenzt und teilweise überlastet, weshalb nach Ansätzen gesucht wird, den Umgang mit großen Datensätzen zu erleichtern.

Ein Arbeitsgebiet stellen dabei Regressionsverfahren dar, welche zu den bedeutendsten Anwendungen in der Statistik gehören und bei denen sehr viele Beobachtungen auftreten können.

Hier hat sich die Unterraum-Einbettung als ein möglicher Weg erwiesen: Aus einem großen Datensatz wird ein kleinerer berechnet, der mit einer sehr viel geringeren Anzahl Beobachtungen dennoch die wesentlichen Informationen enthält. Die Schätzung der Regressionsparameter gelingt bis auf einen kleinen, kontrollierbaren Fehler relativ zum großen Datensatz.

Zusätzlich können Bayessche Verfahren betrachtet werden, bei denen ggf. Vorinformationen einbezogen werden können und als Ergebnis eine Verteilung anstelle eines einzelnen Schätzwerts steht. Auch hier ist die Anwendbarkeit der Einbettungsmethodik für die lineare Regression bereits gezeigt.

Unklar ist dies jedoch im Falle hierarchischer Bayesscher Modelle, in denen die Verteilung der Regressionsparameter durch zusätzliche Hyperparameter beschrieben wird. Ob die Verteilungen beider Parameterklassen auch aus dem reduzierten

Datensatz, der sogenannten Skizze, korrekt geschätzt werden, soll in dieser Arbeit untersucht werden.

Dabei wird auf die Ergebnisse von Geppert *et al.* (2014) für den nicht-hierarchischen Fall aufgebaut. In einigen Situationen lässt sich die Approximierbarkeit der Daten durch die Skizze auf ein hierarchisches Modell übertragen. Im Übrigen führt eine Simulation verschiedener Beispiele mit drei Einbettungsmethoden zu vielversprechenden Ergebnissen. Als ein Problem erweist sich der Übergang zu generalisierten linearen Modellen.

Im folgenden Kapitel 2 werden die verwendeten Methoden der Bayes-Statistik, insbesondere der Parameterschätzung in hierarchischen Regressionsmodellen eingeführt. Kapitel 3 beschreibt die Einbettungsmethodik im Allgemeinen und die drei verwendeten Verfahren im Besonderen und gibt die vorhandenen Erkenntnisse für den nicht-hierarchischen Fall wieder. Als Ergänzung dazu enthält das Kapitel 4 Hinweise zum Forschungsstand in angrenzenden Gebieten.

Den empirischen Hauptteil der Arbeit stellt die Simulationsstudie in Kapitel 5 dar. Sie besteht zunächst aus fünf Beispielen, die später durch zwei weitere ergänzt werden. Unter Beachtung dieser Ergebnisse wird im Kapitel 6 versucht, allgemeinere Aussagen zur Anwendbarkeit der Einbettungsmethodik in hierarchischen Modellen herzuleiten. Dies betrifft die Schätzung der Regressions- und Hyperparameter sowie der gemeinsamen Verteilung aller. Generalisierte lineare Modelle werden gesondert besprochen. Zu ihnen werden in Kapitel 7 zwei zusätzliche Beispiele für den nicht-hierarchischen Fall ohne Bayes-Methodik untersucht. Im letzten Kapitel 8 werden alle Ergebnisse dieser Arbeit zusammengefasst und mögliche künftige Forschungsansätze skizziert.

Kapitel 2

Grundlagen und Bezeichnungen

Die in dieser Arbeit untersuchte Modellklasse erfordert eine kurze Einführung in die verwendeten Methoden der Bayes-Statistik. Dies betrifft zunächst die Parameterschätzung im Allgemeinen. Ein wichtiges Hilfsmittel dabei sind die im zweiten Abschnitt besprochenen numerischen MCMC-Verfahren. Zuletzt wird die Klasse der hierarchischen Regressionsmodelle im Zusammenhang mit der Bayesschen linearen Regression unter Normalverteilungsannahme besprochen.

Dabei werden die in dieser Arbeit verwendeten Notationen eingeführt und ggf. erläutert. Hier und im folgenden Kapitel 3 neu eingeführte Begriffe sind **fett** dargestellt. Regelmäßig verwendete Symbole sind im Anhang A.1 zusammengefasst.

2.1 Bayes-Schätzung

In der Bayes-Statistik (Thomas Bayes, † 1761) werden vorhandene Parameter als Zufallsvariablen aufgefasst. Ein Schätzproblem bedeutet die Bestimmung ihrer Verteilung. Hierzu werden einerseits die beobachteten Daten verwendet, andererseits gewisse Vorinformationen (***prior*** oder ***a priori***-Wissen genannt). Aufgrund der Beobachtungen wird der *prior* korrigiert bzw. ergänzt, es ergibt sich die ***a-posteriori***-Information (***posterior***).

Die untersuchten Verteilungen werden durch Dichten p beschrieben, welche symbolisch sowohl den stetigen als auch den diskreten Fall umfassen und ggf. mehrdimensional sind. Weiter seien $p(\cdot,\cdot)$ eine gemeinsame und $p(\cdot|\cdot)$ eine auf das zweite Argument bedingte Dichte.

Es seien

$$\xi = (\xi_1, \ldots, \xi_l)^\top \in \mathbb{R}^l \tag{2.1}$$

ein Vektor unbekannter Parameter und

$$y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^{n(\times m)} \tag{2.2}$$

eine, ggf. m-dimensionale, Stichprobe. Beide Symbole sind im Allgemeinen als Zufallsvariablen zu lesen, sofern sie allein stehen. Andererseits dienen sie auch als Argument einer Dichte p, welche die entsprechende Verteilung beschreibt. Als Bedingung in $p(\cdot|\cdot)$ sind sie wie realisierte Zufallsvariablen zu lesen.

Nun heißt $p(y|\xi)$ ***likelihood*** der Daten. Der *prior* ist $p(\xi)$. Gesucht ist die *a-posteriori*-Dichte $p(\xi|y)$. Für sie gilt mit Hilfe des Satzes von Bayes:

$$p(\xi|y) = \frac{p(y,\xi)}{p(y)} = \frac{p(y|\xi) \cdot p(\xi)}{p(y)}. \tag{2.3}$$

Wegen

$$p(y) = \int_{\mathbb{R}^l} p(y|\xi)p(\xi)d\xi \tag{2.4}$$

(Gelman *et al.*, 2014, Kap. 1.3) gilt bis auf Normalisierung die Proportionalität

$$p(\xi|y) \propto p(y|\xi) \cdot p(\xi). \tag{2.5}$$

Integrale wie in Gleichung (2.4) werden symbolisch auch für den diskreten Fall verwendet. Tatsächlich sind dort Summen bzw. Reihen zu betrachten.

Ist nur ein Teil der Parameter von Interesse, so gilt mit

$$\xi = \left(\xi_{(1)}^\top, \xi_{(2)}^\top\right)^\top, \quad \xi_{(1)} \in \mathbb{R}^{l_1}, \quad \xi_{(2)} \in \mathbb{R}^{l_2}, \quad l = l_1 + l_2 \tag{2.6}$$

a posteriori für die Rand-, gemeinsame und bedingte Dichte:

$$p(\xi_{(1)}|y) = \int_{\mathbb{R}^{l_2}} p(\xi|y)d\xi_{(2)} = \int_{\mathbb{R}^{l_2}} p(\xi_{(1)}|\xi_{(2)}, y)p(\xi_{(2)}|y)d\xi_{(2)} \tag{2.7}$$

(Gelman *et al.*, 2014, Kap. 3.1). Die Verteilung zu $p(\xi_{(1)}|\xi_{(2)}, y)$ heißt ***full conditional posterior*** von $\xi_{(1)}$.

Oft ist ein *prior* nicht gegeben und deshalb geeignet zu wählen, was erheblichen Einfluss auf das Ergebnis haben kann. Ein **nicht-informativer *prior*** wird so gewählt, dass die *posterior* von ihm wenig beeinflusst wird. Dies kann etwa eine Verteilung mit sehr großer, gegen unendlich gehender Varianz sein (**unzulässiger *prior***, vgl. z. B. Gelman *et al.*, 2014, Kap. 5.7), welche allen möglichen Parametern *a priori* ähnliche Wahrscheinlichkeiten gibt. Ein **konjugierter *prior*** gehört zu einer bestimmten Klasse von Verteilungen, zu der auch die *posterior* gehört und die zu

gewissen Verteilungsklassen der *likelihood* eindeutig existiert (vgl. z. B. Rinne, 2008, Kap. C4.1).

Für ein gegebenes Bayes-Modell aus *likelihood* und *prior* ist die tatsächliche Berechnung von Integralen wie in Gleichung (2.7) das wesentliche Problem. Bei bestimmten Modellen ist eine analytische Lösung der *posterior* berechenbar. Anderenfalls müssen numerische Verfahren angewendet werden: Neben den in Abschnitt 2.2 vorgestellten MCMC-Verfahren kommen je nach Modell auch ABC oder INLA in Frage (vgl. Abschnitt 4.2).

2.2 *Markov-Chain-Monte-Carlo*-Verfahren

In vielen Fällen kann die *posterior* $p(\xi|y)$ nicht analytisch berechnet werden. Sie wird dann numerisch bestimmt. In einem ***Markov-Chain-Monte-Carlo*-Verfahren** (**MCMC**) werden dazu iterativ Zufallszahlen

$$\xi^{(t)} \in \mathbb{R}^l, \quad t = 1, \ldots, T \tag{2.8}$$

aus Verteilungen mit Dichten $\tilde{p}^{(t)}(\xi)$ gezogen. Jeder Wert hängt nur vom zuvor gezogenen ab (**Markov-Eigenschaft**). Das Verfahren ist so konstruiert, dass die *posterior* als **stationäre Verteilung** einer **Markov-Kette** auftritt, d. h. die Verteilungen der Zufallszahlen (2.8) mit den Dichten $\tilde{p}^{(t)}$ konvergieren gegen sie (Gelman *et al.*, 2014, Kap. 11). Nach hinreichend langer Zeit kann die Stichprobe der weiteren $\xi^{(t)}$ die *posterior* so wiedergeben, dass etwa Momente geschätzt werden können oder auf Verteilungsfamilien getestet werden kann.

Es existieren verschiedene Umsetzungen dieses Prinzips. Die in der Software „OpenBUGS“ (OpenBUGS Project Management Group, 2012, *User Manual*, *Introduction*) implementierten Algorithmen gehören zur **Metropolis-Hastings**-Klasse: Hier wird zunächst ein Startwert $\xi^{(0)}$ aus einer Startverteilung gezogen, von ihm ausgehend die weiteren $\xi^{(t)}$. Deren jeweilige Verteilung mit Dichte $\tilde{p}^{(t)}$ heißt ***proposal distribution***. Ihre genaue Wahl hängt vom konkreten Algorithmus ab, sie muss den Träger der *posterior* überdecken und soll ungefähr deren Form besitzen (Congdon, 2006, Kap. 1.3-4).

Sind die Werte bis $\xi^{(t)}$ gegeben, wird der nächste **Kandidat** ξ^* aus $\tilde{p}^{(t+1)}$ gezogen. Für ihn wird die **Akzeptanzwahrscheinlichkeit**

$$\alpha(\xi^*|\xi^{(t)}) := \min\left\{1, \frac{p(\xi^*|y) \cdot \tilde{p}^{(t+1)}(\xi^{(t)}|\xi^*)}{p(\xi^{(t)}|y) \cdot \tilde{p}^{(t+1)}(\xi^*|\xi^{(t)})}\right\} \tag{2.9}$$

bestimmt. Die $\tilde{p}^{(t+1)}(\cdot|\cdot)$ beschreiben dabei Übergangswahrscheinlichkeiten der Markov-Kette. Die *posterior* kann an den gegebenen Stellen ausgewertet werden, da sich die Normalisierungskonstanten $p(y)$ aus Gleichung (2.3) im Quotienten gegenseitig aufheben. Damit wird

$$\xi^{(t+1)} := \begin{cases} \xi^* & \text{mit Wahrscheinlichkeit } \alpha \\ \xi^{(t)} & \text{mit Wahrscheinlichkeit } 1-\alpha \end{cases} \tag{2.10}$$

gesetzt. Dies bedeutet, dass der Markov-Kern

$$\kappa(\xi^{(t)}|\xi) = \begin{cases} \alpha(\xi|\xi^{(t)}) \cdot \tilde{p}^{(t+1)}(\xi|\xi^{(t)}) & : \xi \neq \xi^{(t)} \\ 1 - \int_{\mathbb{R}^l} \alpha(\xi|\xi^{(t)}) \cdot \tilde{p}^{(t+1)}(\xi|\xi^{(t)}) d\xi & : \xi = \xi^{(t)} \end{cases} \tag{2.11}$$

verwendet wird (Congdon, 2006, Kap. 1.4). Näheres zu diesem Verfahren und seinen Konvergenzeigenschaften kann etwa bei Gelman *et al.* (2014, Kap. 11.2-3) und Congdon (2006, Kap. 1.4–5) sowie den dort gegebenen Referenzen nachvollzogen werden.

Ein Spezialfall des allgemeinen Metropolis-Hastings-Algorithmus' ist der **Metropolis-Algorithmus**, bei dem die *proposal distribution* symmetrisch ist und sich die $\tilde{p}^{(t+1)}$ in der Akzeptanzwahrscheinlichkeit (2.9) daher gegenseitig aufheben (Gelman *et al.*, 2014, Kap. 11.2).

„OpenBUGS" verwendet weiterhin den **Gibbs-*sampler***: Hier wird der neue Kandidat ξ^* in mehreren Schritten komponentenweise gewählt und die *proposal distribution* auf die übrigen Komponenten bedingt. Dies führt zu einer Akzeptanzwahrscheinlichkeit $\alpha \equiv 1$ (Congdon, 2006, Kap. 1.4.1).

Beim ***slice sampling*** schließlich wird ein Kandidat aus einer Verteilung mit einer zusätzlichen Dimension gezogen, für die gilt:

$$\tilde{p}^{(t+1)}((\xi^\top, \xi_{l+1})^\top|y) \begin{cases} \propto 1 & : 0 < \xi_{l+1} < \tilde{p}^{(t+1)}(\xi|y) \\ = 0 & : \text{sonst} \end{cases} \quad \forall \xi \tag{2.12}$$

Auf diese Weise kann etwa das Verfahren des Gibbs-*samplers* auf ein eindimensionales ξ übertragen werden (Gelman *et al.*, 2014, Kap. 12.3).

2.3 Hierarchische Bayes-Modelle und Regression

Das **hierarchische Bayes-Modell** beruht auf einer speziellen Einteilung der Parameter entsprechend (2.6) mit

$$\xi = \left(\beta^\top, \theta^\top\right)^\top, \quad \beta \in \mathbb{R}^k, \theta \in \mathbb{R}^{l-k} \tag{2.13}$$

so dass die Verteilung der Daten y von den β direkt beeinflusst wird, während die β wiederum durch die θ beschrieben werden:

$$p(y, \beta, \theta) = p(y|\beta, \theta)p(\beta|\theta)p(\theta). \tag{2.14}$$

Die θ, als „Parametrisierung der Parameter“, heißen **Hyperparameter**. Weitere Hierarchiestufen kommen in Frage. Entsprechend der Gleichungen (2.3) bis (2.7) gilt:

$$p(\theta|y) \propto p(\theta)p(y|\theta) = p(\theta) \int_{\mathbb{R}^k} p(y|\beta, \theta)p(\beta|\theta)d\beta \tag{2.15}$$

(vgl. Congdon, 2010, Kap. 1.4). Die Verteilung zu $p(\beta|\theta)$ wird auch **Populationsverteilung** genannt.

Da die Daten nur mittelbar von den Hyperparametern beeinflusst werden, gilt weiter:

$$p(\beta, \theta|y) \propto p(\beta, \theta)p(y|\beta, \theta) = p(\theta)p(\beta|\theta)p(y|\beta), \tag{2.16}$$

so dass nur für die Hyperparameter ein *prior* gewählt werden muss (Gelman *et al.*, 2014, Kap. 5.2). Im Falle eines konjugierten *priors* kann eine auf die Hyperparameter bedingte *posterior* $p(\beta|\theta, y)$ bestimmt werden. Aus dieser wiederum kann unter anderem bei Normalverteilung die *posterior*

$$p(\theta|y) = \frac{p(\beta, \theta|y)}{p(\beta|\theta, y)} \tag{2.17}$$

der Hyperparameter berechnet werden (ebd., Kap. 5.3).

Im Folgenden wird ein lineares Regressionsmodell mit dem Parametervektor

$$\beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k \tag{2.18}$$

betrachtet. Mit gegebener deterministischer Designmatrix

$$X = (X_{1,\cdot}^\top, \ldots, X_{n,\cdot}^\top)^\top \in \mathbb{R}^{n \times k} \tag{2.19}$$

und Annahme der Normalverteilung und Homoskedastizität ohne Autokorrelation ergibt sich die *likelihood*

$$p(y_i|\beta) = N\left(X_{i,\cdot}\beta, s^2\right), \quad i = 1, \ldots, n, \tag{2.20}$$

oder kurz

$$p(y|\beta) = N\left(X\beta, \mathrm{Id}_n s^2\right) \tag{2.21}$$

(vgl. z. B. Rinne, 2008, Kap. D1.1). Die Abhängigkeit der Verteilungen $p(\cdot)$ bzw. $p(\cdot|\cdot)$ von X wird impliziert, aber in dieser Arbeit nicht durchgehend genannt.

Wird (2.21) als Bayes-Modell interpretiert, so ist die *posterior* der Regressionsparameter β als Schätzwert gesucht. Sie entspricht insofern dem klassischen Resultat, als gilt:

$$p(\beta|y, s^2) = N\left(\hat{\beta}, \hat{V}s^2\right) \tag{2.22}$$

mit den Schätzern

$$\hat{\beta} = (X^\top X)^{-1} X^\top y, \quad \hat{V} = (X^\top X)^{-1} \tag{2.23}$$

(Gelman *et al.*, 2014, Kap. 14.2).

Die unbekannte Verteilung von β sei im Folgenden durch einen oder mehrere Hyperparameter θ beschrieben. Der Varianzparameter s^2 wird entweder als fest und bekannt vorausgesetzt oder ebenfalls durch seine *posterior* geschätzt.

Es existieren zahlreiche solche Modelle. Möglichkeiten werden etwa bei Gelman *et al.* (2014, Kap. 15) oder Congdon (2010) vorgestellt. Besonders hervorzuheben ist eine Formulierung, bei der β und θ ebenfalls normalverteilt sind (vgl. Gelman *et al.*, 2014, Kap. 15.3; Congdon, 2010, Kap. 3.3, oder die Modelle 1a, 1a' und 1b in Kapitel 5).

Gelman *et al.* (2014, Kap. 5.4) zeigen die vollständige Berechnung eines ähnlichen Normalverteilungsmodells ohne den Kontext der Regression. Hier wird auf Schwierigkeiten hingewiesen, wenn ein nicht-informativer *prior* für die Varianzparameter gewählt wird. Dies kann zu ungeeigneten Schätzungen führen, wenn die Varianz gering oder die Anzahl der Parameter (hier k) klein ist (ebd., Kap. 5.7). Bei der Regression ist die Varianz von *intercept* und *slope*-Parametern ggf. gesondert zu modellieren (ebd., Kap. 15.4).

Kapitel 3

Einbettungen

Der wesentliche Gegenstand dieser Arbeit sind Einsatzmöglichkeit und Leistungsfähigkeit eines Einbettungsverfahrens zur Reduktion großer Datensätze. Dieses Verfahren wird im Folgenden vorgestellt. Nach einer allgemeinen Einführung werden einige Ergebnisse von Geppert *et al.* (2014) über den Nutzen bei der (Bayesschen) Regressionsanalyse vorgestellt. Zuletzt werden die hier verwendeten konkreten Realisierungen des Verfahrens besprochen.

Es ist zu beachten, dass k bei der allgemeinen Besprechung der Einbettung die Dimension des gesamten Datensatzes bezeichnet, im Kontext der Regression jedoch nur die Anzahl der Einflussvariablen, wie in der übrigen Arbeit. Der Unterschied von 1 entsteht daraus, dass dort der Vektor der Zielvariablen hinzukommt.

3.1 Motivation und Definition

Es sei ein Datensatz $U \in \mathbb{R}^{n \times k}$ mit $n \gg k$ gegeben. Die sehr große Zahl n der Beobachtungen verursacht neben einem hohen Speicherbedarf auch lange Rechenzeiten bei Analysen. Daher wird eine Möglichkeit gesucht, anstelle des großen einen kleineren Datensatz zu verwenden, der zu möglichst ähnlichen Ergebnissen führt, d. h. sie approximiert (vgl. Frieze *et al.*, 2004).

Da hier eine (Bayessche) Regressionsanalyse betrachtet wird, bestehen diese Ergebnisse aus Schätzwerten der Parameter bzw. ihrer *a posteriori*-Verteilung. Der Datensatz enthält die Designmatrix bzw. die Werte der Einflussvariablen und einen Vektor mit Werten der Zielvariable. Seine für die Regression wesentliche Struktur muss bei Verkleinerung erhalten bleiben – insbesondere die Zahl der Variablen.

Die Verkleinerung geschieht durch Projektion in einen Raum $\mathbb{R}^{n' \times k}$ mit wesentlich kleinerer Dimension $n' \ll n$. Derjenige Unterraum von $\mathbb{R}^{n \times k}$, welcher den großen Datensatz enthält, wird in $\mathbb{R}^{n' \times k}$ so **eingebettet** (vgl. Sarlós, 2006), dass er dort den kleinen Datensatz näherungsweise enthält.

Die Spalten von U werden als zueinander orthonormal angenommen. Bei Datensätzen, für die dies nicht gilt, wird der Spaltenraum aus einer Singulärwertzerlegung bestimmt; die Einbettung wirkt dann auch bezüglich der ursprünglichen Daten (vgl. Sarlós, 2006).

Sei nun $\epsilon \in \left(0, \frac{1}{2}\right]$ ein vorgegebener Parameter, der die Güte der Approximation beschreibt. Dann heißt die Abbildung

$$\Pi : \mathbb{R}^n \to \mathbb{R}^{n'} \tag{3.1}$$

ϵ-Unterraum-Einbettung von U, wenn gilt:

$$(1-\epsilon)\|Ux\|_2^2 \leq \|\Pi Ux\|_2^2 \leq (1+\epsilon)\|Ux\|_2^2 \quad \forall x \in \mathbb{R}^k. \tag{3.2}$$

Sie wird hier kurz (ϵ-)Einbettung genannt, der durch sie verkleinerte Datensatz ΠU heißt **Skizze** (Geppert *et al.*, 2014, Def. 2; vgl. Sarlós, 2006, Def. 1).

Je kleiner demnach ϵ gewählt wird, desto ähnlicher sind sich U und ΠU in ihrer Eigenschaft als lineare Abbildungen im $\mathbb{R}^k$. Auf solchen Abbildungen (Matrix-Multiplikationen) beruht auch die Schätzung der Regressionsparameter – vgl. etwa die Gleichungen (2.23). Ein größer gewähltes ϵ bewirkt dagegen eine kleinere und damit leichter zu verarbeitende Skizze um den Preis größerer Ungenauigkeit, wie im Abschnitt 3.2 genauer gezeigt wird.

3.2 Eigenschaften im nicht-hierarchischen Modell

Im Falle eines Regressionsmodells hat der Datensatz in der Regel keine orthonormalen Spalten. Geppert *et al.* (2014) betrachten deshalb eine ϵ-Einbettung des Spaltenraums von (y, X) mit $\epsilon \in \left(0, \frac{1}{3}\right]$. Damit untersuchen sie die Schätzwerte der Regressionsparameter β.

Dabei wird festgestellt, dass gewisse Unterschiede zwischen dem Schätzwert

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \|X\beta - y\|_2^2, \tag{3.3}$$

welcher aus dem großen Datensatz berechnet wird, und dem Schätzwert

$$\hat{\beta}' = \underset{\beta' \in \mathbb{R}^k}{\operatorname{argmin}} \|\Pi X\beta' - \Pi y\|_2^2 \tag{3.4}$$

aus der Skizze durch ϵ beschränkt sind:

Zunächst gilt für die Residuen nach Geppert *et al.* (2014, Lem. 5):

$$\left\|X\hat{\beta}' - y\right\|_2 \leq \sqrt{1+3\epsilon}\left\|X\hat{\beta} - y\right\|_2 \tag{3.5}$$

und nach dem Beweis an dieser Stelle sogar

$$\left\|X\hat{\beta}' - y\right\|_2 \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}}\left\|X\hat{\beta} - y\right\|_2, \tag{3.6}$$

was für $\epsilon \in \left(0, \frac{1}{3}\right)$ eine stärkere Schranke bedeutet.

Für den Abstand zwischen den Schätzwerten bezeichne $d_{\min}$ den kleinsten echt positiven Singulärwert von X. Dies ist die Wurzel aus dem entsprechenden Eigenwert von $X^\top X$. Alle Singulärwerte sind echt positiv, wenn X vollen Spaltenrang hat (Bronstein *et al.*, 2005, Kap. 4.5.3). Es gilt nach Geppert *et al.* (2014, Lem. 6):

$$\left\|\hat{\beta} - \hat{\beta}'\right\|_2 \leq d_{\min}^{-1}\sqrt{3\epsilon}\left\|X\hat{\beta} - y\right\|_2. \tag{3.7}$$

Nach dem Beweis an dieser Stelle gilt wegen Ungleichung (3.6) mit

$$\begin{aligned}\left\|X(\hat{\beta} - \hat{\beta}')\right\|_2^2 &= \left\|X\hat{\beta}' - y\right\|_2^2 - \left\|X\hat{\beta} - y\right\|_2^2 \\ &\leq \left(\frac{1+\epsilon}{1-\epsilon} - 1\right)\left\|X\hat{\beta} - y\right\|_2^2 \\ &= \frac{2\epsilon}{1-\epsilon}\left\|X\hat{\beta} - y\right\|_2^2\end{aligned} \tag{3.8}$$

sogar

$$\left\|\hat{\beta} - \hat{\beta}'\right\|_2 \leq d_{\min}^{-1}\sqrt{\frac{2\epsilon}{1-\epsilon}}\left\|X\hat{\beta} - y\right\|_2. \tag{3.9}$$

In der Bayes-Regression sind weiter die *a-posteriori*-Verteilungen von β und β' von Interesse. Als Abstandsmaß dieser zwei Wahrscheinlichkeitsmaße verwenden Geppert *et al.* (2014, Def. 3) die **Wasserstein-Distanz** bezüglich der euklidischen Norm:

$$W(\mathbb{P}_1, \mathbb{P}_2) := \inf\left\{\sqrt{\mathbb{E}_{(Z_1,Z_2)}\left(\|Z_1 - Z_2\|_2^2\right)} : Z_1 \sim \mathbb{P}_1, Z_2 \sim \mathbb{P}_2\right\}. \tag{3.10}$$

Seien nun $p(\beta) = p(\beta') = N(m, V)$ ein *prior*, S, $\tilde{X}$ und $\tilde{y}$ definiert durch

$$V = (S^\top S)^{-1}, \quad \tilde{X} = \begin{pmatrix} X \\ S \end{pmatrix}, \quad \tilde{y} = \begin{pmatrix} y \\ Sm \end{pmatrix} \tag{3.11}$$

sowie

$$\hat{\tilde{\beta}} = \underset{\tilde{\beta}\in\mathbb{R}^k}{\operatorname{argmin}}\left\|\tilde{X}\tilde{\beta} - \tilde{y}\right\|_2^2 \tag{3.12}$$

der entsprechende Schätzwert und $\tilde{d}_{\min}$ der kleinste echt positive Singulärwert von $\tilde{X}$. Für eine ϵ-Einbettung Π des Spaltenraums von X mit $\epsilon \in \left(0, \frac{1}{3}\right]$ gilt dann nach Geppert *et al.* (2014, The. 11):

$$W\left(p(\beta|y,X), p(\beta'|\Pi y, \Pi X)\right) \leq \sqrt{\tilde{d}_{\min}^{-2} \cdot 3\epsilon \cdot \left\|\tilde{X}\hat{\tilde{\beta}} - \tilde{y}\right\|_2^2 + 9\epsilon^2 \cdot \operatorname{tr}\left(\left(\tilde{X}^\top \tilde{X}\right)^{-1}\right)} \tag{3.13}$$

bzw. mit der Schranke aus (3.9) sogar

$$W\left(p(\beta|y,X), p(\beta'|\Pi y, \Pi X)\right) \leq \sqrt{\tilde{d}_{\min}^{-2} \cdot \frac{2\epsilon}{1-\epsilon} \cdot \left\|\tilde{X}\hat{\tilde{\beta}} - \tilde{y}\right\|_2^2 + 9\epsilon^2 \cdot \operatorname{tr}\left(\left(\tilde{X}^\top \tilde{X}\right)^{-1}\right)}. \tag{3.14}$$

Die Aussagen (3.5) bis (3.9) sowie (3.13) und (3.14) gelten für beliebige Realisierungen von (y, X). Aufgrund der Realisierung werden die Regressionsschätzer und Residuen hier nicht mehr als Zufallsvariablen betrachtet und die Aussagen sind nicht nur $\mathbb{P}$-fast sicher, sondern sicher.

3.3 Umsetzungen

Es existieren verschiedene Möglichkeiten, eine Abbildung mit der Einbettungseigenschaft (3.2) zu realisieren (vgl. Geppert *et al.*, 2014, Kap. 3, und die dort gegebenen Referenzen). Geppert *et al.* (2014) verwenden drei verschiedene Verfahren. Es wird jeweils eine zufallsabhängige Matrix $\Pi \in \mathbb{R}^{n' \times n}$ definiert, welche die Zeilen von U so transformiert, dass sich ihre Anzahl verringert:

Die **Rademacher-Matrix-Methode** (**BCH**), welche durch BCH-Code umgesetzt wird, verwendet eine umskalierte **Rademacher-Matrix**

$$\Pi = \frac{1}{\sqrt{n'}}(r_{ij})_{i=1,\ldots,n';\ j=1,\ldots,n} \tag{3.15}$$

aus u. i. v. Zufallsvariablen mit

$$\mathbb{P}(r_{ij} = -1) = \mathbb{P}(r_{ij} = 1) = \frac{1}{2}, \quad i = 1, \ldots, n';\ j = 1, \ldots, n \tag{3.16}$$

(Geppert *et al.*, 2014, Kap. 3; vgl. Clarkson und Woodruff, 2009). Die anderen beiden Methoden bauen auf ihr auf (Ailon und Liberty, 2009, und Clarkson und Woodruff, 2013).

Die ***Subsampled-Randomized-Hadamard-Transform*-Methode** (**SRHT**) verwendet

$$\Pi = \frac{1}{\sqrt{n'}} R H_n D. \tag{3.17}$$

Dabei enthält $R \in \{0,1\}^{n' \times n}$ in jeder Zeile genau eine 1, deren Positionen u. i. diskret gleichverteilt auf $\{1,\ldots,n\}$ sind, und im Übrigen 0. $H_n \in \{-1,1\}^{n \times n}$ ist eine **Hadamard-Matrix**:

$$H_1 := (1)$$
$$H_{2^m} := \begin{pmatrix} H_{2^{m-1}} & H_{2^{m-1}} \\ H_{2^{m-1}} & -H_{2^{m-1}} \end{pmatrix}, \quad m \in \mathbb{N} \tag{3.18}$$
$$H_n H_n^\top = n\mathrm{Id}_n$$

(Hedayat und Wallis, 1978, Kap. 3), weshalb n hier eine Zweierpotenz ist und ggf. intern angepasst wird. D ist eine Diagonalmatrix

$$D = \mathrm{diag}(a_1,\ldots,a_n) \in \{-1,0,1\}^{n \times n} \tag{3.19}$$

aus u. i. v. Zufallsvariablen mit

$$\mathbb{P}(a_i = -1) = \mathbb{P}(a_i = 1) = \frac{1}{2}, \quad i = 1,\ldots,n \tag{3.20}$$

(Geppert *et al.*, 2014, Kap. 3; vgl. Boutsidis und Gittens, 2013, Def. 1.2).

Die **Clarkson-Woodruff-Methode** (**CW**) verwendet

$$\Pi = BD \tag{3.21}$$

mit D wie zuvor. Für $B \in \{0,1\}^{n' \times n}$ wird eine zufällige Abbildung

$$h : \{1,\ldots,n\} \to \{1,\ldots,n'\}, \quad j \mapsto h(j), \tag{3.22}$$

betrachtet, wobei $h(j)$ für alle j auf {1, . . . , n'} diskret gleichverteilt ist. Die Einträge von B sind dann

$$b_{ij} = 1 \iff i = h(j) \tag{3.23}$$

(Geppert *et al.*, 2014, Kap. 3; vgl. Clarkson und Woodruff, 2013, und Nelson und Nguyen, 2013).

In allen drei Fällen hängt n' nicht wesentlich – und nach empirischen Ergebnissen nie – von n ab, wohl aber vom zu wählenden ϵ. Hinzu kommen k und ein Parameter δ: Die drei möglichen Π sind aufgrund der Randomisierungen nicht sicher eine ϵ-Einbettung, sondern nur mit einer Wahrscheinlichkeit $1 - \delta$ (Geppert *et al.*, 2014, Kap. 3; vgl. Frieze *et al.*, 2004, und Sarlós, 2006).

Ein wesentlicher Vorteil ist dagegen die Unabhängigkeit der Einträge von Π bei ihrer Berechnung, sowohl voneinander als auch von U, sofern k bekannt ist. Auf diese Weise hängt die Bearbeitung einer Beobachtung aus dem großen Datensatz nicht von den vorangegangenen oder künftigen ab, vielmehr kann der Datensatz eintragsweise

eingebettet werden, was auch den Speicherbedarf verringert. Weiterhin ermöglicht die Linearität von Π eine Aufteilung der Daten zur räumlich oder zeitlich getrennten Berechnung (vgl. Geppert *et al.*, 2014, Kap. 3).

Die erreichbaren Dimensionen n' der Skizzen und die benötigten Rechenzeiten für die Einbettungen werden in den oben jeweils gegebenen Referenzen gezeigt und sind in Tabelle 3.1 zusammengefasst. Bei Verwendung verschiedener Methoden und gleichen ϵ und δ bedeutet eine kleinere Skizze stets eine längere Zeit für ihre Berechnung. Die anschließende Regressionsanalyse selbst läuft bei kleinerem n' natürlich schneller.

Tabelle 3.1: Größenordnungen der Dimension n' der Skizze und der Rechenzeit T für die drei Einbettungsmethoden.

	BCH		SRHT		CW
$n' \in$	$O\left(\frac{k+\ln(1/\delta)}{\epsilon^2}\right)$	$<$	$O\left(\frac{\left(\sqrt{k}+\sqrt{\ln(n)}\right)^2 \ln(k/\delta)}{\epsilon^2}\right)$	$<$	$O\left(\frac{k^2}{\epsilon^2\delta}\right)$
$T \in$	$\Theta(nkn')$	$>$	$O(nk\ln(n'))$	$>$	$O(nk)$

Die drei von Geppert *et al.* (2014) verwendeten Methoden sind für die Software „R“ (R Core Team, 2014) im Paket „ls2mat“ von Quedenfeld *et al.* (2014) realisiert. Hierbei wird der Wahrscheinlichkeitsparameter δ in Abhängigkeit von ϵ so gewählt, dass die Einbettungseigenschaft mit äußerst hoher Wahrscheinlichkeit gilt. Das in der Sprache „C++“ realisierte Paket greift auf Funktionen aus „LAPACK++“ (Stimming *et al.*, 2010) zurück.

Kapitel 4

Verwandte Forschungsarbeiten

In vielen Bereichen der Wissenschaft und ihrer Anwendungen treten große Datensätze auf, die sich als Matrix wiedergeben lassen. Die Größe kann eine Reduktion erforderlich machen, wobei die vorhandene Information möglichst erhalten bleiben soll. Die in dieser Arbeit angewandte Methode der Unterraum-Einbettung stellt eine mögliche Lösung dieses Problems dar.

Dieses Kapitel gibt einen kurzen Überblick über einzelne Forschungsarbeiten, welche sich mit dieser Thematik auseinandersetzen. Zunächst wird die Anwendung von Einbettungsmethoden in gewissen Situationen besprochen. Im Anschluss werden Quellen zu möglichen Alternativen bei der Datenreduktion und deren Anwendungen sowie alternative Rechenverfahren genannt.

4.1 Anwendung von Einbettungsmethoden

Woodruff (2014) gibt einen umfassenden Überblick über die Anwendbarkeit von Einbettungsmethoden in numerischen Verfahren. Dies betrifft zunächst Regressionsprobleme mit verschiedenen Zielfunktionen. Weiter wird die Approximation von Matrizen (*low rank approximation*) und Graphen besprochen. Besonderes Augenmerk liegt auf der Übertragung der ϵ-Approximierbarkeit bei Verwendung anderer Normen und diverser Matrix-Operationen.

Muthukrishnan (2005) setzt sich besonders mit der Problematik der begrenzten Übertragungs-, Berechnungs- und Speicherkapazität auseinander. In vielen Fällen ist es wünschenswert, die Daten nur stückweise zu verarbeiten, ohne den großen

Datensatz als ganzes zu betrachten (*data stream*, *online monitoring*). Einbettungsmethoden sind dazu insofern in der Lage (vgl. Abschnitt 3.3).

Diese Arbeit ist im Rahmen des Sonderforschungsbereiches 876: „Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung“ im Teilprojekt C4: „Regressionsverfahren für sehr große, hochdimensionale Daten“ entstanden. Dieses Projekt befasst sich u. a. mit der Anwendung und Weiterentwicklung der Unterraum-Einbettungsmethode. Stellvertretend sei erneut auf Geppert *et al.* (2014), ferner auf Schwiegelshohn und Sohler (2014), Feldman *et al.* (2010) und Jabs (2012) verwiesen.

4.2 Andere Verfahren

Sehr nah an der Einbettungsmethodik sind die bei Deshpande *et al.* (2006) besprochenen Verfahren (*volume sampling*). Hier wird ebenfalls eine zufällige Auswahl der Zeilen der Matrix getroffen und iterativ verbessert. Ein Schwerpunkt liegt auf Cluster-Problemen.

Zur Erleichterung der in dieser Arbeit verwendeten MCMC-Verfahren schlagen Quiroz *et al.* (2014) auch die Arbeit mit zufällig ausgewählten Untermengen der Daten vor.

Als Alternative zur Lösung des Regressionsproblems bei großen Datensätzen schlagen Boutsidis *et al.* (2013) die Verwendung bestimmter Teilmengen (*coresets*) der Daten vor. Hier werden auch Schranken zur Güte der Approximation angegeben. Eine Gegenüberstellung mit der Unterraum-Einbettung, auch für *data-stream*-Anwendungen, findet sich bei Feldman *et al.* (2010).

Andere Verfahren versuchen nicht die Zahl der Beobachtungen zu reduzieren, sondern diejenige der Einflussvariablen. Jabs (2012) gibt einen Überblick über verschiedene Formen der recheneffizienten Dimensionsreduktion und vergleicht die Leistungsfähigkeit insbesondere bei Regression. Unter anderem wird die Hauptkomponentenanalyse betrachtet. Weitere Hinweise zum Umgang mit großen Datenmengen finden sich bei Hastie *et al.* (2009), insbesondere zur gleichfalls Einbettung genannten nicht-linearen Dimensionsreduktion (ebd., Kap. 14.9).

Shah und Meinshausen (2013) betrachten ebenfalls verschiedene Regressionsprobleme, wobei die Datenmatrix in beiden Richtungen groß, jedoch dünn besetzt ist. Die Reduktion geschieht durch sogenanntes *min-wise hashing*, welches hier auf reellwertige Daten anwendbar gemacht wird.

Schließlich lassen sich andere Möglichkeiten zur Berechnung der *posterior* in Bayesschen Modellen finden. Insbesondere bei großen Datensätzen ist die Recheneffizienz ein wesentliches Problem. Neben den hier verwendeten MCMC-Methoden kommen approximative Verfahren in Frage: Ebenfalls auf Simulation beruht *approximate Bayesian computation* (ABC), wie sie z. B. bei Marjoram *et al.* (2003) beschrieben ist. Kombinationen mit MCMC sind denkbar (ebd.). Wesentlich effizienter zu berechnen sind die *a-posteriori*-Randdichten über *integrated nested Laplace approximations* (INLA) nach Rue *et al.* (2009), sofern nur wenige Hyperparameter vorhanden sind.

Kapitel 5

Simulationsstudie

Um einen Eindruck von der tatsächlichen Leistungsfähigkeit des Einbettungsverfahrens zu gewinnen, wird eine empirische Untersuchung an ausgewählten hierarchischen Modellen durchgeführt. Dabei werden die *posteriors* sowohl aus einem großen Datensatz als auch aus den daraus erzeugten Skizzen berechnet und die Ergebnisse verglichen.

Die verwendeten Modelle und der allgemeine Aufbau der Simulationsstudie werden im nächsten Abschnitt eingeführt. Es folgen die Darstellung der Ergebnisse und eine erste Auswertung. Aufgrund dieser scheint eine kurze zusätzliche Untersuchung sinnvoll, die im Anschluss vorgestellt wird. Zuletzt erfolgt eine Zusammenfassung der wichtigsten Beobachtungen.

Die näherungsweise Bestimmung der *posteriors* mittels eines MCMC-Verfahrens erfolgt mit Hilfe der Software „OpenBUGS“ (OpenBUGS Project Management Group, 2012), die Vor- und Nachbereitung der Daten sowie die Darstellung der Ergebnisse mit der Software „R“ (R Core Team, 2014); insbesondere wird das „R“-Paket „ls2mat“ (Quedenfeld *et al.*, 2014, vgl. Abschnitt 3.3) für die Berechnung der Skizzen verwendet.

5.1 Aufbau

Die Ergebnisse von Geppert *et al.* (2014) beziehen sich auf eine Regressionsanalyse mit Designmatrix $X \in \mathbb{R}^{n \times k}$, Regressionsparametern $\beta \in \mathbb{R}^k$ und einer Zielvariablen $y \in \mathbb{R}^n$ mit

$$p(y_i|\beta) = N(X_{i,\cdot}\beta, s^2), \quad i = 1, \ldots, n. \tag{5.1}$$

Die verwendeten Datensätze haben daher die Form (y, X). In dieser Arbeit wird β zusätzlich von Hyperparametern θ bestimmt. Das Modell 4 (s. u.) enthält dagegen ein verallgemeinertes lineares Modell mit einer Einflussvariable und mehreren Zielvariablen.

Die Daten werden auf Grundlage des jeweiligen Modells simuliert: zuerst θ, daraus β, daraus zuletzt y. Dabei werden die Parameter als Konstante oder gemäß einer bestimmten Verteilung fest gewählt. Im Vergleich mit diesen bekannten wahren Werten lässt sich die Güte der Regression einschätzen, nachdem die *posterior* der Parameter aus den Daten berechnet wurde.

Als *prior* für die Hyperparameter wird in der Regel ein nicht-informativer gewählt. Hiervon wird abgewichen, wenn dies schon bei Verwendung des großen Datensatzes zu keinen plausiblen Ergebnissen führt, was vor allen bei Varianzparametern der Fall ist. Hier wird nötigenfalls die wahre Verteilung als *prior* verwendet.

In jedem Fall enthält der große Datensatz (y, X) $1 + k = 6$ Spalten und $n = 10^4$ u.i.v. Beobachtungen. Dies entspricht nicht der Größenordnung, mit der in Anwendungen tatsächlich zu rechnen ist, ermöglicht aber die vergleichende Berechnung der *posterior* auch aus dem großen Datensatz (vgl. Geppert *et al.*, 2014, Kap. 5). Zur Erzeugung der Skizzen werden die drei Einbettungsmethoden BCH, CW und SRHT sowie je die Approximationsparameter $\epsilon = 0.1$ und $\epsilon = 0.2$ verwendet. Tabelle 5.1 zeigt die jeweils nötige Anzahl Beobachtungen in den Skizzen.

Tabelle 5.1: Größe der Skizzen nach Einbettungsmethode und Approximationsparameter.

	BCH	CW	SRHT
$\epsilon = 0.1$	1075	8192	1076
$\epsilon = 0.2$	268	2048	269

Abgesehen vom Modell 4 enthält die Designmatrix X eine konstante Spalte $x_1 = 1$ (*intercept*) und vier stochastische Einflussvariablen: Realisierungen von

$$x_2 \sim R(-5, 5), \quad x_3 \sim R(0, 10), \quad x_4 \sim N(0, 3^2), \quad x_5 \sim N(-5, 3^2). \tag{5.2}$$

Zur Vereinfachung, zur leichteren Vergleichbarkeit der Modelle und um Einflüsse besser nachvollziehbar zu machen, wird die Varianz der *likelihood* (5.1) mit $s = 0.5$ festgehalten.

Nachfolgend werden die Modelle im Einzelnen beschrieben. Ggf. sind Normalvertei-

lungen so skaliert, dass auch eine Halbachse als Parameterraum „praktisch sicher“ eingehalten wird, falls beispielsweise nur positive Zahlen erreicht werden sollen.

Modell 0: Nicht-Hierarchisch Zur Erprobung des Verfahrens und als allgemeiner Vergleichsmaßstab wird ein einfaches Regressionsmodell ohne Hyperparameter betrachtet:

$$\begin{aligned} \theta &\in \emptyset \\ p(\beta|\theta) &= N((1,2,-1,-3,1)^\top, 0.5^2 \mathrm{Id}_5) \end{aligned} \tag{5.3}$$

Modell 1a: Normalverteilung mit gleichen Varianzen Hier sind die Verteilungsparameter von β nicht mehr fest, sondern ihrerseits verteilt:

$$\begin{aligned} \theta = (\mu, \sigma) &\in \mathbb{R}^5 \times \mathbb{R}_+ \\ p(\beta|\theta) &= N(\mu, \sigma^2 \mathrm{Id}_5) \\ p(\mu) &= N((1,2,-1,-3,1)^\top, 0.3^2 \mathrm{Id}_5) \\ p(\sigma) &= N(0.2, 0.025^2) \end{aligned} \tag{5.4}$$

Modell 1b: Normalverteilung mit verschiedenen Varianzen

$$\begin{aligned} \theta = (\mu, \sigma) &\in \mathbb{R}^5 \times \mathbb{R}_+^5 \\ p(\beta|\theta) &= N(\mu, \mathrm{diag}(\sigma)^2) \\ p(\mu) &= N((1,2,-1,-3,1)^\top, 0.3^2 \mathrm{Id}_5) \\ p(\sigma) &= N\left(\begin{pmatrix} 0.2 \\ 0.4 \\ 0.05 \\ 0.2 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.025^2 & & & & 0 \\ & 0.025^2 & & & \\ & & 0.01^2 & & \\ & & & 0.01^2 & \\ 0 & & & & 0.1^2 \end{pmatrix} \right) \end{aligned} \tag{5.5}$$

Modell 2a: Gammaverteilung Hier und im Folgenden werden andere Verteilungsfamilien verwendet:

$$\begin{aligned} \theta = (\alpha, \gamma) &\in \mathbb{R}_+^5 \times \mathbb{R}_+ \\ p(\beta_j|\theta) &= \Gamma(\alpha_j, \gamma), \quad j = 1, \ldots, 5 \\ p(\alpha) &= N((2,4,2,6,2)^\top, 0.3^2 \mathrm{Id}_5) \\ p(\gamma) &= N(2, 0.3^2) \end{aligned} \tag{5.6}$$

Modell 2b: Exponentialverteilung

$$\begin{aligned} \theta = \lambda &\in \mathbb{R}_+^5 \\ p(\beta_j|\theta) &= Exp(\lambda_j), \quad j = 1, \ldots, 5 \\ p(\lambda) &= N((1, 0.5, 1, 0.\bar{3}, 1)^\top, 0.05^2 \mathrm{Id}_5) \end{aligned} \tag{5.7}$$

Modell 3: Diskrete Verteilung

$$\begin{gathered}\theta = (\mu, \pi) \in \mathbb{R}^5 \times (0,1) \\ p(\beta|\theta) = N(\mu, 0.3^2 \mathrm{Id}_5) \\ P(\mu = (1,2,-1,-3,1)^\top) = 1-\pi, \quad P(\mu = (-1,0,-3,-5,-1)^\top) = \pi \\ \pi = 0.25\end{gathered} \tag{5.8}$$

Hier wie auch im folgenden Modell sind Hyperparameter für die Simulation fest gewählt. Bei der Berechnung der *posterior* werden sie aber als Unbekannte interpretiert und geschätzt.

Modell 4: Logistisches Wachstum Dieses Modell stammt aus einem Anwendungsbeispiel in „OpenBUGS“ (OpenBUGS Project Management Group, 2012, *Examples Volume II*, *Multivariate Orange trees*). Die ursprünglichen Daten stammen von (Draper und Smith, 1981, Kap. 24, *Exercise N*) und wurden von Lindstrom und Bates (1990, Kap. 7.1) in einem logistischen Regressionsmodell angewendet. Sie beschreiben das Wachstum von $k = 5$ Orangenbäumen in Abhängigkeit von der Zeit. Der große Datensatz in dieser Arbeit ist ihnen nachempfunden, wobei einige Hyperparameter festgehalten sind, um eine realistische Verteilung zu erzielen. Tatsächlich werden die Bäume zu einigen Zeitpunkten wieder kleiner, was aber in der Größenordnung möglicher Messfehler liegt.

Die Designmatrix besteht hier nur aus einer Spalte $x \in \mathbb{R}^n$ mit $n = 10^4$ geordneten stochastischen Zeitpunkten aus $R(0, 2000)$. Es gibt zu jedem Baum eine Zielvariable $y_j \in \mathbb{R}^n$, $j = 1, \ldots, k$. Mit den Transformationen

$$\begin{gathered}\eta_{j(i)} = \frac{\beta_{j1}}{1 + \beta_{j2} \cdot \exp(\beta_{j3} \cdot x_{(i)})}, \quad j = 1, \ldots, k, \quad i = 1, \ldots, n \\ \beta_{j1} = \exp(\lambda_{j1}), \quad \beta_{j2} = \exp(\lambda_{j2}) - 1, \quad \beta_{j3} = -\exp(\lambda_{j3}), \quad j = 1, \ldots, k\end{gathered} \tag{5.9}$$

sind nun $\lambda \in \mathbb{R}^{5\times 3}$ und $\chi \in \mathbb{R}_+$ die interessierenden Parameter, $\beta_{j3} \in \mathbb{R}$, $j = 1, \ldots, k$, die eigentlichen Regressionsparameter und $\eta_j \in \mathbb{R}^n$ die Erwartungswerte der Zielvariablen. Damit ist das Modell

$$\begin{gathered}p(y_{j(i)}|\lambda, \chi) = N(\eta_{j(i)}, \chi^2), \quad j = 1, \ldots, k, \quad i = 1, \ldots, n \\ \theta = (\mu, \tau) \in \mathbb{R}^3 \times \mathbb{R}^{3\times 3} \\ p(\lambda_j|\theta) = N(\mu, \tau^{-1}), \quad j = 1, \ldots, k \\ \mu = (5, 2, -6)^\top \\ \tau = \begin{pmatrix} 38 & -10 & -1 \\ -10 & 40 & -7 \\ -1 & -7 & 48 \end{pmatrix} \\ \chi = 2\end{gathered} \tag{5.10}$$

gegeben.

5.2 Ergebnisse

In jedem Modell sind die *a-posteriori*-Verteilungen der jeweiligen Parameter, auch der Hyperparameter, zu berechnen bzw. empirisch anzunähern. Dies geschieht durch einen MCMC-Algorithmus von „OpenBUGS“ (OpenBUGS Project Management Group, 2012) jeweils für den großen Datensatz und für die sechs Skizzen.

Die Tabellen 5.3 bis 5.14 in diesem Abschnitt zeigen die Ergebnisse und werden im Folgenden erläutert. Wegen des großen Umfangs sind sie zum Teil nur beispielhaft und werden durch die Tabellen B.1 bis B.13 im Anhang B vervollständigt.

Es sind verschieden lange *burn-in*-Phasen notwendig: Sie sind so bestimmt, dass die Markov-Kette nach Augenschein konvergiert ist, also über einen hinreichend langen Zeitraum keine systematischen Änderungen der empirischen Verteilungen mehr feststellbar sind, mindestens aber 1000 Iterationen lang. Tabelle 5.2 gibt einen Überblick der benötigten *burn-in*-Phasen.

Tabelle 5.2: Längen der *burn-in*-Phasen in Einheiten von 1000 Iterationen.

	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
Modell	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
0	1	1	1	1	1	1	1
1a	1	1	1	1	1	1	1
1b	1	1	1	1	1	1	1
2a	4	2	3	16	4	3	10
2b	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1
4	10	5	2	2	8	10	3

Nach der *burn-in*-Phase werden die nächsten 10^4 Werte als empirische *posterior* betrachtet und die angegebenen Kennzahlen der Verteilung aus ihr berechnet. Die in den Tabellen als *posterior* angegebene Verteilungsfamilie ist lediglich empirisch festgestellt und lässt sich etwa durch *qq plots* überprüfen.

Weiter zeigt jeder Tabellenabschnitt den betrachteten Parameter des Modells und seinen wahren Wert: eine Konstante oder Verteilung, die bei der Simulation der Daten verwendet wurde. Der bei der Berechnung verwendete *prior* schließlich ist in den meisten Fällen nicht-informativ: etwa eine Normalverteilung um 0 mit sehr großer Varianz („N flach“) oder eine Gleichverteilung auf der „fast ganzen“ Halbachse. Wenn sich bei versuchsweiser Verwendung des großen Datensatzes auf andere

Weise kein plausibles Ergebnis erzielen lässt, wird ein informativer *prior*, nötigenfalls der wahre Wert, verwendet. Ein *prior* ist nur bei denjenigen Parametern angegeben, die nicht durch Hyperparameter bestimmt werden.

Alle Tabellen zeigen das arithmetische Mittel, die empirische Standardabweichung und den Stichprobenmedian. So ist auch eine Asymmetrie bei Abweichung von der Normalverteilung zu erkennen. Ggf. sind zusätzlich Schätzwerte derjenigen Parameter angegeben, welche die Verteilungsfamilie der *posterior* charakterisieren.

Vor einer Auswertung der Ergebnisse im nächsten Abschnitt verbleiben noch einige Resultate zu notieren, die aus den Tabellen allein nicht zu erkennen sind:

Tabelle 5.3: *A-posteriori*-Verteilungen der Parameter in **Modell 0**. Weitere Ergebnisse im Anhang auf Seite 77.

$\boldsymbol{\beta_1}$							wahr: $N(1, 0.5^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.010	0.751	0.890	1.299	1.124	0.767	1.861
Stdabw.	0.013	0.013	0.019	0.013	0.014	0.031	0.013
Median	1.010	0.751	0.889	1.299	1.124	0.767	1.861

$\boldsymbol{\beta_4}$							wahr: $N(-3, 0.5^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.020	-3.032	-3.048	-3.041	-2.964	-2.953	-3.063
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	-3.020	-3.032	-3.048	-3.041	-2.964	-2.953	-3.063

Bei allen Modellen außer 4 kommt es vor, dass die Markov-Kette der eigentlichen Regressionsparameter β nur sehr wenige Werte annimmt und die Verteilung eher an eine umskalierte Binomialverteilung erinnert. Die Varianz ist relativ zur gegebenen Größenordnung jedoch sehr klein (vgl. z. B. Tabelle 5.4). Damit kann in jedem Fall davon gesprochen werden, dass eher ein konstanter Wert als eine Verteilung geschätzt wird. Dies ist durchaus das Ziel einer – auch Bayesschen – Regressionsanalyse.

Die *posterior* der Skalenparameter σ in den Modellen 1a und 1b hängt erkennbar nicht vom verwendeten Datensatz ab. Dies geht so weit, dass die jeweiligen Markov-Ketten zu gleichen Zeitpunkten genau gleiche Werte annehmen. Die als *prior* verwendete wahre Verteilung bleibt erhalten (vgl. z. B. Tabelle 5.6).

Tabelle 5.4: *A-posteriori*-Verteilungen der β-Parameter in **Modell 1a**. Weitere Ergebnisse im Anhang auf Seite 78.

$\boldsymbol{\beta_1}$					wahr: $N(\mu_1, \sigma^2)$, $\mathbb{E}(\mu_1) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.042	1.870	1.040	0.025	1.706	0.837	0.784
Stdabw.	0.013	0.013	0.015	0.013	0.013	0.029	0.013
Median	1.042	1.869	1.040	0.025	1.706	0.837	0.784

$\boldsymbol{\beta_2}$					wahr: $N(\mu_2, \sigma^2)$, $\mathbb{E}(\mu_2) = 2$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.993	2.033	1.985	2.045	1.856	1.959	2.067
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	1.993	2.033	1.985	2.045	1.856	1.959	2.067

Tabelle 5.5: *A-posteriori*-Verteilungen der μ-Parameter in **Modell 1a**. Weitere Ergebnisse im Anhang auf Seite 79.

$\boldsymbol{\mu_1}$					wahr: $N(1, 0.3^2)$		
prior: N flach					*posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.045	1.872	1.042	0.028	1.708	0.839	0.787
Stdabw.	0.201	0.201	0.201	0.201	0.201	0.203	0.201
Median	1.045	1.873	1.042	0.028	1.709	0.840	0.787

$\boldsymbol{\mu_2}$					wahr: $N(2, 0.3^2)$		
prior: N flach					*posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.991	2.032	1.984	2.043	1.854	1.957	2.066
Stdabw.	0.201	0.201	0.201	0.201	0.201	0.201	0.201
Median	1.993	2.033	1.985	2.045	1.856	1.959	2.067

Tabelle 5.6: *A-posteriori*-Verteilung von σ in **Modell 1a**.

$\boldsymbol{\sigma}$							wahr: $N(0.2, 0.025^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.199	0.199	0.199	0.199	0.199	0.199	0.199
Stdabw.	0.025	0.025	0.025	0.025	0.025	0.025	0.025
Median	0.200	0.200	0.200	0.200	0.200	0.200	0.200

Tabelle 5.7: *A-posteriori*-Verteilungen der Parameter in **Modell 1b**. Weitere Ergebnisse im Anhang ab Seite 80.

$\boldsymbol{\beta_3}$							wahr: $N(\mu_3, \sigma_3^2)$, $\mathbb{E}(\mu_3) = -1$
							posterior: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.006	-1.045	-1.009	-0.903	-1.015	-0.994	-0.914
Stdabw.	0.002	0.002	0.003	0.002	0.002	0.004	0.002
Median	-1.006	-1.045	-1.009	-0.903	-1.015	-0.994	-0.914

$\boldsymbol{\mu_3}$							wahr: $N(-1, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.006	-1.046	-1.010	-0.904	-1.016	-0.995	-0.914
Stdabw.	0.052	0.052	0.052	0.052	0.052	0.052	0.052
Median	-1.006	-1.046	-1.010	-0.904	-1.016	-0.995	-0.915

$\boldsymbol{\sigma_3}$							wahr: $N(0.05, 0.01^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.050	0.050	0.050	0.050	0.050	0.050	0.050
Stdabw.	0.010	0.010	0.010	0.010	0.010	0.010	0.010
Median	0.050	0.050	0.050	0.050	0.050	0.050	0.050

Tabelle 5.8: *A-posteriori*-Verteilungen der Parameter in **Modell 2a**. Weitere Ergebnisse im Anhang ab Seite 84.

$\boldsymbol{\beta_5}$							wahr: $\Gamma(\alpha_5, \gamma)$, $\mathbb{E}(\beta_5) = 1$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.036	1.047	0.967	0.923	1.038	0.962	0.845
Stdabw.	0.004	0.003	0.003	0.005	0.004	0.003	0.001
Median	1.035	1.047	0.969	0.921	1.035	0.962	0.845

$\boldsymbol{\alpha_5}$							wahr: $N(2, 0.3^2)$
prior: $R(0.1, 20)$							*posterior*: $\Gamma(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	4.447	4.458	4.366	4.242	4.415	4.406	4.129
$\hat{b}$	1.349	1.352	1.397	1.396	1.346	1.404	1.444
Mittel	3.296	3.298	3.126	3.038	3.280	3.138	2.859
Stdabw.	1.563	1.562	1.496	1.475	1.561	1.495	1.407
Median	3.119	3.122	2.958	2.854	3.087	2.998	2.703

$\boldsymbol{\gamma}$							wahr: $N(2, 0.3^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.212	2.209	2.203	2.219	2.197	2.206	2.211
Stdabw.	0.283	0.286	0.286	0.283	0.281	0.286	0.289
Median	2.217	2.210	2.203	2.224	2.195	2.216	2.211

Tabelle 5.9: *A-posteriori*-Verteilungen der Parameter in **Modell 2b**. Weitere Ergebnisse im Anhang ab Seite 86.

$\boldsymbol{\beta_2}$							wahr: $Exp(\lambda_2)$, $\mathbb{E}(\beta_2) = 2$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.046	1.903	2.017	2.097	1.798	2.061	1.859
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	2.046	1.903	2.017	2.097	1.798	2.061	1.859

$\boldsymbol{\lambda_2}$							wahr: $N(0.5, 0.05^2)$
prior: $R(10^{-5}, 10^5)$							*posterior*: $\Gamma(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	2.023	2.013	2.012	2.051	1.933	2.018	1.995
$\hat{b}$	2.064	1.917	2.027	2.157	1.762	2.068	1.861
Mittel	0.980	1.050	0.993	0.951	1.097	0.976	1.072
Stdabw.	0.689	0.740	0.700	0.664	0.789	0.687	0.759
Median	0.828	0.875	0.836	0.808	0.911	0.820	0.895

Tabelle 5.10: *A-posteriori*-Verteilungen der Parameter in **Modell 3**. Weitere Ergebnisse im Anhang auf Seite 88.

$\boldsymbol{\beta_3}$							wahr: $N(\mu_3, 0.3^2)$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.474	-1.529	-1.484	-1.501	-1.466	-1.462	-1.656
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	-1.474	-1.529	-1.484	-1.501	-1.466	-1.462	-1.656

$\boldsymbol{\pi}$							wahr: 0.25
prior: $R(0, 1)$							*posterior*: $B(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	1.016	1.016	1.016	1.016	1.016	1.016	1.016
$\hat{b}$	2.018	2.018	2.018	2.018	2.018	2.018	2.018
Mittel	0.335	0.335	0.335	0.335	0.335	0.335	0.335
Stdabw.	0.235	0.235	0.235	0.235	0.235	0.235	0.235
Median	0.298	0.298	0.298	0.298	0.298	0.298	0.298

Im Modell 3 sind die μ-Parameter hier nicht wiedergegeben. Ihre Markov-Kette bleibt typischerweise in einem der beiden möglichen Punkte konstant. Dies lässt das gesamte Ergebnis etwas fragwürdig erscheinen. Der Parameter π hängt wiederum nicht von den Daten ab, seine Verteilung entfernt sich aber vom *prior* hin zu einer offenbar typischen betaverteilten Form (Tabelle 5.10). Es ist denkbar, dass die Markov-Kette hier absorbierende Zustände erreicht. Wahrscheinlicher aber ist, dass der MCMC-Algorithmus nicht für solch ein teilweise diskretes Modell geeignet ist, da die *proposal distribution* nur äußerst selten einen der möglichen Werte erreicht und in der Folge kein neuer mehr akzeptiert wird.

Tabelle 5.11: *A-posteriori*-Verteilungen der λ_1-Parameter in **Modell 4**. Übrige λ entsprechend.

$\boldsymbol{\lambda_{11}}$							wahr: $\lambda_1 \sim N(\mu, \tau^{-1})$, $\mu_1 = 5$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	5.192	6.667	5.138	6.620	7.215	1.583	7.252
Stdabw.	0.000	0.047	0.011	0.053	0.106	0.285	0.093
Median	5.192	6.667	5.138	6.621	7.210	1.613	7.252

$\boldsymbol{\lambda_{12}}$							wahr: $\lambda_1 \sim N(\mu, \tau^{-1})$, $\mu_2 = 2$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.200	3.176	3.274	3.173	3.191	-7.585	3.249
Stdabw.	0.001	0.118	0.103	0.137	0.269	0.476	0.239
Median	2.199	3.174	3.270	3.167	3.173	-7.633	3.237

$\boldsymbol{\lambda_{13}}$							wahr: $\lambda_1 \sim N(\mu, \tau^{-1})$, $\mu_3 = -6$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-5.774	-7.021	-5.322	-6.984	-7.591	-11.014	-7.604
Stdabw.	0.001	0.061	0.039	0.070	0.137	0.285	0.119
Median	-5.774	-7.020	-5.322	-6.985	-7.596	-10.990	-7.605

Die Normalverteilungsannahme der *posterior* für die τ_{kl}-Parameter im Modell 4 mit $k \neq l$ kann nicht völlig richtig sein, wie auch am Median zu sehen ist (Tabelle 5.13). Vielmehr ist hier eine leichte Schiefe zu beobachten. Der Parameter χ zeigt beim großen Datensatz die schon für die β-Parameter der anderen Modelle beschriebene Beschränkung auf wenige Werte.

Tabelle 5.12: *A-posteriori*-Verteilungen der μ-Parameter in **Modell 4**.

$\boldsymbol{\mu_1}$							wahr: 5
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	5.020	6.397	4.896	6.361	6.990	1.387	7.029
Stdabw.	0.103	0.122	0.108	0.126	0.149	0.282	0.135
Median	5.021	6.397	4.897	6.363	6.989	1.404	7.029

$\boldsymbol{\mu_2}$							wahr: 2
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.098	3.176	3.235	3.193	3.198	-7.528	3.251
Stdabw.	0.099	0.128	0.116	0.141	0.254	0.471	0.220
Median	2.098	3.175	3.235	3.189	3.181	-7.485	3.246

$\boldsymbol{\mu_3}$							wahr: -6
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-6.006	-7.037	-5.435	-7.005	-7.638	-10.949	-7.657
Stdabw.	0.140	0.097	0.119	0.100	0.149	0.274	0.130
Median	-6.005	-7.037	-5.436	-7.004	-7.641	-10.930	-7.658

Tabelle 5.13: *A-posteriori*-Verteilungen der τ-Parameter in **Modell 4**. Weitere Ergebnisse im Anhang auf Seite 89.

$\boldsymbol{\tau_{11}}$							wahr: 38
prior: $\tau \sim Wish(10 \cdot \mathrm{Id}_3, 3)$							*posterior*: $\Gamma(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	3.442	3.099	3.358	3.044	2.604	2.921	2.712
$\hat{b}$	0.070	0.091	0.079	0.084	0.068	0.071	0.069
Mittel	49.249	34.202	42.650	36.215	38.219	41.016	39.593
Stdabw.	26.544	19.430	23.273	20.757	23.685	23.997	24.041
Median	44.250	30.565	38.340	31.930	33.000	36.170	34.615

$\boldsymbol{\tau_{12}}$							wahr: -10
prior: $\tau \sim Wish(10 \cdot \mathrm{Id}_3, 3)$							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-15.670	-0.851	-2.862	0.652	-0.705	-6.143	-0.778
Stdabw.	20.314	18.168	19.806	18.800	18.096	16.847	18.389
Median	-13.860	-0.553	-2.680	0.602	-0.793	-5.340	-0.817

$\boldsymbol{\tau_{13}}$							wahr: -1
prior: $\tau \sim Wish(10 \cdot \mathrm{Id}_3, 3)$							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-2.971	0.246	-7.957	-0.219	2.957	-4.630	0.862
Stdabw.	12.985	18.254	15.903	18.909	19.005	18.686	19.028
Median	-2.520	0.252	-6.594	0.039	3.008	-4.290	1.048

Tabelle 5.14: *A-posteriori*-Verteilung von χ in **Modell 4**.

$\boldsymbol{\chi}$							wahr: 2
prior: $\chi^{-2} \sim \Gamma(10^{-3}, 10^{-3})$							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.998	200.761	68.165	221.200	472.673	18.678	410.608
Stdabw.	0.006	1.947	0.239	2.132	9.148	0.129	7.937
Median	1.998	200.700	68.160	221.200	472.600	18.680	410.500

5.3 Auswertung

Es ist zu untersuchen, in welchen Situationen die Schätzung der *posterior* auf Grundlage der Skizzen der Schätzung aus dem großen Datensatz nahekommt. Insbesondere für die Hyperparameter ist hier keine Gesetzmäßigkeit bekannt und es wird aufgrund der Ergebnisse des vorigen Abschnitts zunächst versucht, deskriptiv zu einer Einschätzung zu gelangen.

Für die Güte der Einbettungsmethode ist es unerheblich, ob das Ergebnis der in der Simulation bekannten Wahrheit entspricht, solange es mit demjenigen des großen Datensatzes übereinstimmt. Faktoren wie eine ungünstige Wahl des Modells oder des *priors* sowie der in „OpenBUGS“ nicht genau zu kontrollierende MCMC-Algorithmus können eine schlechte Schätzung trotz korrekter Daten verursachen. Hier können die Skizzen dennoch den großen Datensatz korrekt repräsentieren. In Fällen, bei denen das Ergebnis nicht oder kaum von den Daten abhängt (s. o.), lässt sich dies nicht sagen.

Es ist weiter anzumerken, dass alle Datensätze jeweils eine einzige Stichprobe darstellen. Eine asymptotische Näherung an die wahre Verteilung wird zwar durch die MCMC-Berechnung der *posterior* erreicht, dies kann aber zufällige Abweichungen in den Daten nicht völlig ausgleichen. Der große Datensatz besteht jedoch aus sehr vielen u.i.v Beobachtungen, weshalb er im Mittel der wahren Verteilung nahekommen sollte. Die Skizzen sind wiederum aus ihm berechnet und könnten mögliche Zufallsschwankungen nachvollziehen. Insgesamt deuten kleine Abweichungen der Ergebnisse vom wahren Wert und untereinander noch nicht auf eine grundsätzlich falsche Berechnung. Den Skizzen selbst liegt ein weiterer Zufallsprozess zugrunde (vgl. Kapitel 3), der für gewisse leichte Abweichungen mitverantwortlich sein kann.

Modell 0 bestätigt nach Augenschein die Resultate von Geppert *et al.* (2014), wonach die Abweichungen der Regressionsparameter zwischen Skizze und großem Datensatz beschränkt sind. Tabelle 5.3 zeigt dennoch eine deutliche Abweichung für den *intercept*. In der Tat hat die „Einserspalte“ bei den Skizzen nicht mehr diese Form, sondern enthält auch andere Werte (vgl. Kapitel 3). Die übrigen Parameter zeigen keine nennenswerten Abweichungen und eine durchgehend sehr geringe Varianz, wie schon im vorigen Abschnitt ausgeführt wird.

Dieses Ergebnis gilt grundsätzlich auch für die β-Parameter in den übrigen Modellen mit Ausnahme von 4. Tendenziell zeigen komplexere Modelle (2a, 2b, 3) die größeren Abweichungen zwischen großem Datensatz und Skizzen. Hier ist auch die zu erwartende Verschlechterung bei großem ϵ zu erkennen.

Eine Möglichkeit, die Güte der Einbettungsmethoden quantitativ zu beurteilen, bietet die Ungleichung (3.6) aufgrund der Ergebnisse von Geppert *et al.* (2014). Alle Modelle außer 4 sind gewöhnliche lineare Regressionsmodelle, für deren Parametervektoren β (großer Datensatz) bzw. β' (Skizze) demnach gelten muss:

$$\frac{\|X\hat{\beta}' - y\|_2}{\|X\hat{\beta} - y\|_2} \leq \sqrt{\frac{1+\epsilon}{1-\epsilon}} \tag{5.11}$$

Wegen der sehr geringen Varianz wird hier nicht die ganze *posterior* der β als Schätzer verwendet, sondern lediglich das arithmetische Mittel. Tabelle 5.15 zeigt die linke Seite der Ungleichung (5.11) (Residuumsquotient) bei den einzelnen Modellen und Einbettungsmethoden. Die Bedingung ist in jedem Fall erfüllt. In Modell 2a ist das Residuum mit β' sogar kleiner, im Übrigen oft gleich.

Tabelle 5.15: Residuumsquotienten nach Modell und Einbettungsmethode. Nach Ungleichung (5.11) dürfen diese Werte höchstens $\sqrt{\frac{1+\epsilon}{1-\epsilon}}$ (rund 1.106 bzw. 1.225) sein.

	$\epsilon = 0.1$			$\epsilon = 0.2$		
Modell	BCH	CW	SRHT	BCH	CW	SRHT
0	1.002	1.000	1.001	1.010	1.001	1.007
1a	1.005	1.000	1.009	1.017	1.001	1.006
1b	1.004	1.000	1.005	1.010	1.001	1.013
2a	0.994	0.991	0.974	1.016	0.973	0.966
2b	1.001	1.000	1.000	1.010	1.001	1.008
3	1.005	1.000	1.000	1.009	1.000	1.009

Weiter kann der Abstand zwischen den Parametervektoren mit der Ungleichung (3.9) beurteilt werden. Nach den Ergebnissen in Tabelle 5.16 ist auch diese immer erfüllt. Beides ist noch kein Beweis für die Gültigkeit der Resultate von Geppert *et al.* (2014) im Falle eines hierarchischen Modells. Dass die beobachteten Werte diese oberen Schranken deutlich unterschreiten, deutet jedoch auf Gültigkeit hin. Die Abweichungen zwischen großem Datensatz und Skizze sind, wie zu erwarten ist, in der Regel für $\epsilon = 0.2$ größer, wobei es Ausnahmen gibt.

Zum Vergleich zeigt Tabelle 5.17 die Abstände zwischen den Parametervektoren, wenn der *intercept* aus ihnen entfernt wird. Demnach hat dieser einen besonders großen Anteil am Unterschied der Schätzwerte zwischen großem Datensatz und Skizze, wie schon an den obigen Ergebnissen zu erkennen ist.

All diese Befunde für die Regressionsparameter β sind durchaus nicht überraschend: Geppert *et al.* (2014, Kap. 3) setzen für die Gültigkeit ihrer Resultate lediglich

Tabelle 5.16: Differenzen der Schätzwerte der Regressionsparameter zwischen großem Datensatz und Skizzen nach Modell und Einbettungsmethode mit ihrer oberen Schranke nach Ungleichung (3.9).

	$\epsilon = 0.1$				$\epsilon = 0.2$			
Modell	Schranke	BCH	CW	SRHT	Schranke	BCH	CW	SRHT
0	5.751	0.270	0.126	0.300	8.626	0.224	0.256	0.868
1a	4.228	0.834	0.010	1.029	6.342	0.693	0.214	0.279
1b	4.962	0.617	0.026	0.258	7.442	0.241	0.147	0.319
2a	29.236	0.094	0.675	1.051	43.854	0.304	1.112	1.294
2b	17.497	0.589	0.121	0.452	26.246	3.537	1.014	1.542
3	7.308	0.222	0.090	0.146	10.962	0.768	0.098	1.011

Tabelle 5.17: Differenzen der geschätzten Parametervektoren ohne den *intercept* analog Tabelle 5.16.

	$\epsilon = 0.1$			$\epsilon = 0.2$		
Modell	BCH	CW	SRHT	BCH	CW	SRHT
0	0.074	0.038	0.083	0.193	0.079	0.173
1a	0.105	0.010	0.159	0.199	0.059	0.106
1b	0.070	0.026	0.123	0.195	0.071	0.213
2a	0.028	0.091	0.168	0.251	0.203	0.454
2b	0.188	0.046	0.092	0.480	0.212	0.460
3	0.165	0.012	0.041	0.218	0.028	0.233

voraus, dass die β einer – unbekannten – Verteilung folgen und die Daten gemäß der *likelihood* (5.1) beschrieben werden. Dafür scheint es unerheblich, ob die unbekannte Verteilung von Hyperparametern bestimmt oder fest ist. Unterschiede zwischen den Schätzwerten aus großem Datensatz und Skizzen sind daher im Folgenden bei den Hyperparametern zu suchen.

Während für β eher ein Punkt geschätzt wird als eine Verteilung (s. o.), ist die Variabilität in den zugehörigen Hyperparametern zu finden. Bei Modellen mit einem Lokationsparameter μ (1a und 1b) weist dieser gerade die Varianz der wahren Verteilung von β auf; sie wird desweiteren durch die σ-Parameter geschätzt (vgl. etwa Tabelle 5.7). Dies alles gilt für die Skizzen und den großen Datensatz gleichermaßen. Die Erwartungswerte der μ geben die Wahrheit in fast allen Fällen gut wieder, weniger gut bei den Skizzen für den *intercept*, wie dies schon bei den β der Fall ist.

Auch bei den Hyperparametern der Modelle 2a und 2b sind keine auffälligen Unterschiede zwischen Skizzen und großem Datensatz zu erkennen (Tabellen 5.8 und 5.9). Diese Form- und (inversen) Skalenparameter werden jedoch nicht durchgehend in der richtigen Größenordnung geschätzt. Dies spricht nicht gegen die Einbettung, da diese offenbar die für den MCMC-Algorithmus wesentliche Struktur des Datensatzes erhält. Im Modell 3 wird π, wie im vorigen Abschnitt erwähnt, anscheinend unabhängig von den konkreten Daten bestimmt (Tabelle 5.10).

Nach den bisher besprochenen Ergebnissen erstreckt sich eine Übereinstimmung von großem Datensatz und Skizze bei den Regressionsparametern β auch auf die sie bestimmenden Hyperparameter. Die *posteriors* stimmen in der Regel überein, was sowohl die Verteilungsfamilie als auch deren Parameter betrifft.

Modell 4 besitzt eine andere Struktur. Die Rolle der Regressionsparameter wird hier von den λ bzw. ihren Transformationen gemäß der Gleichungen (5.9) eingenommen. Der große Datensatz liefert für sie Schätzwerte, welche der Wahrheit recht nahe kommen, bei den Skizzen wird die richtige Größenordnung dagegen zum Teil verfehlt, in einem Fall bis hin zum Vorzeichen (Tabelle 5.11). Hier zeigt sich auch erstmals eine Erhöhung der Varianz der *posterior* bei den Skizzen. Die Abweichung ist für $\epsilon = 0.1$ noch schwach ausgeprägt, für $\epsilon = 0.2$ dagegen deutlich. Dies trifft vor allem auf die Einbettungsmethode CW zu, obwohl ihre Skizze einen wesentlich größeren Umfang als die anderen hat.

Der selbe Befund gilt für die μ-Parameter in Modell 4, welche den Erwartungswert der λ modellieren (Tabelle 5.12). Sie unterscheiden sich von letzteren fast nur durch die erhöhte Varianz, wie es auch bei den entsprechenden Lokationsparametern in den Modellen 1a und 1b der Fall ist.

Für die Präzisionmatrix τ gelingt eine korrekte Schätzung bestenfalls bei den Diagonalelementen, wo die Größenordung der *posteriors* ungefähr mit den wahren Werten und untereinander übereinstimmt (Tabelle 5.13). Die Präzision nimmt hier mit ϵ gegenüber dem großen Datensatz ab, was mit der steigenden Varianz der λ korrespondiert. Es muss aber auf die besonders große Varianz der *posterior* hingewiesen werden. Bei den übrigen Elementen von τ sind die Schätzungen schon für den großen Datensatz nicht immer sinnvoll, die Skizzen weichen von diesem und noch weiter von den wahren Werten ab.

Die Varianzerhöhung im gesamten Modell 4 wird am Parameter χ besonders deutlich, welcher direkt die Standardabweichung in der *likelihood* beschreibt. Sie ist beim großen Datensatz realistisch, bei den Skizzen extrem erhöht, vor allem bei den kleineren nach BCH und SRHT (Tabelle 5.14). Wie schon bei τ handelt es sich um einen

reinen Skalenparameter. Dies kann ein Hinweis sein, dass die Einbettungsmethoden die Informationen über solche Größen nicht sehr gut erhalten, wenn sie nicht wie in den anderen Modellen als fest angenommen oder mit dem wahren *prior* geschätzt werden. Andererseits ist ein negativer Zusammenhang mit der Stichprobengröße denkbar: Weniger Information bedeutet größere Unsicherheit.

Bei Betrachtung der Daten in den Skizzen ist schließlich leicht zu sehen, dass die Zahlen in der Regel eine andere, höhere Größenordnung haben als im großen Datensatz. Ebenso sind andere Vorzeichen zu finden. Dies scheint für die anderen Modelle keine Schwierigkeiten zu verursachen, im generalisierten linearen Modell 4 jedoch schon, wo die „Messwerte“ natürlicherweise positiv sein sollten. Die Verteilungen und die *link*-Funktion sind zwar auch für andere Werte definiert, es kann so aber nicht mehr mit sinnvollen Ergebnissen gerechnet werden.

5.4 Folgestudie

Nach den obigen Ergebnissen der Hauptstudie erweist sich die Schätzung von Skalenparametern im komplexen Modell 4 als problematisch. Deren Schätzwerte wie auch die Varianzen der anderen Schätzer nehmen zum Teil deutlich zu, wenn eine Skizze anstelle des großen Datensatzes verwendet wird. Dagegen zeigt beispielsweise das Modell 1a, in dem Varianzen nur eine geringe Rolle spielen, bei allen Datensätzen günstige Ergebnisse.

Um einen möglichen Einfluss von Skalenparametern auf die Güte der Einbettung näher zu untersuchen, werden Modifikationen dieser beiden Modelle betrachtet. Der Aufbau dieser Folgestudie ist derselbe wie zuvor.

Modell 1a’: Varianzschätzung in der *likelihood* Das Modell 1a wird so abgeändert, dass die Varianz in der *likelihood* nicht mehr fest ist, sondern ebenfalls einer Verteilung folgt, welche durch die *posterior* zu schätzen ist:

$$\begin{aligned} p(y_i|\beta,\chi) &= N(X_{i,\cdot}\beta,\chi^2), \quad i=1,\dots,n \\ p(\chi) &= N(0.5, 0.025^2) \\ \theta &= (\mu,\sigma) \in \mathbb{R}^5 \times \mathbb{R}_+ \\ p(\beta|\theta) &= N(\mu,\sigma^2 \mathrm{Id}_5) \\ p(\mu) &= N((1,2,-1,-3,1)^\top, 0.3^2\mathrm{Id}_5) \\ p(\sigma) &= N(0.2, 0.025^2) \end{aligned} \tag{5.12}$$

Der neue Skalenparameter χ hängt nicht von Hyperparametern ab.

Modell 4': Fixierte Varianzen Modell 4 wird dagegen vereinfacht: Die Skalenparameter χ und τ werden bei der Berechnung der *posterior* nicht mehr als Variablen, sondern als feste bekannte Größen interpretiert:

$$\begin{gathered} p(y_{j(i)}|\lambda) = N(\eta_{j(i)}, 2^2), \quad j = 1, \ldots, k, \quad i = 1, \ldots, n \\ \theta = (\mu) \in \mathbb{R}^3 \\ p(\lambda_j|\theta) = N\left(\mu, \begin{pmatrix} 38 & -10 & -1 \\ -10 & 40 & -7 \\ -1 & -7 & 48 \end{pmatrix}^{-1}\right), \quad j = 1, \ldots, k \\ \mu = (5, 2, -6)^\top \end{gathered} \tag{5.13}$$

Die Durchführung und Auswertung der Studie erfolgt wie zuvor. Tabelle 5.18 zeigt die benötigten *burn-in*-Phasen im Vergleich mit den Modellen 1a und 4. Wie in Modell 1a wird der Parameter σ auch in Modell 1a' offenbar unabhängig von den gegebenen Daten geschätzt.

Tabelle 5.18: Längen der *burn-in*-Phasen in Einheiten von 1000 Iterationen.

	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
Modell	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
1a	1	1	1	1	1	1	1
1a'	1	1	1	1	1	1	1
4	10	5	2	2	8	10	3
4'	2	1	3	2	7	2	4

Die Ergebnisse für β, μ und σ in Modell 1a' (Tabellen 5.19 bis 5.21 sowie B.14 und B.15 im Anhang B) entsprechen weitgehend denjenigen von Modell 1a. Die *posterior* der β nimmt allerdings nicht mehr nur einzelne Werte an und zeigt eine etwas höhere Varianz. Es kann somit eher davon gesprochen werden, dass für die Regressionsparameter ganze Verteilungen geschätzt werden.

Die neu untersuchte Varianz χ in der *likelihood* (Tabelle 5.21) wird sowohl mit dem großen Datensatz als auch mit den Skizzen deutlich überschätzt. Dies korrespondiert mit der Varianzerhöhung bei β gegenüber Modell 1a, wo χ als feste Konstante angenommen wird. Die Varianz der *posterior* von χ wird jedoch kleiner geschätzt als sie tatsächlich ist.

Anders als zwischen den Modellen 1a und 1a' werden beim Übergang von Modell 4 zu 4' Skalenparameter entfernt. Dies hat gewissermaßen die gegenteilige Wirkung: Die *posterior* der verallgemeinerten Regressionsparameter λ (Tabelle 5.22) zeigt tatsächlich keine Normalverteilung mehr, sondern es werden nur noch wenige verschiedene

Tabelle 5.19: *A-posteriori*-Verteilungen der β-Parameter in **Modell 1a'**. Weitere Ergebnisse im Anhang auf Seite 90.

$\boldsymbol{\beta_1}$					wahr: $N(\mu_1, \sigma^2)$, $\mathbb{E}(\mu_1) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.064	1.408	0.922	0.917	1.560	0.917	1.434
Stdabw.	0.067	0.085	0.076	0.081	0.074	0.114	0.080
Median	1.063	1.406	0.922	0.916	1.559	0.916	1.433

$\boldsymbol{\beta_2}$					wahr: $N(\mu_2, \sigma^2)$, $\mathbb{E}(\mu_2) = 2$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.991	1.985	1.963	2.036	1.951	1.964	2.044
Stdabw.	0.009	0.011	0.010	0.011	0.010	0.015	0.010
Median	1.991	1.985	1.963	2.036	1.951	1.964	2.044

Tabelle 5.20: *A-posteriori*-Verteilungen der μ-Parameter in **Modell 1a'**. Weitere Ergebnisse im Anhang auf Seite 91.

$\boldsymbol{\mu_1}$					wahr: $N(1, 0.3^2)$		
prior: N flach					*posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.061	1.405	0.920	0.914	1.557	0.914	1.431
Stdabw.	0.214	0.221	0.217	0.219	0.217	0.233	0.219
Median	1.062	1.403	0.919	0.913	1.557	0.913	1.430

$\boldsymbol{\mu_2}$					wahr: $N(2, 0.3^2)$		
prior: N flach					*posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.986	1.980	1.959	2.031	1.946	1.959	2.039
Stdabw.	0.199	0.199	0.199	0.199	0.199	0.199	0.199
Median	1.986	1.980	1.958	2.031	1.946	1.959	2.039

Tabelle 5.21: *A-posteriori*-Verteilung von σ und χ in **Modell 1a'**.

$\boldsymbol{\sigma}$							wahr: $N(0.2, 0.025^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.200	0.200	0.200	0.200	0.200	0.200	0.200
Stdabw.	0.024	0.024	0.024	0.024	0.024	0.024	0.024
Median	0.200	0.200	0.200	0.200	0.200	0.200	0.200

$\boldsymbol{\chi}$							wahr: $N(0.5, 0.025^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.526	3.036	2.126	3.012	2.986	1.920	3.014
Stdabw.	0.011	0.013	0.010	0.012	0.013	0.013	0.013
Median	2.527	3.037	2.126	3.011	2.986	1.920	3.013

Werte angenommen, wie es bereits bei den Modellen 1a und 1b der Fall ist. Die beobachteten Größenordnungen bleiben jedoch dieselben wie bei Modell 4, auch was die Unterschiede zwischen großem Datensatz und Skizzen betrifft.

Letzteres gilt auch für die μ in Modell 4' (Tabelle 5.23). Im Unterschied zu Modell 4 ist die Varianz ihrer *posterior* etwas geringer. Dies passt wie bei 1a und 1a' dazu, dass nun weniger Skalenparameter modelliert sind.

5.5 Zusammenfassung

Es werden Datensätze aus verschiedenen Bayesschen Regressionsmodellen simuliert, wobei die wahren Parameter bzw. Hyperparameter, als Konstanten oder gemäß einer Verteilung, bekannt sind. Aus diesen großen Datensätzen werden mit den Einbettungsmethoden BCH, CW und SRHT Skizzen erzeugt. Bei allen Modellen und so erzeugten Datensätzen werden die vorhandenen Parameter durch ein MCMC-Verfahren geschätzt. Dabei sollen die wahren Werte möglichst wiedergefunden werden.

Sind die geschätzten *a-posteriori*-Verteilungen aus den Skizzen derjenigen aus dem großen Datensatz in gewisser Weise ähnlich, kann darauf geschlossen werden, dass die wesentliche Information trotz Reduktion der Daten erhalten ist. Dies ist nach

Tabelle 5.22: *A-posteriori*-Verteilungen der λ_1-Parameter in **Modell 4'**. Übrige λ entsprechend.

$\boldsymbol{\lambda_{11}}$					wahr: $\lambda_1 \sim N(\mu, S^2)$, $\mu_1 = 5$, S vgl. (5.13) *posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	5.192	6.755	5.138	6.695	7.467	0.810	7.449
Stdabw.	0.000	0.001	0.001	0.001	0.002	0.054	0.000
Median	5.192	6.755	5.137	6.695	7.469	0.804	7.449

$\boldsymbol{\lambda_{12}}$					wahr: $\lambda_1 \sim N(\mu, S^2)$, $\mu_2 = 2$, S vgl. (5.13) *posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.202	3.149	3.074	3.188	3.171	-8.848	3.157
Stdabw.	0.002	0.005	0.009	0.000	0.002	0.109	0.004
Median	2.202	3.150	3.075	3.188	3.171	-8.872	3.156

$\boldsymbol{\lambda_{13}}$					wahr: $\lambda_1 \sim N(\mu, S^2)$, $\mu_3 = -6$, S vgl. (5.13) *posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-5.772	-7.120	-5.390	-7.044	-7.785	-11.604	-7.807
Stdabw.	0.001	0.002	0.003	0.002	0.002	0.052	0.001
Median	-5.772	-7.120	-5.390	-7.045	-7.786	-11.610	-7.808

Tabelle 5.23: *A-posteriori*-Verteilungen der μ-Parameter in **Modell 4'**.

$\boldsymbol{\mu_1}$							wahr: 5
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	5.019	6.462	4.898	6.409	7.166	2.415	7.172
Stdabw.	0.076	0.075	0.076	0.076	0.075	0.076	0.076
Median	5.019	6.461	4.897	6.409	7.165	2.415	7.171

$\boldsymbol{\mu_2}$							wahr: 2
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.098	3.134	3.046	3.188	3.151	-3.671	3.154
Stdabw.	0.074	0.074	0.074	0.074	0.074	0.078	0.075
Median	2.098	3.134	3.046	3.188	3.151	-3.671	3.154

$\boldsymbol{\mu_3}$							wahr: -6
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-6.002	-7.118	-5.504	-7.051	-7.784	-9.069	-7.822
Stdabw.	0.065	0.064	0.066	0.065	0.065	0.067	0.065
Median	-6.003	-7.117	-5.505	-7.052	-7.784	-9.070	-7.822

Geppert *et al.* (2014) bei nicht-hierarchischen Modellen der Fall. Hierarchische Modelle bleiben hier zu untersuchen.

In den meisten Fällen können keine bedeutsamen Unterschiede zwischen großem Datensatz und Skizzen festgestellt werden. Dies betrifft zunächst die eigentlichen Regressionsparameter β (*intercept*, *slope*) bei gewöhnlichen linearen Modellen:

Nicht nur im nicht-hierarchischen Modell 0, sondern auch in den hierarchischen Modellen 1a, 1b, 2a, 2b und 3 liegt die Abweichung der β-Schätzer deutlich innerhalb der von Geppert *et al.* (2014) bestimmten Schranken. Unterschiede sind vor allem beim *intercept* zu erkennen, was mit der Struktur der Skizzen erklärt werden kann. Tendenziell wird die Abweichung größer, wenn der Wert ϵ, der Approximationsparameter der Einbettung, erhöht wird; die Skizze wird dabei kleiner. Auch ist die Abweichung bei komplexeren hierarchischen Modellen (2a, 2b und 3) deutlicher zu erkennen.

In all diesen Modellen ist auch die Abweichung bei den Hyperparametern im Allgemeinen gering. Beschreiben sie den Erwartungswert der β (Hyperparameter μ), so entspricht die geschätzte Varianz der μ der wahren Varianz der β. Die Regressionsparameter selbst werden dagegen oft mit einer sehr geringen Varianz, geradezu als Konstanten geschätzt; ihre *posteriors* nehmen dabei nur wenige verschiedene Werte an. Unterschiede zwischen großem Datensatz und Skizzen beim Erwartungswert der *μ-posterior* finden sich wiederum vor allem beim *intercept*.

Hyperparameter, welche die Varianz der Regressionsparameter beschreiben, werden oft überschätzt (vgl. Modelle 2a, 2b und 4). Werden sie dagegen im Modell als Konstanten fixiert (fester wahrer Wert), tritt die genannte sehr geringe Varianz der β-Schätzer auf (3 und 4'). Allgemein scheint die Schätzung von Skalenparametern durch MCMC-Verfahren hier nur sehr begrenzt möglich. In einigen Fällen erfolgt sie offensichtlich datenunabhängig (1a, 1b und 3).

Ähnliche Schwierigkeiten zeigen sich auch in Modellen, in welchen nicht alle Parameter normalverteilt sind (2a, 2b und 3). Hier werden Hyperparameter sowohl mit dem großen Datensatz als auch mit den Skizzen verschätzt. Wird die Varianz aus der *likelihood* nicht festgehalten, sondern geschätzt (Modelle 1a' und 4), sind die Schätzwerte zu hoch.

Im generalisierten linearen Modell 4 zeigen sich dagegen Unterschiede zwischen den Datensätzen. Hier werden die wahren Werte der nicht-linear transformierten Regressionsparameter λ bei Schätzung mit dem großen Datensatz annähernd wiedergefunden, mit den Skizzen dagegen sehr stark verschätzt. Insbesondere erhöht sich

die Varianz der *posterior* mit ϵ. Die Einbettungsmethode CW liefert hier besonders ungünstige Ergebnisse, obwohl ihre Skizze die größte ist. Die Skalenparameter des Modells werden mit den Skizzen in eine völlig unrealistische Größenordnung geschätzt. Die Änderung einiger Vorzeichen in den Daten der Skizzen scheint hier ebenfalls eine Rolle zu spielen.

Demnach sind Einbettungen nur bei solch komplexen generalisierten linearen Modellen grundsätzlich ungeeignet. Im Übrigen sind die Abweichungen der Schätzwerte bzw. *posteriors* zwischen großem Datensatz und Skizzen kleiner als angenommen oder kaum vorhanden. Verschiedene Probleme bei der MCMC-Schätzung treten bei allen Datensätzen auf und sagen daher nichts über die Güte der Einbettung aus.

Kapitel 6

Folgerungen

Geppert *et al.* (2014) zeigen Konsequenzen der Einbettung im nicht-hierarchischen Regressionsmodell mit normalverteilten Daten. Dies wird im Abschnitt 3.2 zusammengefasst. Dort wird die *posterior* $p(\beta|y)$ bzw. der Schätzer $\hat{\beta}$ betrachtet und danach unterschieden, ob diese aus dem großen Datensatz oder einer Skizze berechnet werden. Es wird auf verschiedene Weise ein Abstand dieser Größen zwischen großem Datensatz und Skizze gemessen, der sich jeweils als durch den Approximationsparameter ϵ beschränkt erweist.

In diesem Kapitel wird versucht, aus den Ergebnissen von Geppert *et al.* (2014) Folgerungen für den Fall eines hierarchischen Regressionsmodells herzuleiten. Dabei wird auf die Simulationsstudie in Kapitel 5 Bezug genommen. Empirische Ergebnisse können Vermutungen einerseits widerlegen, andererseits stützen oder neue anregen.

Zu Beginn des Kapitels wird eine grundsätzliche Approximierbarkeit der Parameterschätzung bei Einbettung herausgearbeitet, soweit es sich um ein gewöhnliches lineares Modell mit Hyperparametern handelt. Danach werden im Besonderen die Regressionsparameter betrachtet und Folgerungen aus den Verteilungsannahmen gezogen. Ähnliches geschieht für die Schätzung der Hyperparameter. In beiden Fällen werden Situationen mit Normalverteilung und solche mit sonstigen Verteilungen unterschieden. Zuletzt wird auf Besonderheiten bei generalisierten linearen Modellen eingegangen, bevor zum Schluss des Kapitels die wichtigsten Ergebnisse zusammengefasst werden.

In Fortsetzung der Notation aus den Kapiteln 2 und 3 werden einige Abkürzungen verwendet: Der große Datensatz wird mit y bezeichnet, die Skizze mit Πy; Parameter, die aus der Skizze berechnet werden, sind in der Form ξ' gekennzeichnet; haben zwei Größen p und p' höchstens einen Abstand, der in Abhängigkeit von ϵ beschränkt

ist, wie etwa in Ungleichung (5.11), wird dies in der Form $p \overset{\epsilon}{\approx} p'$ notiert. Es ist zu beachten, dass $p \overset{\epsilon}{\approx} p'$ nicht in jedem Fall dieselbe ϵ-abhängige Schranke bedeutet.

6.1 Approximierbarkeit der Parameterschätzung

Die Simulationsstudie im Kapitel 5 legt nahe, dass sich die Schätzung der Regressionsparameter β im hierarchischen Fall ähnlich verhält wie im nicht-hierarchischen: Unterschiede zwischen großem Datensatz und Skizze sind gering und deutlich unterhalb der von Geppert *et al.* (2014) bestimmten Schranken, obwohl diese dort nur für nicht-hierarchische Modelle bewiesen werden (Tabellen 5.15 bis 5.17).

Ebenso zeigt die Schätzung der Hyperparameter θ in der Regel keine deutlichen Abweichungen zwischen großem Datensatz und Skizze. Nur im generalisierten linearen Modell ist dies anders.

Die Ähnlichkeit der Ergebnisse wird verständlich, wenn die gemeinsame Verteilung von β und θ betrachtet wird. Für ihre *posterior* gilt mit Gleichung (2.16):

$$p(\beta, \theta|y) = \frac{p(y|\beta)p(\beta|\theta)p(\theta)}{p(y)} \tag{6.1}$$

bzw.

$$p(\beta', \theta'|\Pi y) = \frac{p(\Pi y|\beta')p(\beta'|\theta')p(\theta')}{p(\Pi y)}. \tag{6.2}$$

Dabei hängt zunächst der *prior* nicht von den Daten ab:

$$p(\theta') = p(\theta) \Rightarrow p(\beta'|\theta') = p(\beta|\theta). \tag{6.3}$$

Aus den Ergebnissen von Geppert *et al.* (2014) für den nicht-hierarchischen Fall, vgl. Abschnitt 3.2:

$$p(\beta|y) \overset{\epsilon}{\approx} p(\beta'|\Pi y), \quad p(y) \overset{\epsilon}{\approx} p(\Pi y), \quad p(\beta) = p(\beta') \tag{6.4}$$

folgt weiter

$$p(y|\beta) = \frac{p(\beta|y)p(y)}{p(\beta)} \overset{\epsilon}{\approx} \frac{p(\beta'|\Pi y)p(\Pi y)}{p(\beta')} = p(\Pi y|\beta') \tag{6.5}$$

und somit insgesamt

$$p(\beta', \theta'|\Pi y) = \frac{p(\Pi y|\beta')p(\beta|\theta)p(\theta)}{p(\Pi y)} \overset{\epsilon}{\approx} \frac{p(y|\beta)p(\beta|\theta)p(\theta)}{p(y)} = p(\beta, \theta|y). \tag{6.6}$$

Die Approximation pflanzt sich so auf den hierarchischen Fall fort. Ist der Abstand zwischen den *posteriors* $p(\beta|y)$ und $p(\beta'|\Pi y)$ Abhängigkeit von ϵ beschränkt, wie

von Geppert *et al.* (2014) gezeigt wird, so auch der zwischen den *posteriors* (6.1) und (6.2).

Welches Ausmaß der Abstand quantitativ hat, bleibt dabei offen; solche Aussagen wären für die gemeinsame Verteilung von β und θ schwer zu interpretieren und hängen eventuell von den konkreten Verteilungen ab.

Aus der Approximierbarkeit bei der gemeinsamen Verteilung folgt dasselbe für die Randverteilung

$$p(\theta'|\Pi y) = \int_{\mathbb{R}^k} p(\beta', \theta'|\Pi y) d\beta' \overset{\epsilon}{\approx} \int_{\mathbb{R}^k} p(\beta, \theta|y) d\beta = p(\theta|y), \tag{6.7}$$

denn alle Information über θ ist bereits in der gemeinsamen *posterior* enthalten.

6.2 Regressionsparameter

Eine der Gleichung (6.7) entsprechende Aussage für die Randverteilungen

$$p(\beta|y) \overset{\epsilon}{\approx} p(\beta'|\Pi y) \tag{6.8}$$

ist in (6.4) bereits gegeben; Hyperparameter kommen hier nicht vor. Es scheint offensichtlich, was in Abschnitt 5.3 bereits vermutet wird: dass die Schätzung der unbekannten Verteilung von β nicht davon abhängt, ob sie von festen oder variablen (Hyper-)Parametern bestimmt ist.

Dazu passt, dass Geppert *et al.* (2014, Kap. 3) zunächst keine Verteilungsannahmen über die *posterior* $p(\beta|y)$ machen. Die *likelihood* (2.21) ist allerdings eine Normalverteilung; ebenso wird dies für den *prior* $p(\beta)$ gefordert, wenn die ϵ-Approximation bezüglich der Wasserstein-Distanz betrachtet wird (ebd., Kap. 4.2).

Die Klasse der Normalverteilungen ist ihr eigener konjugierter *prior*, wenn die Varianz der *likelihood* fest ist (Bernardo und Smith, 1994, Anh. A.2); dies ist in mehreren Modellen der Simulationsstudie der Fall. Eine lineare Transformation durch die Designmatrix X erhält die Normalverteilung (Rinne, 2008, Kap. B3.10.4). Somit ist auch die *posterior* normalverteilt, was eine Interpretation gemäß der Schätzung (2.22) erleichtert.

Im hierarchischen Modell (2.14) tritt

$$p(\beta|\theta)p(\theta) = p(\beta, \theta) \tag{6.9}$$

gewissermaßen an die Stelle des normalverteilten nicht-hierarchischen *priors* $p(\beta)$. Ist nun $p(\beta|\theta)$ wiederum normalverteilt (Modelle 1a und 1b), so ist weiterhin ein

konjugierter *prior* gegeben, der die Voraussetzung zur ϵ-Approximation bezüglich der Wasserstein-Distanz erfüllt: Wegen

$$p(y|\beta, \theta) = p(y|\beta) \tag{6.10}$$

wirkt sich nur die Verteilung von $p(\beta|\theta)$ direkt auf die Bestimmung der *posterior* $p(\beta|y)$ aus, $p(\theta)$ dagegen nicht.

Demnach gilt die Schranke (3.14) auch im hierarchischen Modell, falls der *prior* für β normalverteilt ist. Die Schätzer $\hat{\beta}$ bzw. $\hat{\beta}'$ sind nach Gleichung (2.22) Erwartungswerte von $p(\beta|y)$ bzw. $p(\beta'|\Pi y)$ und ihr Abstand ist dementsprechend beschränkt, vgl. Abschnitt 3.2.

Die ϵ-Approximierbarkeit bezüglich der Wasserstein-Distanz muss dagegen nicht gelten, wenn der *prior* für β einer anderen Verteilung folgt, wie es etwa in den Modellen 2a, 2b und 3 der Simulationsstudie der Fall ist. Es bleibt die Möglichkeit, $\hat{\beta}$ und $\hat{\beta}'$ zu betrachten, welche empirisch die Schranken aus (3.6) und (3.9) erfüllen.

Bezüglich dieser Schranken ist bei Geppert *et al.* (2014) noch nicht explizit von einer Bayesschen Regressionsanalyse die Rede. Die Schätzer sind lediglich als Erwartungswerte der *posteriors* interpretierbar. Letztere müssen nun nicht mehr normalverteilt sein; etwa bei symmetrischen Verteilungen ist es allerdings denkbar, dass sich „im Erwartungswert“ ein ähnliches Verhalten zeigt.

Allgemein stellt sich die Frage nach der tatsächlichen Bedeutung des *priors* für die Schätzung von β. Je mehr Daten gegeben sind, desto stärker überwiegt in der Regel die *likelihood* (vgl. Congdon, 2006, Kap. 1.2); das asymptotische Verhalten einer *posterior* für $n \to \infty$ kann beispielsweise bei Bernardo und Smith (1994, Kap. 5.3) nachvollzogen werden. In jedem Fall ist die *likelihood* weiterhin normalverteilt. In der Simulation haben auch die Skizzen noch eine beträchtliche Größe (Tabelle 5.1), so dass asymptotische Überlegungen in Frage kommen.

Wird der Einfluss des *priors* vernachlässigbar gering, so wirkt sich auch die Verteilungsklasse möglicher Hyperparameter nicht mehr wesentlich auf die Schätzung aus. Diese erfolgt dann nahezu klassisch und es spricht nichts mehr gegen die Gültigkeit der Schranken aus dem nicht-hierarchischen Fall. In der Simulationsstudie wird dies dadurch gestützt, dass die MCMC-simulierte *posterior* von β meist nur wenige Werte annimmt und eine sehr geringe Varianz hat; sie ähnelt in diesem Sinne einem klassischen Punktschätzer.

6.3 Hyperparameter

Ist die Schätzung der Regressionsparameter wesentlich durch die Daten und weniger durch den *prior* bestimmt, stellt sich weiterhin die Frage nach dem Verhältnis von Daten und Hyperparametern. Nach Abschnitt 6.1 hat die Schätzung der *posterior* aus einer Skizze nur einen beschränkten Abstand zu derjenigen aus dem großen Datensatz.

Für den Fall der Normalverteilung ergibt sich aus den Modellen 1a und 1b der Simulationsstudie empirisch folgendes Bild:

$$\hat{\beta} = \mathbb{E}(\mu) =: m, \quad p(\mu|y) = N(m, \mathrm{diag}(\mathbb{E}^2(\sigma))), \quad p(\beta|y) = \delta_m, \quad p(\sigma|y) = p(\sigma) \tag{6.11}$$

sowohl beim großen Datensatz als auch bei den Skizzen. Beim *intercept* ist der Abstand etwas größer als bei den *slope*-Parametern, vgl. Abschnitt 5.3. Wird im Modell 1a' zusätzlich die Varianz in der *likelihood* geschätzt, so ist deren *posterior* zu groß, im Übrigen ändert sich wenig, vgl. Abschnitt 5.4.

Die *a-posteriori*-Normalverteilung von μ erklärt sich aus der *likelihood*

$$p(y|\mu) = N(X\mathbb{E}(\mu), \mathrm{Id}_n s^2) = N(X\beta, \mathrm{Id}_n s^2) \tag{6.12}$$

für den Fall eines nichtinformativen, aber auch eines konjugierten *priors*: Aus einer Normal-Invers-Wishart-Verteilung von

$$p(\mu, \sigma) = p(\mu|\sigma)p(\sigma) \tag{6.13}$$

ergibt sich $p(\mu|\sigma)$ als Normalverteilung um m mit einer zu σ proportionalen Kovarianzmatrix (Gelman *et al.*, 2014, Kap. 3.6, dort ohne den Kontext der Regression), vgl. Anhang A.2.

Ein dementsprechendes Resultat in einem vereinfachten eindimensionalen Fall liefert Congdon (2010, Kap. 3.3) für die *full conditional posterior* von μ, die geringfügig von n, jedoch nur über $\hat{\beta} = m$ im Erwartungswert von den eigentlichen Daten abhängt, was für den großen Datensatz und die Skizzen ϵ-approximativ dasselbe ergibt. Hier ist σ wiederum in der Varianz von μ enthalten.

Als *prior* für σ wird in den Simulationen die wahre Verteilung gewählt, um zumindest für den großen Datensatz plausible Ergebnisse zu erreichen. Die *posterior* ist daraufhin in allen Fällen von den Daten unbeeinflusst, was gut mit den obigen auf σ bedingten Resultaten übereinstimmt.

Aus der *a-posteriori*-Normalverteilung von μ um den Schätzwert von β ergibt sich, dass die von Geppert *et al.* (2014) bestimmten Schranken (3.6) und (3.9) im hierar-

chischen Normalverteilungsmodell auch für den Abstand der $\mathbb{E}(\mu)$ zwischen großem Datensatz und Skizze gelten.

Ist $p(\beta|\theta)$ nicht normalverteilt, so ist ohne Weiteres keine solch konkrete Aussage möglich. Beispiele sind die Modelle 2a, 2b und 3 in der Simulationsstudie. Dort ist wiederum zu sehen, dass sich die *posterior* von θ nur geringfügig zwischen großem Datensatz und Skizze unterscheidet, wobei der Abstand teilweise erkennbar mit ϵ zunimmt.

Genauere Aussagen wären für die jeweils gegebene Verteilungsklasse im Einzelnen zu beweisen. Mit dem obigen Resultat für μ ist aber weiterhin plausibel, dass eine Funktion f der Hyperparameter mit

$$f(\theta|y) = \mathbb{E}(\beta|y,\theta) = \hat{\beta} \overset{\epsilon}{\approx} \hat{\beta}' = f(\theta'|\Pi y) \quad \forall\theta \tag{6.14}$$

gemäß (3.6) und (3.9) zwischen großem Datensatz und Skizze beschränkt bleibt.

6.4 Generalisierte lineare Modelle

Anders als in den übrigen Modellen der Simulationsstudie stimmen im generalisierten linearen Modell 4 die Ergebnisse aus großem Datensatz und Skizzen augenscheinlich nicht überein. Die Abweichungen bei den Regressionsparametern scheinen zu groß, um eine von ϵ abhängige Beschränkung anzunehmen, auch wenn dies bei den zum Teil komplexen Transformationen schwierig zu interpretieren ist. Auch im vereinfachten Modell 4' zeigt sich ähnliches.

Allgemein formuliert werden hier lineare Prädiktoren $X\beta$ durch eine *link*-Funktion g transformiert:

$$\eta = g(X\beta), \tag{6.15}$$

so dass mit einem Fehler e gilt:

$$y = \eta + e, \tag{6.16}$$

(vgl. McCulloch und Searle, 2001, Kap. 5), in diesem Fall mit normalverteilter *likelihood*

$$p(y|\beta, X) = p(y|\eta) = N(\eta, \mathrm{Cov}(e)). \tag{6.17}$$

Das im Kapitel 5 betrachtete Modell ist tatsächlich noch etwas komplexer, da es zusätzliche Parameter etwa in der Form

$$\eta = g(X\beta, \lambda) \tag{6.18}$$

enthält (vgl. Rinne, 2008, Kap. D1.2).

Weiter sei daran erinnert, dass die Aussagen nach Geppert *et al.* (2014) im Abschnitt 3.2 auf den Schätzern (3.3) und (3.4) beruhen, welche jeweils die Residualquadratsumme minimieren. Im generalisierten Modell ist dies

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \|\eta - y\|_2^2 = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \|g(X\beta) - y\|_2^2 \tag{6.19}$$

(vgl. Rinne, 2008, Kap. D1.2.1) für den großen Datensatz. Die entsprechende Formulierung für eine Skizze wäre

$$\hat{\beta}' = \underset{\beta' \in \mathbb{R}^k}{\operatorname{argmin}} \|\Pi\eta' - \Pi y\|_2^2 = \underset{\beta' \in \mathbb{R}^k}{\operatorname{argmin}} \|\Pi g(X\beta') - \Pi y\|_2^2 . \tag{6.20}$$

Ob ein solcher Ausdruck tatsächlich berechnet wird, ist allerdings zweifelhaft: Geppert *et al.* (2014, vor Lem. 5) sprechen für den nicht-hierarchischen Fall über die Projektion von Πy in den Spaltenraum von ΠX, was den Schätzer (3.4) begründet. Diese Herleitung ist im generalisierten Modell nicht möglich. Π ist eine Einbettung des Spaltenraums von (y, X), welche die Existenz einer *link*-Funktion mit der Eigenschaft (6.16) nicht berücksichtigt.

Sei (y, X) ein Datensatz, der Gleichung (6.16) erfüllt. Er werde zur Skizze $\Pi(y, X) = (\Pi y, \Pi X)$ eingebettet, wie es in der Simulationsstudie geschieht. Um die wesentliche Datenstruktur dabei zu erhalten, muss die Skizze ebenfalls Gleichung (6.16) zumindest ϵ-approximativ erfüllen:

$$\Pi y \overset{\epsilon}{\approx} g(\Pi X\beta) + e. \tag{6.21}$$

Soll ein Schätzer aus ihr berechnet werden, wird (6.19) analog angewendet mit $(\Pi y, \Pi X)$ anstelle von (y, X):

$$\hat{\beta}' = \underset{\beta' \in \mathbb{R}^k}{\operatorname{argmin}} \|g(\Pi X\beta') - \Pi y\|_2^2 . \tag{6.22}$$

In der Simulationsstudie sind solche Ergebnisse zu sehen. Von Interesse sind nun Aussagen über den Abstand zwischen den Schätzern (6.19) und (6.22).

Letzterer hat aber nicht die von Geppert *et al.* (2014) geforderte Form analog (3.4). Diese wäre vielmehr in (6.20) zu finden. Damit beides äquivalent und sinnvoll ist, müsste die Einbettung mit der *link*-Funktion kommutieren:

$$\Pi g(X\beta) = g(\Pi X\beta) \quad \forall \beta, X. \tag{6.23}$$

Dies ist im Allgemeinen nicht der Fall ist, das g nicht linear ist.

Dieses Problem wird im Kapitel 7 anhand zweier einfacher Modelle mit einem *logit-link* veranschaulicht. Dort ist zu sehen, dass die Struktur der Daten in den Skiz-

zen eine andere ist als im großen Datensatz. Sie entspricht so nicht mehr den Modellannahmen und führt dementsprechend zu falschen Schätzwerten. Eine Ausnahme ist jedoch das Verhalten der Einbettungsmethode CW, welche unter gewissen Symmetrie-Bedingungen weiterhin geeignet zu sein scheint.

Im Allgemeinen lassen sich somit keine Aussagen zu durch ϵ beschränkten Abständen aus denjenigen von Geppert *et al.* (2014) herleiten. Dies heißt nicht, dass keine Approximation möglich ist. Die besondere Struktur des Modells, charakterisiert durch die *link*-Funktion, müsste jedoch bei der Berechnung des Schätzers bzw. bei der Formulierung der Einbettung berücksichtigt werden.

Wie in den vorigen Abschnitten stellt sich auch hier die Frage nach dem Einfluss der Daten einerseits und des *priors* andererseits auf die Schätzung der verschiedenen Parametertypen. Je nach Situation und Stärke des Zusammenhangs der Parameter kann der tatsächliche Abstand zwischen Schätzwerten größer oder kleiner sein.

Vor diesem Hintergrund überrascht es nicht, wenn sich großer Datensatz und Skizzen in den Modellen 4 und 4' unterschiedlich verhalten und ihre Schätzwerte zum Teil große Abstände erreichen, wenn auch nicht bei jedem Parameter. Zusätzliche Parameter wie in der Form (6.18) können die Situation zusätzlich erschweren.

Schließlich ist in jedem konkreten Fall zu beachten, ob die Daten nach Anwendung der Einbettung noch die richtige Struktur aufweisen: Manche Modelle setzen die Zielvariable als nicht-negativ oder als ganzzahlig voraus; beides muss in der Skizze nicht mehr durchgehend erfüllt sein. Die Ergebnisse von Geppert *et al.* (2014) setzen aber ohnehin eine normalverteilte Zielvariable voraus.

6.5 Zusammenfassung

Es wird untersucht, inwieweit sich die Ergebnisse von Geppert *et al.* (2014) auf Parameter in hierarchischen Modellen übertragen lassen. Dabei lassen sich die wesentlichen empirischen Ergebnisse der Simulationsstudie grundsätzlich erklären.

Daraus, dass der Abstand der geschätzten Regressionsparameter zwischen einem großen Datensatz und seiner Skizze im nicht-hierarchischen Fall beschränkt ist, folgt eine allgemeine Aussage über die *posteriors* in hierarchischen linearen Modellen: Für Regressions- und Hyperparameter sowie für beide gemeinsam sind die Abstände ebenfalls beschränkt, in zunächst unbestimmtem Ausmaß.

Genauer ergibt sich eine Verallgemeinerung der Aussagen aus Abschnitt 3.2 über

die Verteilungen bzw. Punktschätzer der Regressionsparameter: Ist deren *prior* im hierarchischen Modell normalverteilt, sind die Abstände weiter in der von Geppert *et al.* (2014) beschriebenen Weise beschränkt. Bei anderen Verteilungen kann vermutet werden, dass das Ergebnis dem sehr nahe kommt, da die Bedeutung der *likelihood* aufgrund der großen Stichproben wesentlich überwiegt. Dann wird weniger eine *posterior* als vielmehr ein Wert erzielt, der einem klassischen Punktschätzer ähnelt.

Bei Normalverteilung der *prior* lässt sich für *location*-Hyperparameter zeigen, dass ihre Erwartungswerte *a posteriori* derselben Beschränkung unterliegen wie die von ihnen kontrollierten Regressionsparameter. Bei anderen Modellen ist dies für solche Transformationen von Hyperparametern denkbar, die den Erwartungswert eines Regressionsparameters ergeben.

Für generalisierte lineare Modelle lassen sich dagegen keine Beschränkungen der Abstände für die verschiedenen Parameter herleiten, wenn die gegebenen Regressionsschätzer verwendet werden, sofern nicht Einbettung und *link*-Funktion kommutieren.

In allen Fällen sollte beachtet werden, welchen tatsächlichen Einfluss *likelihood* und *prior* auf die Berechung der *posterior* eines Parameters ausüben. Haben die Daten nur geringe Bedeutung, so können die Schätzungen aus großem Datensatz und Skizze entsprechend ähnlicher sein. Anderenfalls kann eher die Verteilung der *prior* in einem Modell vernachlässigt werden.

Unter Berücksichtigung all dessen findet sich eine große Übereinstimmung der Überlegungen dieses Kapitels mit den Ergebnissen der Simulationsstudie, wonach sich die Parameterschätzungen aus einem großem Datensatz und seinen Skizzen in vielen Fällen sehr ähnlich verhalten.

Kapitel 7

Studie zu generalisierten linearen Modellen

Zur empirischen Untersuchung, wie sich ein generalisiertes lineares Modell bei Einbettung der Daten verhält, werden in Ergänzung zu den Überlegungen in Abschnitt 6.4 zwei weitere Modelle betrachtet. Sie sind nicht hierarchisch und ihre Parameter lassen sich klassisch schätzen.

Insofern entfernt sich dieses Kapitel vom Gegenstand dieser Arbeit. Er enthält zunächst die Beschreibung der Modelle und eine Auswertung der Ergebnisse. Anschließend wird eine Einordnung in das Gesamtresultat der Kapitel 5 und 6 versucht. Dabei ergibt sich eine zusätzliche Anwendbarkeit der Einbettungsmethode CW in gewissen Situationen, die künftig näher zu untersuchen ist.

7.1 Aufbau und Ergebnisse

In beiden Fällen wird eine stochastische Designmatrix $X \in \mathbb{R}^{n\times k}$ mit einer Einserspalte und $k-1$ Spalten mit Realisierungen von $x \sim R(-5,5)$ verwendet. Es sind wiederum $n = 10000$ und $k = 5$. Als „wahrer“ Parametervektor wird $\beta = 1 \in \mathbb{R}^k$ gewählt.

Die beiden Modelle haben die Form

$$y_{(h)} = g_{(h)}(X\beta) + e_{(h)}, \quad e_{(h)} \sim N(0, s^2_{(h)}), \quad h = 1,2 \tag{7.1}$$

und unterscheiden sich geringfügig bezüglich des verwendeten *links*:

Modell 5a: Normalverteilung mit *logit-link*

$$g_{(1)}(X\beta) = \frac{1}{1 + \exp(-X\beta)} \tag{7.2}$$

mit $s_{(1)} = 0.1$. Die Parameter lassen sich ähnlich zur Methodik der logistischen Regression schätzen, womit allerdings nicht die kanonische *link*-Funktion zur Normalverteilung verwendet wird (vgl. McCulloch und Searle, 2001, Kap. 5).

Während sich die Regressionsparameter aus dem großem Datensatz leicht und korrekt schätzen lassen, führt das übliche Berechnungsverfahren für generalisierte lineare Modelle (R Core Team, 2014, Dokumentation zu `glm`) bei keiner der Skizzen zu Konvergenz. Wird dagegen die Residualquadratsumme aus (6.19) bzw. (6.20) numerisch minimiert (R Core Team, 2014, Dokumentation zu `optim`), ergeben sich die in Tabelle 7.1 zusammengefassten Schätzwerte, nun mit Konvergenz.

Tabelle 7.1: Schätzwerte der Regressionsparameter in **Modell 5a**.

	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
β_1	1.012	3.461	3.133	22.495	3.713	2.945	3.053
β_2	1.008	0.395	0.571	2.506	0.465	0.432	0.302
β_3	1.011	0.454	0.261	2.546	0.427	0.381	0.430
β_4	1.007	0.425	0.389	2.778	0.624	-0.114	0.345
β_5	1.006	0.425	0.411	2.043	0.214	0.497	0.353

Diese lassen sich ohne Angaben zur Varianz zwar nicht vollständig bewerten, es wird aber deutlich, wie stark die Schätzwerte aus den Skizzen von den wahren Parametern abweichen, insbesondere für den *intercept*.

Dieses Ergebnis stützt die Vermutung aus Abschnitt 6.4 über die Unverträglichkeit der Einbettungsmatrix mit einer nicht-linearen *link*-Funktion: Die Transformation der Daten verändert deren Struktur derart, dass sie nicht mehr zu den Modellannahmen passen. Dies wird in Abbildung 7.1 illustriert. Insbesondere führt die Veränderung der Einserspalte in den Skizzen zu falschen Schätzwerten beim *intercept*.

Die Abbildung 7.1 zeigt jedoch auch eine Besonderheit der Einbettungsmethode CW: Hier wird die Struktur der Daten nicht völlig aufgelöst, wie bei den beiden anderen, sondern bleibt in gewisser Weise erhalten. Rund die Hälfte der Datenpunkte ist so verschoben, dass sich das Bild aus dem Bereich positiver y im negativen wiederholt. Es entsteht eine Form von Punktsymmetrie um 0. In diesen beiden

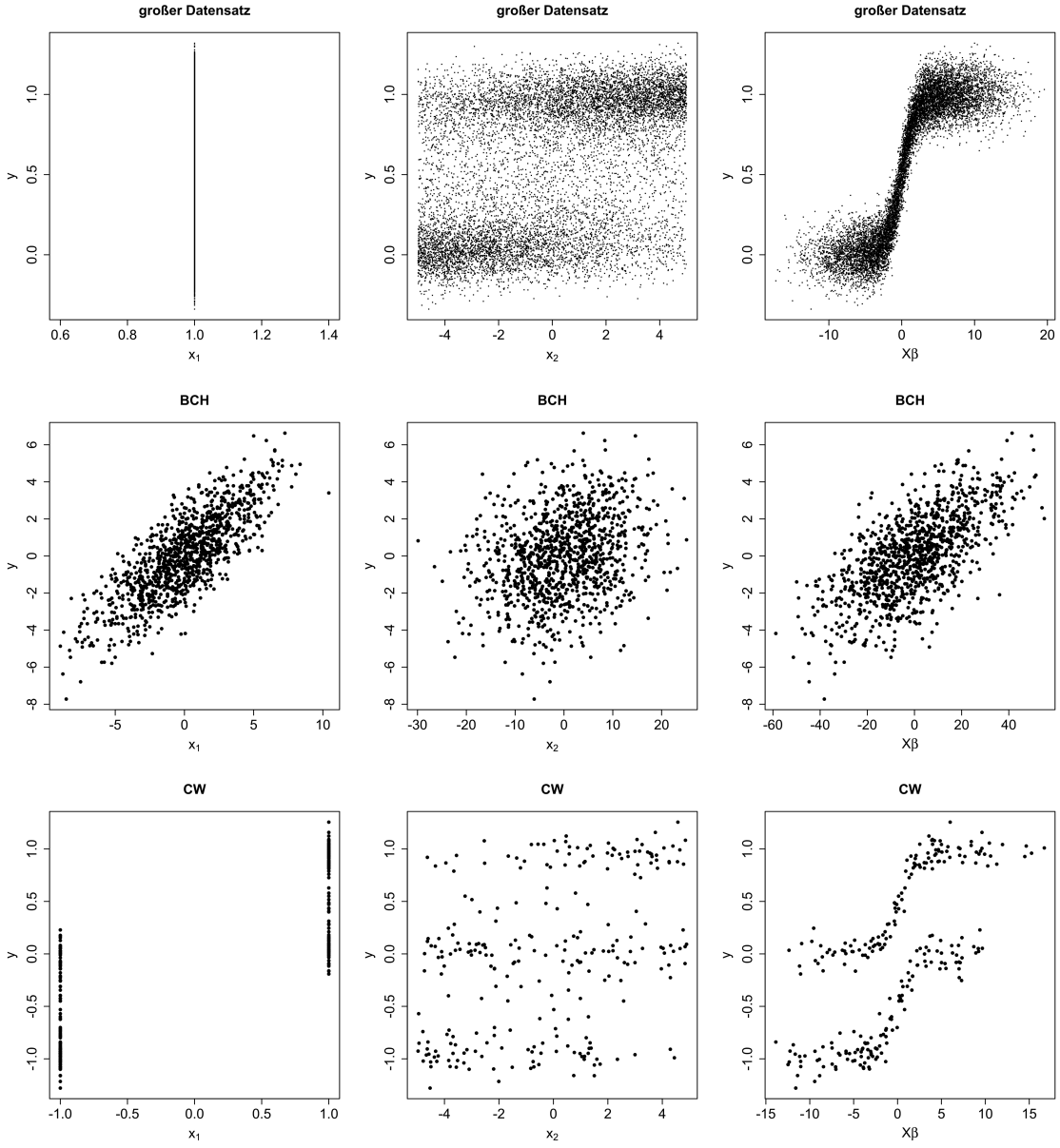

Abbildung 7.1: Die Einbettung ($\epsilon = 0.1$) verändert die Struktur der Daten in **Modell 5a**. Es ist jeweils die Zielvariable gegen die „Einserspalte“ (links), gegen eine Einflussvariable (Mitte) und gegen den gesamten linearen Prädiktor mit „wahrem“ β (rechts) dargestellt. Im ursprünglichen Datensatz (oben) ist die Form der *link*-Funktion gut zu erkennen, während die Einbettungsmethode BCH (Mitte, für SRHT ist das Bild qualitativ dasselbe) diesen Zusammenhang auflöst und bestenfalls einen linearen erhält. Die Methode CW scheint dagegen rund die Hälfte der Datenpunkte so zu verschieben, dass eine Symmetrie um 0 entsteht, und erhält für beide Hälften den Zusammenhang.

Gruppen von Datenpunkten lässt sich der Zusammenhang aus dem großen Datensatz leicht wiederfinden. Dieser Befund wird im folgenden Modell wieder aufgenommen.

Modell 5b: Normalverteilung mit „symmetrischem *logit-link*“

$$g_{(2)}(X\beta) = \frac{1 - \exp(-X\beta)}{1 + \exp(-X\beta)} \tag{7.3}$$

mit $s_{(2)} = 0.2$. Wegen

$$g_{(2)}(X\beta) = \frac{2}{1 + \exp(-X\beta)} - \frac{1 + \exp(-X\beta)}{1 + \exp(-X\beta)} = 2g_{(1)}(X\beta) - 1 \tag{7.4}$$

und

$$y_{(2)} = 2g_{(1)}(X\beta) - 1 + e_{(2)} \sim 2g_{(1)}(X\beta) - 1 + 2e_{(1)} = 2y_{(1)} - 1 \tag{7.5}$$

lässt sich die Schätzung der Parameter durch die Transformation $\tilde{y} := (y+1)/2$ auf den Fall von Modell 5a zurückführen:

$$\begin{aligned} \hat{\beta}_{(2)}(y) &= \underset{\beta}{\operatorname{argmin}} \, \|y - g_{(2)}(X\beta)\|_2^2 \\ &= \underset{\beta}{\operatorname{argmin}} \, \|2\tilde{y} - 1 - 2g_{(1)}(X\beta) + 1\|_2^2 \\ &= \underset{\beta}{\operatorname{argmin}} \, 4\|\tilde{y} - g_{(1)}(X\beta)\|_2^2 \\ &= \hat{\beta}_{(1)}(\tilde{y}). \end{aligned} \tag{7.6}$$

Die Berechnung der Schätzwerte scheint insgesamt etwas leichter und korrekter zu sein als im vorigen Modell. Die Einbettungsmethode CW führt auch mit `glm` zu Konvergenz und zeigt eine gute Näherung der Ergebnisse aus dem großen Datensatz, wie in Abbildung 7.2 veranschaulicht wird. Im Übrigen ergeben sich mit `optim` die in Tabelle 7.2 gezeigten Werte.

Tabelle 7.2: Schätzwerte der Regressionsparameter in **Modell 5b**.

	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
β_1	1.003	0.627	0.944	1.040	12.213	0.899	1.627
β_2	0.998	0.502	0.920	0.816	8.421	0.988	1.736
β_3	0.999	0.482	0.918	0.832	7.353	0.981	1.070
β_4	0.997	0.490	0.934	0.720	5.142	0.970	0.879
β_5	0.995	0.508	0.920	0.676	8.738	0.985	1.631

Die Abbildung 7.3 weist auf eine Erklärung für die besondere Leistungsfähigkeit von CW hin, die sich in Modell 5a bereits angedeutet hat. Die *link*-Funktion in diesem

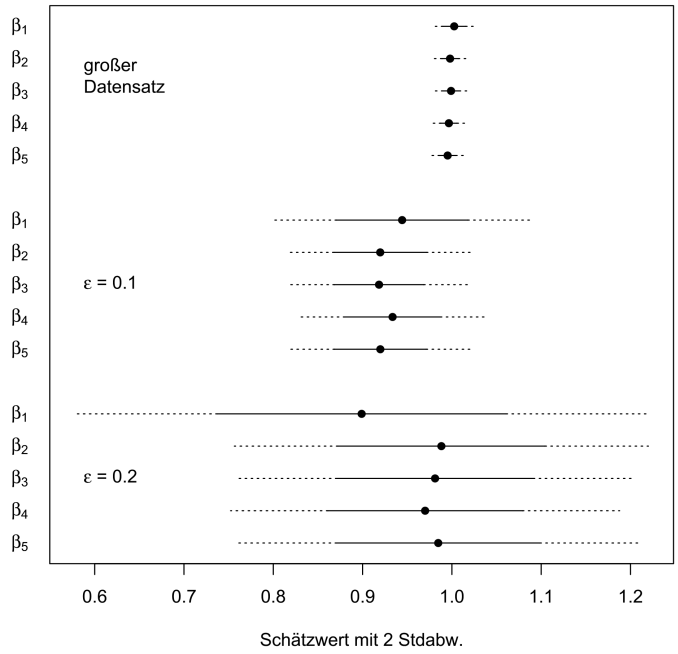

Abbildung 7.2: Parameterschätzung durch `glm` für **Modell 5b** bei Einbettungsmethode **CW**.

Modell ist ungerade, womit die Daten bereits eine punktsymmetrische Struktur um 0 haben. Wo im vorigen Modell noch ein Teil der Datenpunkte von CW verschoben wurde, bleibt das Bild nun im Wesentlichen erhalten. Da die Skizze insofern den Modellannahmen entspricht, lassen sich die Parameter korrekt schätzen.

7.2 Bewertung und Ausblick

Die Studie bestätigt einerseits die Vermutung aus Abschnitt 6.4, dass Skizzen im Allgemeinen nicht mehr die Ergebnisse des großen Datensatzes approximieren können, wenn im Modell eine nicht-lineare *link*-Funktion vorhanden ist:

Die hier untersuchten Situationen sind wesentlich einfacher als das Beispiel eines generalisierten linearen Modells aus Kapitel 5. Dort haben zusätzliche Parameter, insbesondere für die Varianz, und hierarchische Strukturen die Schätzung erschwert. Zudem sind MCMC-Verfahren notwendig gewesen, welche erst spät zu Konvergenz geführt haben.

Hier lassen sich dagegen etablierte Schätzverfahren für nicht Bayessche, generali-

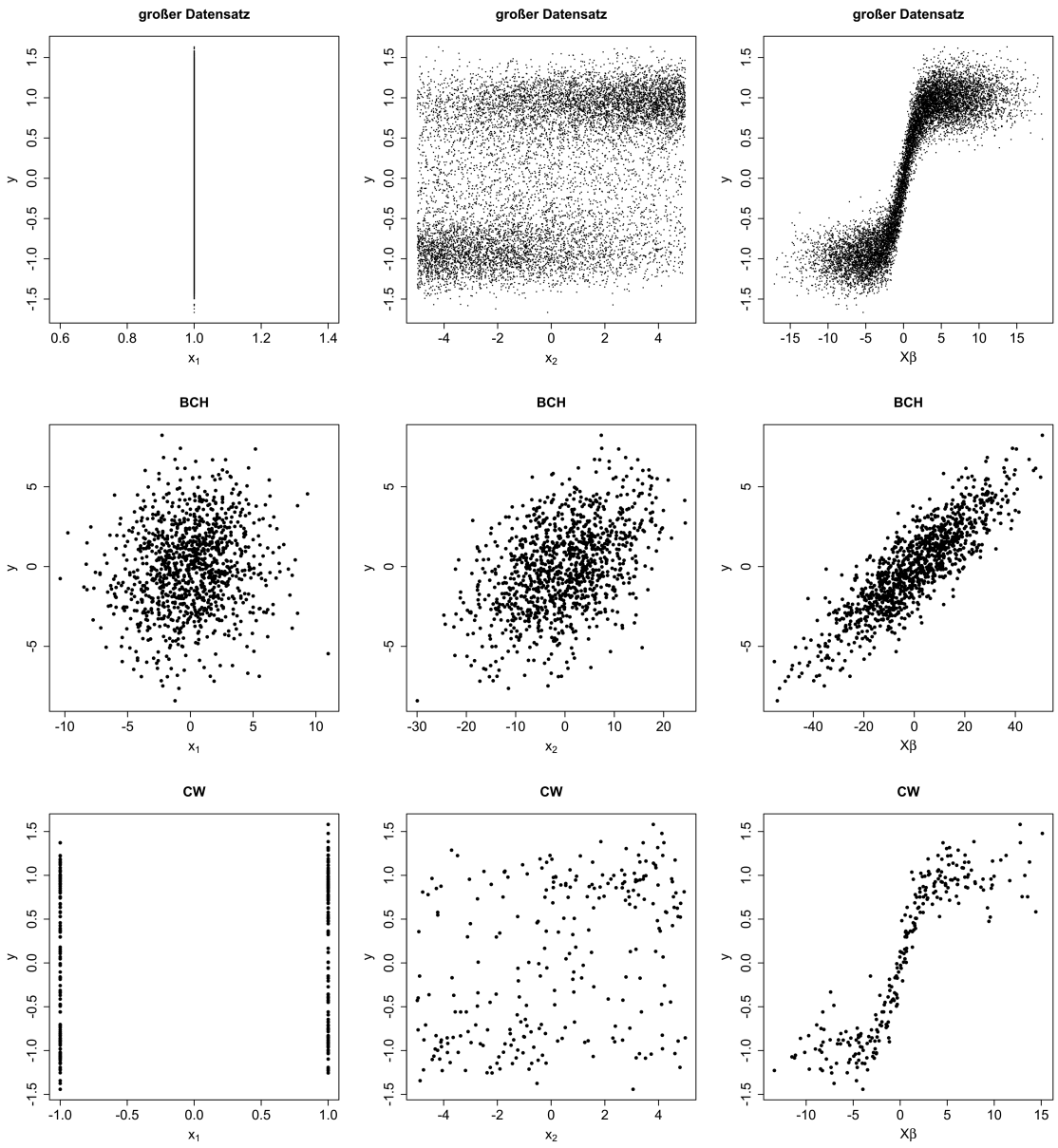

Abbildung 7.3: Daten aus **Modell 5b** analog Abbildung 7.1. Die Einbettungsmethode CW kann nun die Struktur der Daten bis auf die Einserspalte im Wesentlichen erhalten.

sierte lineare Modelle anwenden. Kommt es dabei weiter zu falschen Schätzungen, können andere Ursachen als die *link*-Funktion fast ausgeschlossen werden. In der Tat sind die durch den *link* beschriebenen Zusammenhänge in den Skizzen der Einbettungsmethoden BCH und SRHT nicht mehr zu erkennen und Fehler daher nicht überraschend.

Andererseits zeigt die Methode CW eine Besonderheit und unterscheidet sich damit erstmals in dieser Arbeit wesentlich von den beiden anderen:

Die Zusammenhänge zwischen den Variablen bleiben in gewisser Weise erhalten, es findet lediglich eine symmetrische Verschiebung etwa der Hälfte der Punkte statt. Ist der *link* sogar selbst punktsymmetrisch um 0, ist die Struktur der Skizzen dieselbe wie im großen Datensatz. Die Parameter lassen sich dann weiterhin korrekt schätzen, was gute Näherungen der Ergebnisse liefert.

In diesem besonderen Fall geht die Anwendbarkeit der Methode CW also über gewöhnliche lineare Modelle hinaus. Er erfordert damit auch eine gesonderte zukünftige Untersuchung:

Die Voraussetzung einer ungeraden *link*-Funktion ist möglicherweise weniger restriktiv als es zunächst scheint, da sich auch andere Datensätze entsprechend transformieren lassen, wie schon ab Gleichung (7.4) angedeutet ist. Es bleibt zu untersuchen, ob sich beispielsweise Verfahren für die logistische Regression und andere Modelle mit diskreter Zielvariable herleiten lassen. Eine Lösung für die logistische Regression als Klassifikationsproblem mit einer anderen Form der Skizzierung ist bei Schwiegelshohn und Sohler (2014) zu finden.

Sodann sind die Struktur der Einbettungsmatrix von CW und ihre Auswirkungen in solchen Modellen genauer zu betrachten. Insbesondere ist die Beschränkung des Fehlers in Abhängigkeit vom Approximationsparameter ϵ zu untersuchen. Dies alles führt zunächst von Bayesschen Verfahren wie auch von hierarchischen Modellen weg und hin zu einfacheren Fällen jenseits des eigentlichen Gegenstandes dieser Arbeit.

Kapitel 8

Zusammenfassung und Ausblick

Beim Umgang mit großen Datensätzen wird nach Methoden der Datenreduktion gesucht, welche die wesentlichen Informationen erhalten. Zur Parameterschätzung in linearen Regressionsmodellen und deren Bayesschen Varianten hat sich bei Geppert *et al.* (2014) die Unterraum-Einbettung als eine Möglichkeit herausgestellt: Bei Verwendung des reduzierten Datensatzes (Skizze) wird das Ergebnis aus dem großen Datensatz bis auf einen geringen, kontrollierbaren Fehler approximiert.

Darauf aufbauend wird in dieser Arbeit untersucht, ob sich die Anwendbarkeit der Methodik auf hierarchische Bayes-Modelle ausweiten lässt. Dies betrifft sowohl die Schätzung von Regressions- als auch von Hyperparametern. Die Untersuchung geschieht anhand verschiedener simulierter Beispiel-Datensätze und daran anschließender allgemeiner Überlegungen. Es werden die drei Einbettungsverfahren „Rademacher-Matrix mit BCH-Code“ (BCH), „*Subsampled Randomized Hadamard Transform*“ (SRHT) und „Clarkson-Woodruff“ (CW) zur Erzeugung der Skizzen verwendet.

Für die Regressionsparameter sind im nicht-hierarchischen Fall Schranken der Schätzfehler bekannt, welche bei Verwendung einer Skizze relativ zum großen Datensatz auftreten können. Die Approximierbarkeit ist empirisch auch in allen hierarchischen linearen Modellen erfüllt, sofern die Zielvariable weiterhin normalverteilt ist. Dies ist im Falle normalverteilter *prior* auch theoretisch begründbar, im Übrigen scheinen die Daten den *prior* ohnehin zu überwiegen.

Bei der gemeinsamen *posterior* aller Parameter lässt sich die Approximierbarkeit ebenfalls aus den bekannten Ergebnissen herleiten, es fehlt jedoch eine konkrete Schranke. Für den speziellen Fall, dass ein Hyperparameter den Erwartungswert eines Regressionsparameters beschreibt, lässt sich für ihn dieselbe Fehlerschranke

zeigen. Skalenparameter erlauben dagegen keine allgemeinen Aussagen und werden aus den Skizzen häufig zu hoch geschätzt.

Für generalisierte lineare Modelle scheint die Methodik im Allgemeinen ungeeignet. Die Einbettung wird durch eine linerare Transformation bewirkt, während die *link*-Funktion nicht-linear ist. Bei der Formulierung der Einbettungsmatrix kann der nicht-lineare Zusammenhang in den Daten nicht berücksichtigt werden. Dies wird durch Simulation eines vereinfachten, nicht-hierarchischen Modells ohne Bayes-Methoden bestätigt, um andere Ursachen der Fehlschätzung auszuschließen.

Diese Arbeit beginnt mit einer Simulationstudie, deren Ergebnisse in der Mathematik natürlich keine Beweiskraft besitzen. Vielmehr geben sie Hinweise, in welchen Fällen eine genauere Suche nach allgemeinen Aussagen vielversprechend ist und wo eine Vermutung widerlegt werden könnte. Die erforderliche theoretische Untersuchung wird hier bestenfalls skizziert. Theoretische Ergebnisse sind zwar punktuell durchaus vorhanden, sind aber noch zu vertiefen und zu ergänzen.

Einen Ansatz dazu bieten konjugierte hierarchische Bayes-Modelle, welche eine analytische Berechnung der *posterior* erlauben und so die Approximierbarkeit auch ohne Simulationen bewerten lassen. Insbesondere ließe sich die Varianz der *likelihood* in die Untersuchung einbeziehen, welche bisher nicht geschätzt, sondern festgehalten wird. Auch ergeben sich empirisch noch keine befriedigenden Aussagen für Skalen-Hyperparameter.

Einen interessanten Hinweis liefert die Simulation dagegen für das Verhältnis von *likelihood* und *prior* bei der Berechnung der *posterior* der Regressionsparameter: Auch für Modelle, die nicht den bisher theoretisch geforderten normalverteilten *prior* aufweisen, wird sie aus den Skizzen korrekt geschätzt, wobei in der Regel keine echte Verteilung, sondern vielmehr ein Punktschätzer auftritt. Da die Bedeutung des *priors* bei größer werdenden Datensätzen in der Regel abnimmt, ist die noch immer beträchtliche Größe der Skizzen zu beachten. Es stellt sich die Frage, ob die Skizzen noch groß genug sind, um den *prior* asymptotisch vernachlässigbar zu machen. Dies würde die Modellklasse, für welche die Einbettungsmethodik anwendbar ist, vergrößern.

Unabhängig vom Thema dieser Arbeit können andere Verallgemeinerungen ins Auge gefasst werden. So wird in der Zielfunktion des Regressionsproblems bisher die Euklidische Norm verwendet, andere bleiben zu untersuchen. Ein weiteres Problem stellt die Normalverteilung der Daten da, was bedeutende Modellklassen wie die logistische oder die Poisson-Regression bisher ausschließt.

Was generalisierte lineare Modelle im Allgemeinen betrifft, scheint die Einbettungsmethode CW eine Besonderheit aufzuweisen, die weitere Untersuchungen motiviert: Sofern die *link*-Funktion ungerade ist, wird die Struktur normalverteilter Daten im Wesentlichen erhalten. Damit lassen sich Parameter, zunächst für nicht-hierarchische Modelle ohne Bayes-Methoden, aus einer CW-Skizze zuverlässig schätzen. Bei anderen *link*-Funktionen kommen möglicherweise Vorab-Transformationen der Daten in Frage. Ob sich diese Form der Einbettung damit auch für Datenstrukturen mit nicht normalverteilter Zielvariable nutzbar machen lässt, ist offen.

So zeigt sich erneut, wie aus der Beantwortung gewisser Fragen stets neue entstehen, was die Erforschung desselben oder eines angrenzenden Gebiets weiter voranbringen kann.

Literaturverzeichnis

Ailon, N. und Liberty, E. (2009): Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes, *Discrete & Computational Geometry* **42**(4), 615–630.

Bernardo, J. M. und Smith, A. F. M. (1994): *Bayesian Theory*, Wiley, Chichester.

Boutsidis, C., Drineas, P. und Magdon-Ismail, M. (2013): Near-Optimal Coresets for Least-Squares Regression, *IEEE Transactions on Information Theory* **59**(10), 6880–6892.

Boutsidis, C. und Gittens, A. (2013): Improved Matrix Algorithms via the Subsampled Randomized Hadamard Transform, *SIAM Journal on Matrix Analysis and Applications* **34**(3), 1301–1340.

Bronstein, I. N., Semendjajew, K. A., Musiol, G. und Mühlig, H. (2005): *Taschenbuch der Mathematik*, 6. Aufl., Harri Deutsch, Frankfurt am Main.

Clarkson, K. L. und Woodruff, D. P. (2009): Numerical Linear Algebra in the Streaming Model, in: *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, 205–214.

Clarkson, K. L. und Woodruff, D. P. (2013): Low Rank Approximation and Regression in Input Sparsity Time, in: *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, 81–90.

Congdon, P. D. (2006): *Bayesian Statistical Modelling*, 2. Aufl., Wiley, Chichester.

Congdon, P. D. (2010): *Applied Bayesian Hierarchical Methods*, CRC, Boca Raton.

Deshpande, A., Rademacher, L., Vempala, S. und Wang, G. (2006): Matrix Approximation and Projective Clustering via Volume Sampling, in: *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1117–1126.

Draper, N. R. und Smith, H. (1981): *Applied Regression Analysis*, 2. Aufl., Wiley, New York.

Feldman, D., Monemizadeh, M., Sohler, C. und Woodruff, D. P. (2010): Coresets and Sketches for High Dimensional Subspace Approximation Problems, in: *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, 630–649.

Frieze, A., Kannan, R. und Vempala, S. (2004): Fast Monte-Carlo Algorithms for Finding Low-Rank Approximations, *Journal of the ACM* **51**(6), 1025–1041.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. und Rubin, D. B. (2014): *Bayesian Data Analysis*, 3. Aufl., CRC, Boca Raton.

Geppert, L., Ickstadt, K., Munteanu, A. und Sohler, C. (2014): Random Projections for Bayesian Regression, Technical report, SFB 876 – C4, Technische Universität Dortmund.

Hastie, T., Tibshirani, R. und Friedman, J. (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2. Aufl., Springer, New York.

Hedayat, A. und Wallis, W. D. (1978): Hadamard Matrices and Their Applications, *The Annals of Statistics* **6**(6), 1184–1238.

Jabs, V. (2012): *Vergleich von Methoden zur Dimensionsreduktion unter Berücksichtigung der Rechenzeit und des Speicherbedarfs*, Bachelorarbeit, Technische Universität Dortmund, Fakultät Statistik.

Lindstrom, M. J. und Bates, D. M. (1990): Nonlinear Mixed Effects Models for Repeated Measures Data, *Biometrics* **46**(3), 673–687.

Marjoram, P., Molitor, J., Plagnol, V. und Tavaré, S. (2003): Markov Chain Monte Carlo Without Likelihoods, *Proceedings of the National Academy of Sciences of the USA* **100**(26), 15324–15328.

McCulloch, C. E. und Searle, S. R. (2001): *Generalized, Linear and Mixed Models*, Wiley, New York.

Muthukrishnan, S. (2005): Data Streams: Algorithms and Applications, *Foundations and Trends in Theoretical Computer Science* **1**(2), 117–236.

Nelson, J. und Nguyen, H. L. (2013): OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings, in: *Proceedings of the Fifty-Fourth Annual IEEE Symposium on Foundations of Computer Science*, 117–126.

OpenBUGS Project Management Group (2012): *OpenBUGS*, Version 3.2.2 rev 1063.

Quedenfeld, J., Geppert, L., Ickstadt, K., Munteanu, A. und Sohler, C. (2014): *ls2mat: Random Projections for Bayesian Regression*, Technische Universität Dortmund, Fakultät für Informatik, Lehrstuhl für Effiziente Algorithmen und Komplexitätstheorie, R-Paket, Version 1.0, bisher unveröffentlicht.

Quiroz, M., Villani, M. und Kohn, R. (2014): Speeding up MCMC by Efficient Data Subsampling, bisher unveröffentlicht, `http://arxiv.org/abs/1404.4178`.

R Core Team (2014): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Wien, Version R 3.1.2 GUI 1.65 (6833 Snow Leopard build).

Rinne, H. (2008): *Taschenbuch der Statistik*, 4. Aufl., Harri Deutsch, Frankfurt am Main.

Rue, H., Martino, S. und Chopin, N. (2009): Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations, *Journal of the Royal Statistical Society – Series B: Statistical Methodology* **71**(2), 319–392.

Sarlós, T. (2006): Improved Approximation Algorithms for Large Matrices via Random Projections, in: *Proceedings of the Forty-Seventh Annual IEEE Symposium on Foundations of Computer Science*, 143–152.

Schwiegelshohn, C. und Sohler, C. (2014): Logistic Regression in Datastreams, Technical report, SFB 876 – C4, Technische Universität Dortmund.

Shah, R. D. und Meinshausen, N. (2013): Min-Wise Hashing for Large-Scale Regression and Classification with Sparse Data, bisher unveröffentlicht, `http://arxiv.org/abs/1308.1269`.

Stimming, C., Wagenführ, D., White, B., Baudrez, E., Gryn, J., Nakata, M., Acary, V. und Varjokallio, M. (2010): *LAPACK++*, Version 2.5.4.

Woodruff, D. P. (2014): Sketching as a Tool for Numerical Linear Algebra, *Foundations and Trends in Theoretical Computer Science* **10**(1–2), 1–157.

Tabellenverzeichnis

Anhang A

Erläuterungen

A.1 Symbolverzeichnis

Operatoren

Id_n	Einheitsmatrix der Dimension n
$\mathrm{diag}(a_1, \ldots, a_n)$	Diagonalmatrix mit Diagonalelementen $a_1, \ldots, a_n$
$(\cdot)^\top$	Transposition einer Matrix oder eines Vektors
$\|\cdot\|_2$	Euklidische Norm
$\mathbb{E}$	Erwartungswert
$\mathbb{P}(A)$	Wahrscheinlichkeit des Ereignisses A
$\hat{a}$	Schätzwert des Parametervektors a
$\propto$	proportional zu
$\overset{\epsilon}{\approx}$	Abstand abhängig vom Einbettungsparameter ϵ beschränkt

Verteilungen

$N(m, s^2)$	Normalverteilung mit Erwartungswert m und Varianz s^2
$N(m, S)$	Multivar. Norm'vert. mit E'vektor m und Kovarianzmatrix S
„N flach“	Normalverteilung mit $m = 0$ und $s^2 = 10^5$
$R(a, b)$	stetige Gleichverteilung auf $[a, b]$
$Exp(\lambda)$	Exponentialverteilung mit *rate*-Parameter λ
δ_m	Einpunktverteilung in m
$\Gamma(a, b)$	Gamma-Verteilung, vgl. Anhang A.2
$B(a, b)$	Beta-Verteilung, vgl. Anhang A.2
$Wish(S, d)$	Wishart-Verteilung, vgl. Anhang A.2
$NIW(m, l, S, d)$	Normal-Invers-Wishart-Verteilung, vgl. Anhang A.2

Bayes-Statistik

p	stetige oder diskrete Dichte einer Verteilung
$p(\cdot, \cdot)$	gemeinsame Dichte zweier Verteilungen
$p(\cdot \mid z)$	auf z bedingte Dichte
y	„Daten“: Vektor oder Matrix
ξ	Parameter-Vektor
$p(y \mid \xi)$	*likelihood*
$p(\xi)$	*prior*
$p(\xi \mid y)$	*posterior*

MCMC

$\xi^{(t)}$	Ziehung in t-ter Iteration
$\tilde{p}^{(t)}$	*proposal distribution* von $\xi^{(t)}$
ξ^*	Kandidat für $\xi^{(t+1)}$
$\alpha(\xi^* \mid \xi^{(t)})$	Akzeptanzwahrscheinlichkeit für ξ^*
$\tilde{p}^{(t)}(\cdot \mid \cdot)$	Übergangswahrscheinlichkeiten der Markov-Kette

Regression

X	Designmatrix
y	Vektor der Zielvariablen
$X_{i,\cdot}$, y_i	i-tes Datum
x_j	j-te Spalte von X
β	Vektor der Regressionsparameter
s^2	fester Varianzparameter
e	Fehlervektor
g	*link*-Funktion

Einbettung

U	großer Datensatz
ϵ	Einbettungsparameter
Π	ϵ-Einbettung(smatrix)
ΠU	Skizze von U
$d_{\min}$	kleinster echt positiver Singulärwert
δ	Wahrscheinlichkeit nicht erfolgreicher Einbettung
β', n', ...	auf Skizze bezogene Größen

Dimensionen

n	Anzahl Beobachtungen
l	Anzahl Parameter
k	Anzahl Regressionsparameter bzw. Spalten im Datensatz
n'	Anzahl „Beobachtungen“ in der Skizze

Modellparameter

β	Regressionsparameter
θ	alle Hyperparameter, darunter:
μ	Erwartungswerte
σ	Standardabweichungen
τ	Präzisionsmatrix (inverse Kovarianz)
$\alpha, \gamma, \lambda, \pi, \eta$	sonstige Hyperparameter
χ	Varianz in der *likelihood*

Sonstiges

$\mathbb{R}_+$	echt positive reelle Zahlen
$n \in O(m)$	Landau-Symbol: n wächst nicht schneller als in Größenordnung m
$n \in \Theta(m)$	Landau-Symbol: n wächst genau in Größenordnung m
$\gg, \ll$	„wesentlich“ größer / kleiner

A.2 Verwendete Verteilungen

Die Verteilungen, welche in dieser Arbeit erwähnt, aber im Allgemeinen selten verwendet werden, werden mit ihren Parametern und einigen Eigenschaften kurz vorgestellt. Dabei bezeichnet f jeweils die Dichtefunktion und $\hat{\xi}_{\text{Mom}}$ den Momentenschätzer von ξ aus arithmetischem Mittel $\bar{x}$ und Stichprobenvarianz s^2.

Gamma-Verteilung $\Gamma(a, b)$ (Rinne, 2008, Kap. B3.9.4):

shape $a \in \mathbb{R}_+$, *rate* $b \in \mathbb{R}_+$,
Träger: $\mathbb{R}_+$,

$$f(x) = \frac{1}{g_1(a)} b^a x^{a-1} e^{-bx} \tag{A.1}$$

mit der **Gamma-Funktion**

$$g_1(t) := \int_0^{+\infty} y^{t-1} e^{-y} dy \quad \forall t \in \mathbb{R}_+, \quad g_1(n) = (n-1)! \quad \forall n \in \mathbb{N}, \tag{A.2}$$

Erwartungswert $\frac{a}{b}$, Varianz $\frac{a}{b^2}$,

$$\hat{a}_{\text{Mom}} = \left(\frac{\bar{x}}{s}\right)^2, \quad \hat{b}_{\text{Mom}} = \frac{\bar{x}}{s^2}. \tag{A.3}$$

Beta-Verteilung $B(a, b)$ (Rinne, 2008, Kap. B3.11.1):

shape $a \in \mathbb{R}_+$, *shape* $b \in \mathbb{R}_+$,
Träger: $[0, 1]$,

$$f(x) = \frac{1}{g_2(a, b)} x^{a-1}(1-x)^{b-1}, \tag{A.4}$$

mit der **Beta-Funktion**

$$g_2(s, t) := \int_0^1 y^{s-1}(1-y)^{t-1} dy = \frac{g_1(s)g_1(t)}{g_1(s+t)} \quad \forall (s, t) \in \mathbb{R}_+ \times \mathbb{R}_+, \quad g_1 \text{ aus (A.2)}, \tag{A.5}$$

Erwartungswert $\frac{a}{a+b}$, Varianz $\frac{ab}{(a+b)^2(a+b+1)}$,

$$\hat{a}_{\text{Mom}} = \bar{x}\left(\frac{\bar{x}}{s^2}(1-\bar{x}) - 1\right), \quad \hat{b}_{\text{Mom}} = (1-\bar{x})\left(\frac{\bar{x}}{s^2}(1-\bar{x}) - 1\right). \tag{A.6}$$

Wishart-Verteilung $Wish(S, d)$ (Rinne, 2008, Kap. B3.10.8):

Verteilung einer $(k \times k)$-Zufallsmatrix, Verallgemeinerung der χ^2-Verteilung im $\mathbb{R}^{k \times k}$,
scale $S \in \mathbb{R}^{k \times k}$ positiv definit, $d \in (k-1, +\infty)$ Freiheitsgrade,
Träger: $\{Z \in \mathbb{R}^{k \times k} : Z \text{ pos. def.}\}$,

$$f(Z) = \frac{\det(Z)^{(d-k-1)/2} \exp\left(-\frac{1}{2}\text{tr}(S^{-1}Z)\right)}{2^{dk/2} \det(S)^{d/2} \pi^{k(k-1)/4} \prod_{j=1}^{k} g_1\left(-\frac{d+1-j}{2}\right)}, \quad g_1 \text{ aus (A.2)}, \tag{A.7}$$

Erwartungswert dS,

$$X_1, \ldots, X_n \text{ u. i. v.}, \quad X_1 \sim N(0, S) \quad \Rightarrow \sum_{i=1}^{n} X_i X_i^\top \sim Wish(S, n). \tag{A.8}$$

Normal-Invers-Wishart-Verteilung $NIW(m, l, S, d)$
(Gelman *et al.*, 2014, Kap. 3.6):

gemeinsame Verteilung eines k-Zufallsvektors und einer $(k \times k)$-Zufallsmatrix,
location $m \in \mathbb{R}^k$, Anzahl l; S und d wie bei der Wishart-Verteilung,
Träger: $\{(x, Z) \in \mathbb{R}^k \times \mathbb{R}^{k \times k} : Z \text{ pos. def.}\}$,

$$f(x, Z) \propto \det(Z)^{-(d+k)/2-1} \exp\left(-\frac{1}{2}\text{tr}(SZ^{-1}) - \frac{l}{2}(y-m)^\top Z^{-1}(y-m)\right), \tag{A.9}$$

$$Z^{-1} \sim Wish(S^{-1}, d), \quad X|Z \sim N(m, l^{-1}Z) \quad \Rightarrow (X, Z) \sim NIW(m, l, S, d). \tag{A.10}$$

Anhang B

Weitere Ergebnisse der Simulationsstudie

Fortsetzungen der Tabellen 5.3 bis 5.14 im Abschnitt 5.2 und der Tabellen 5.19 bis 5.21 im Abschnitt 5.4; zu lesen, wie am Anfang von Abschnitt 5.2 erläutert ist.

Tabelle B.1: *A-posteriori*-Verteilungen der Parameter in **Modell 0**.

$\boldsymbol{\beta_2}$							wahr: $N(2, 0.5^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.978	1.935	1.978	1.970	1.796	1.959	2.051
Stdabw.	0.002	0.002	0.003	0.002	0.002	0.004	0.002
Median	1.978	1.935	1.978	1.970	1.796	1.959	2.051

$\boldsymbol{\beta_3}$							wahr: $N(-1, 0.5^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.000	-0.976	-0.974	-0.984	-0.992	-0.986	-1.094
Stdabw.	0.002	0.002	0.003	0.002	0.002	0.004	0.002
Median	-1.001	-0.976	-0.974	-0.984	-0.992	-0.986	-1.094

$\boldsymbol{\beta_5}$							wahr: $N(1, 0.5^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.988	0.933	0.990	1.066	0.958	0.955	1.105
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	0.988	0.933	0.990	1.066	0.958	0.955	1.105

Tabelle B.2: *A-posteriori*-Verteilungen der β-Parameter in **Modell 1a**.

$\boldsymbol{\beta_3}$					wahr: $N(\mu_3, \sigma^2)$, $\mathbb{E}(\mu_3) = -1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.003	-1.077	-1.005	-0.906	-1.138	-1.008	-0.953
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	-1.003	-1.077	-1.005	-0.906	-1.138	-1.008	-0.953

$\boldsymbol{\beta_4}$					wahr: $N(\mu_4, \sigma^2)$, $\mathbb{E}(\mu_4) = -3$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.013	-3.025	-3.011	-3.002	-3.007	-3.030	-2.976
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	-3.013	-3.025	-3.011	-3.002	-3.007	-3.030	-2.976

$\boldsymbol{\beta_5}$					wahr: $N(\mu_5, \sigma^2)$, $\mathbb{E}(\mu_5) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.996	1.057	1.001	0.882	1.046	0.952	0.952
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	0.996	1.057	1.001	0.882	1.046	0.952	0.952

Tabelle B.3: *A-posteriori*-Verteilungen der μ-Parameter in **Modell 1a**.

$\boldsymbol{\mu_3}$							wahr: $N(-1, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.007	-1.081	-1.010	-0.910	-1.142	-1.012	-0.958
Stdabw.	0.203	0.203	0.203	0.203	0.203	0.203	0.203
Median	-1.006	-1.080	-1.008	-0.909	-1.141	-1.011	-0.956

$\boldsymbol{\mu_4}$							wahr: $N(-3, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.018	-3.030	-3.016	-3.007	-3.013	-3.036	-2.982
Stdabw.	0.200	0.200	0.200	0.200	0.200	0.200	0.200
Median	-3.020	-3.032	-3.018	-3.009	-3.015	-3.038	-2.983

$\boldsymbol{\mu_5}$							wahr: $N(1, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.997	1.058	1.002	0.883	1.047	0.953	0.953
Stdabw.	0.203	0.203	0.203	0.203	0.203	0.203	0.203
Median	0.998	1.059	1.004	0.884	1.049	0.954	0.954

Tabelle B.4: *A-posteriori*-Verteilungen der Parameter in **Modell 1b** (1/4).

$\boldsymbol{\beta_1}$							wahr: $N(\mu_1, \sigma_1^2)$, $\mathbb{E}(\mu_1) = 1$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.060	1.673	1.059	0.833	1.203	0.932	1.298
Stdabw.	0.013	0.013	0.020	0.013	0.014	0.030	0.013
Median	1.060	1.673	1.059	0.834	1.203	0.933	1.298

$\boldsymbol{\mu_1}$							wahr: $N(1, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.057	1.670	1.056	0.831	1.200	0.930	1.296
Stdabw.	0.203	0.203	0.204	0.203	0.203	0.205	0.203
Median	1.057	1.669	1.054	0.830	1.199	0.927	1.295

$\boldsymbol{\sigma_1}$							wahr: $N(0.2, 0.025^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.200	0.200	0.200	0.200	0.200	0.200	0.200
Stdabw.	0.025	0.025	0.025	0.025	0.025	0.025	0.025
Median	0.201	0.201	0.201	0.201	0.201	0.201	0.201

Tabelle B.5: *A-posteriori*-Verteilungen der Parameter in **Modell 1b** (2/4).

$\boldsymbol{\beta_2}$							wahr: $N(\mu_2, \sigma_2^2)$, $\mathbb{E}(\mu_2) = 2$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.992	2.039	1.977	2.050	1.799	1.948	2.111
Stdabw.	0.002	0.002	0.003	0.002	0.002	0.004	0.002
Median	1.992	2.039	1.977	2.050	1.799	1.948	2.111

$\boldsymbol{\mu_2}$							wahr: $N(2, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.982	2.029	1.968	2.040	1.789	1.938	2.102
Stdabw.	0.395	0.395	0.395	0.395	0.395	0.395	0.395
Median	1.977	2.025	1.963	2.035	1.784	1.933	2.097

$\boldsymbol{\sigma_2}$							wahr: $N(0.4, 0.025^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.400	0.400	0.400	0.400	0.400	0.400	0.400
Stdabw.	0.025	0.025	0.025	0.025	0.025	0.025	0.025
Median	0.400	0.400	0.400	0.400	0.400	0.400	0.400

Tabelle B.6: *A-posteriori*-Verteilungen der Parameter in **Modell 1b** (3/4).

$\boldsymbol{\beta_4}$							wahr: $N(\mu_4, \sigma_4^2)$, $\mathbb{E}(\mu_4) = -3$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.015	-3.034	-3.034	-3.038	-3.010	-3.069	-3.054
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	-3.015	-3.034	-3.034	-3.039	-3.010	-3.069	-3.054

$\boldsymbol{\mu_4}$							wahr: $N(-3, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.014	-3.033	-3.034	-3.038	-3.009	-3.068	-3.053
Stdabw.	0.201	0.201	0.201	0.201	0.201	0.201	0.201
Median	-3.014	-3.033	-3.034	-3.038	-3.010	-3.069	-3.054

$\boldsymbol{\sigma_4}$							wahr: $N(0.2, 0.01^2)$
prior: wahr							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.200	0.200	0.200	0.200	0.200	0.200	0.200
Stdabw.	0.010	0.010	0.010	0.010	0.010	0.010	0.010
Median	0.200	0.200	0.200	0.200	0.200	0.200	0.200

Tabelle B.7: *A-posteriori*-Verteilungen der Parameter in **Modell 1b** (4/4).

$\boldsymbol{\beta_5}$					wahr: $N(\mu_5, \sigma_5^2)$, $\mathbb{E}(\mu_5) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.996	1.023	1.004	1.023	1.017	0.991	1.140
Stdabw.	0.002	0.002	0.003	0.002	0.002	0.004	0.002
Median	0.996	1.023	1.004	1.023	1.017	0.991	1.140

$\boldsymbol{\mu_5}$					wahr: $N(1, 0.3^2)$		
prior: N flach					*posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.997	1.024	1.005	1.024	1.018	0.992	1.141
Stdabw.	0.417	0.417	0.417	0.417	0.417	0.417	0.417
Median	0.997	1.025	1.006	1.025	1.019	0.993	1.142

$\boldsymbol{\sigma_5}$					wahr: $N(0.4, 0.1^2)$		
prior: wahr					*posterior*: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.402	0.402	0.402	0.402	0.402	0.402	0.402
Stdabw.	0.101	0.101	0.101	0.101	0.101	0.101	0.101
Median	0.403	0.403	0.403	0.403	0.403	0.403	0.403

Tabelle B.8: *A-posteriori*-Verteilungen der Parameter in **Modell 2a** (1/2).

$\boldsymbol{\beta_1}$							wahr: $\Gamma(\alpha_1, \gamma)$, $\mathbb{E}(\beta_1) = 1$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.451	1.361	0.782	0.413	1.622	0.357	0.240
Stdabw.	0.007	0.008	0.003	0.015	0.004	0.008	0.004
Median	1.448	1.365	0.782	0.410	1.625	0.352	0.239

$\boldsymbol{\alpha_1}$							wahr: $N(2, 0.3^2)$
prior: $R(0.1, 20)$							*posterior*: $\Gamma(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	5.280	4.923	3.939	3.303	5.350	3.379	2.971
$\hat{b}$	1.256	1.229	1.451	1.803	1.172	1.956	2.126
Mittel	4.205	4.005	2.715	1.832	4.566	1.728	1.398
Stdabw.	1.830	1.805	1.368	1.008	1.974	0.940	0.811
Median	4.033	3.800	2.540	1.671	4.414	1.586	1.266

$\boldsymbol{\beta_2}$							wahr: $\Gamma(\alpha_2, \gamma)$, $\mathbb{E}(\beta_2) = 2$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.020	2.005	2.026	1.962	1.781	2.030	1.681
Stdabw.	0.003	0.009	0.005	0.001	0.006	0.005	0.002
Median	2.021	2.009	2.025	1.961	1.785	2.031	1.681

$\boldsymbol{\alpha_2}$							wahr: $N(4, 0.3^2)$
prior: $R(0.1, 20)$							*posterior*: $\Gamma(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	6.317	6.241	6.430	6.091	5.753	6.153	5.660
$\hat{b}$	1.154	1.151	1.171	1.130	1.166	1.122	1.201
Mittel	5.474	5.421	5.490	5.390	4.934	5.482	4.713
Stdabw.	2.178	2.170	2.165	2.184	2.057	2.210	1.981
Median	5.312	5.229	5.312	5.199	4.712	5.277	4.524

Tabelle B.9: *A-posteriori*-Verteilungen der Parameter in **Modell 2a** (2/2).

$\boldsymbol{\beta_3}$					wahr: $\Gamma(\alpha_3, \gamma)$, $\mathbb{E}(\beta_3) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.097	1.083	1.048	0.995	1.169	0.926	1.037
Stdabw.	0.002	0.003	0.005	0.001	0.002	0.004	0.001
Median	1.096	1.085	1.046	0.995	1.170	0.925	1.037

$\boldsymbol{\alpha_3}$					wahr: $N(2, 0.3^2)$		
prior: $R(0.1, 20)$					*posterior*: $\Gamma(a, b)$		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	4.604	4.567	4.347	4.404	4.759	4.262	4.490
$\hat{b}$	1.347	1.357	1.320	1.388	1.351	1.417	1.377
Mittel	3.418	3.366	3.292	3.173	3.523	3.008	3.261
Stdabw.	1.593	1.575	1.579	1.512	1.615	1.457	1.539
Median	3.264	3.189	3.114	2.985	3.356	2.849	3.096

$\boldsymbol{\beta_4}$					wahr: $\Gamma(\alpha_4, \gamma)$, $\mathbb{E}(\beta_4) = 3$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	2.984	2.968	3.017	2.940	3.014	3.064	2.756
Stdabw.	0.004	0.004	0.004	0.002	0.004	0.007	0.002
Median	2.982	2.967	3.017	2.939	3.011	3.064	2.756

$\boldsymbol{\alpha_4}$					wahr: $N(6, 0.3^2)$		
prior: $R(0.1, 20)$					*posterior*: $\Gamma(a, b)$		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	8.085	7.843	7.952	7.966	8.347	8.053	7.595
$\hat{b}$	1.060	1.043	1.045	1.053	1.097	1.041	1.074
Mittel	7.626	7.522	7.611	7.564	7.610	7.736	7.069
Stdabw.	2.682	2.686	2.699	2.680	2.634	2.726	2.565
Median	7.441	7.286	7.406	7.383	7.447	7.527	6.876

Tabelle B.10: *A-posteriori*-Verteilungen der Parameter in **Modell 2b** (1/2).

$\boldsymbol{\beta_1}$					wahr: $Exp(\lambda_1)$, $\mathbb{E}(\beta_1) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.472	0.914	1.360	1.914	4.977	0.480	0.000
Stdabw.	0.013	0.013	0.015	0.013	0.014	0.029	0.000
Median	1.472	0.914	1.360	1.914	4.977	0.480	0.000

$\boldsymbol{\lambda_1}$					wahr: $N(1, 0.05^2)$		
prior: $R(10^{-5}, 10^5)$					*posterior*: $\Gamma(a, b)$		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	2.006	2.000	1.950	1.944	2.016	1.995	3.165
$\hat{b}$	1.479	0.911	1.334	1.852	5.053	0.478	0.000
Mittel	1.357	2.196	1.462	1.050	0.399	4.177	50311.792
Stdabw.	0.958	1.553	1.047	0.753	0.281	2.957	28280.651
Median	1.135	1.831	1.221	0.879	0.338	3.498	50195.000

$\boldsymbol{\beta_3}$					wahr: $Exp(\lambda_3)$, $\mathbb{E}(\beta_3) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.990	1.058	0.985	0.927	0.816	1.133	1.237
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.001
Median	0.990	1.058	0.985	0.927	0.816	1.133	1.237

$\boldsymbol{\lambda_3}$					wahr: $N(1, 0.05^2)$		
prior: $R(10^{-5}, 10^5)$					*posterior*: $\Gamma(a, b)$		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	2.011	2.000	1.990	1.988	2.022	1.990	1.948
$\hat{b}$	0.995	1.054	0.968	0.928	0.823	1.125	1.201
Mittel	2.021	1.898	2.057	2.143	2.457	1.769	1.622
Stdabw.	1.425	1.342	1.458	1.520	1.728	1.254	1.162
Median	1.697	1.599	1.724	1.796	2.066	1.481	1.351

Tabelle B.11: *A-posteriori*-Verteilungen der Parameter in **Modell 2b** (2/2).

$\boldsymbol{\beta_4}$					wahr: $Exp(\lambda_4)$, $\mathbb{E}(\beta_4) = 3$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	3.136	3.045	3.152	3.096	2.858	3.239	2.858
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	3.136	3.045	3.152	3.096	2.858	3.239	2.858

$\boldsymbol{\lambda_4}$					wahr: $N(0.\bar{3}, 0.05^2)$		
prior: $R(10^{-5}, 10^5)$					*posterior*: $\Gamma(a, b)$		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	1.931	2.001	2.002	1.972	2.042	2.001	1.964
$\hat{b}$	3.040	3.055	3.172	3.053	2.904	3.222	2.759
Mittel	0.635	0.655	0.631	0.646	0.703	0.621	0.712
Stdabw.	0.457	0.463	0.446	0.460	0.492	0.439	0.508
Median	0.532	0.549	0.532	0.541	0.585	0.521	0.596

$\boldsymbol{\beta_5}$					wahr: $Exp(\lambda_5)$, $\mathbb{E}(\beta_5) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.106	1.152	1.074	1.089	1.354	0.989	1.301
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.001
Median	1.106	1.152	1.074	1.089	1.354	0.989	1.301

$\boldsymbol{\lambda_5}$					wahr: $N(1, 0.05^2)$		
prior: $R(10^{-5}, 10^5)$					*posterior*: $\Gamma(a, b)$		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	1.995	1.971	2.004	2.007	1.935	1.953	2.024
$\hat{b}$	1.098	1.139	1.077	1.095	1.307	0.966	1.318
Mittel	1.818	1.731	1.860	1.833	1.480	2.022	1.535
Stdabw.	1.287	1.233	1.314	1.294	1.064	1.447	1.079
Median	1.528	1.450	1.558	1.522	1.237	1.701	1.298

Tabelle B.12: *A-posteriori*-Verteilungen der Parameter in **Modell 3**.

$\boldsymbol{\beta_1}$							wahr: $N(\mu_1, 0.3^2)$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.166	0.018	0.255	0.306	0.902	0.260	1.150
Stdabw.	0.013	0.013	0.015	0.013	0.013	0.029	0.014
Median	0.166	0.018	0.255	0.306	0.902	0.260	1.150

$\boldsymbol{\beta_2}$							wahr: $N(\mu_2, 0.3^2)$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	1.469	1.327	1.473	1.472	1.258	1.481	1.550
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	1.469	1.327	1.473	1.472	1.258	1.481	1.550

$\boldsymbol{\beta_4}$							wahr: $N(\mu_4, 0.3^2)$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.513	-3.575	-3.511	-3.485	-3.520	-3.495	-3.613
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	-3.513	-3.575	-3.511	-3.485	-3.520	-3.495	-3.613

$\boldsymbol{\beta_5}$							wahr: $N(\mu_5, 0.3^2)$ *posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.465	0.452	0.466	0.457	0.522	0.477	0.396
Stdabw.	0.002	0.002	0.002	0.002	0.002	0.004	0.002
Median	0.465	0.452	0.466	0.457	0.522	0.477	0.396

Tabelle B.13: *A-posteriori*-Verteilungen der τ-Parameter in **Modell 4**.

$\boldsymbol{\tau_{22}}$							wahr: 40
prior: $\tau \sim Wish(10 \cdot \mathrm{Id}_3, 3)$							*posterior*: $\Gamma(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	3.478	2.996	2.994	2.797	2.300	2.431	2.342
$\hat{b}$	0.064	0.058	0.056	0.058	0.057	0.066	0.058
Mittel	54.677	51.314	53.560	47.944	40.064	37.066	40.108
Stdabw.	29.320	29.645	30.952	28.667	26.416	23.772	26.209
Median	49.445	45.600	47.265	42.045	34.050	31.595	34.055

$\boldsymbol{\tau_{23}}$							wahr: -7
prior: $\tau \sim Wish(10 \cdot \mathrm{Id}_3, 3)$							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	6.817	-4.127	0.807	-4.459	-5.000	-12.250	-5.203
Stdabw.	13.907	21.888	19.219	21.569	19.473	19.346	19.745
Median	5.819	-4.034	0.368	-3.891	-4.345	-11.115	-4.854

$\boldsymbol{\tau_{33}}$							wahr: 48
prior: $\tau \sim Wish(10 \cdot \mathrm{Id}_3, 3)$							*posterior*: $\Gamma(a, b)$
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
$\hat{a}$	3.444	3.441	3.138	3.367	2.985	3.244	3.059
$\hat{b}$	0.144	0.056	0.083	0.056	0.059	0.059	0.057
Mittel	23.967	61.302	38.008	60.154	50.991	54.849	53.489
Stdabw.	12.914	33.049	21.456	32.783	29.512	30.454	30.585
Median	21.690	55.490	34.160	54.305	45.135	49.155	47.465

Tabelle B.14: *A-posteriori*-Verteilungen der β-Parameter in **Modell 1a'**.

$\boldsymbol{\beta_3}$					wahr: $N(\mu_3, \sigma^2)$, $\mathbb{E}(\mu_3) = -1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.004	-1.023	-0.998	-0.989	-1.068	-0.995	-1.080
Stdabw.	0.009	0.011	0.010	0.011	0.011	0.015	0.011
Median	-1.004	-1.023	-0.998	-0.989	-1.068	-0.995	-1.080

$\boldsymbol{\beta_4}$					wahr: $N(\mu_4, \sigma^2)$, $\mathbb{E}(\mu_4) = -3$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.015	-3.003	-3.014	-3.010	-2.918	-3.016	-3.060
Stdabw.	0.008	0.010	0.010	0.010	0.009	0.014	0.010
Median	-3.015	-3.003	-3.014	-3.010	-2.918	-3.016	-3.060

$\boldsymbol{\beta_5}$					wahr: $N(\mu_5, \sigma^2)$, $\mathbb{E}(\mu_5) = 1$		
					posterior: N		
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.998	1.047	0.977	1.018	0.994	0.988	1.053
Stdabw.	0.008	0.011	0.010	0.010	0.010	0.014	0.010
Median	0.998	1.047	0.977	1.018	0.994	0.988	1.052

Tabelle B.15: *A-posteriori*-Verteilungen der μ-Parameter in **Modell 1a'**.

$\boldsymbol{\mu_3}$							wahr: $N(-1, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-1.006	-1.025	-1.000	-0.990	-1.070	-0.996	-1.081
Stdabw.	0.200	0.200	0.200	0.200	0.200	0.200	0.200
Median	-1.006	-1.025	-1.000	-0.990	-1.070	-0.996	-1.081

$\boldsymbol{\mu_4}$							wahr: $N(-3, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	-3.013	-3.001	-3.011	-3.007	-2.916	-3.014	-3.057
Stdabw.	0.202	0.202	0.202	0.202	0.202	0.203	0.202
Median	-3.014	-3.002	-3.013	-3.008	-2.917	-3.015	-3.059

$\boldsymbol{\mu_5}$							wahr: $N(1, 0.3^2)$
prior: N flach							*posterior*: N
	großer	$\epsilon = 0.1$			$\epsilon = 0.2$		
	Datensatz	BCH	CW	SRHT	BCH	CW	SRHT
Mittel	0.999	1.048	0.978	1.019	0.995	0.989	1.054
Stdabw.	0.202	0.202	0.202	0.202	0.202	0.202	0.202
Median	0.998	1.047	0.977	1.018	0.994	0.988	1.052

Eidesstattliche Versicherung

Name, Vorname:	Rathjens, Jonathan
Matr.-Nr.:	139791

Ich versichere hiermit an Eides statt, dass ich die vorliegende Masterarbeit mit dem Titel

„Hierarchische Bayes-Regression bei Einbettung großer Datensätze“

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

(Ort, Datum)

Belehrung:

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50.000,00 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden. (§ 63 Abs. 5 Hochschulgesetz – HG –)

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird gfls. elektronische Vergleichswerkzeuge (wie z.B. die Software „turnitin“) zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen:

(Ort, Datum)