TU Dortmund
Fakultät Statistik

Masterarbeit

Untersuchung von Regression auf eingebetteten Datensätzen unter Verwendung von verschiedenen Abstandsnormen und Penalisierungstermen

vorgelegt von
Steffen Müller

am 15. März 2016

Betreuung:
Prof. Dr. Katja Ickstadt

Inhaltsverzeichnis

1 Einleitung

Die Bedeutung und die Größe von zu analysierenden Datenmengen nehmen in vielen Anwendungsbereichen stetig zu. Die Verarbeitung von riesigen Datenmengen stellt dabei eine große Herausforderung für die Bereiche der Informatik und der Statistik dar. Eine statistische Analyse großer Datenmengen ist zum einen zeitlich aufwendig und zum anderen kostenintensiv. Zudem sind auch die Kapazitäten in der Computerverarbeitung begrenzt. Es werden von daher immer neue und leistungsfähigere Methoden gesucht, um große Datenmengen mit möglichst wenig Informationsverlust zu reduzieren, um diese effizienter analysieren zu können. Gerade in einem der größten Anwendungsgebiete der Statistik, der Regressionsanalyse, können solche Methoden von Vorteil sein.
In dieser Arbeit werden Bayes-Regressionsmodelle betrachtet, deren Vorteil es ist, dass Vorinformationen über Parameter im Modell, bereits mit eingebaut werden können. Ein entscheidender Nachteil hingegen ist eine teils sehr hohe Laufzeit, die durch die in der Bayes-Regression zugrundeliegenden Verfahren entstehen. Gerade bei sehr großen Datenmengen kann die Anwendung von Bayes-Regression zu hohen Laufzeiten und zu einem hohen Speicherverbrauch führen.
Aufgrund dieser Tatsache wird in dieser Arbeit die Anwendbarkeit von sogenannten Unterraumeinbettungen bezüglich der Bayes-Regression untersucht. Bei einer Unterraumeinbettung wird mittels zufälliger Projektionen aus einem großen Datensatz ein kleinerer Datensatz erzeugt, so dass dieser deutlich reduzierte Datensatz, approximativ die gleiche Information bezüglich der Regression besitzt.
Hauptteil und Ziel dieser Arbeit ist zum einen die Untersuchung einer bestimmten Implementierung einer Unterraumeinbettung und deren Anwendbarkeit bei der Bayes-Regression, sowie zum anderen die Untersuchung einer Modifizierung dieser Unterraumeinbettung, um die Gültigkeit der Methode auch für verallgemeinerte Verteilungen zu gewährleisten. Hierfür werden in dieser Arbeit zahlreiche Simulationen durchgeführt und die daraus resultierenden Erkenntnisse erläutert.
Im zweiten Kapitel wird die vorliegende Problemstellung vertiefend beschrieben. Kapitel 3 wird alle in dieser Arbeit verwendeten statistischen Methoden bezüglich der Regression beschreiben. In Kapitel 4 werden die Methoden der ε-Unterraumeinbettungen erläutert, sowie eine Modifizierung dieser. Kapitel 5 umfasst die wichtigsten Resultate aller durchgeführten Simulationen. Abschließend werden in Kapitel 6 die zentralen Ergebnisse zusammengefasst.

2 Problemstellung

Die dieser Arbeit zugrunde liegende Ausgangsposition, sowie die sich ergebenen Problemstellungen werden in diesem Kapitel genauer beschrieben.
Diese Arbeit wird im Wesentlichen auf den Ergebnissen der Arbeit von Geppert et al. (2015) aufbauen, in der die Anwendbarkeit von Unterraumeinbettungen für die Bayes-Regression unter bestimmten Annahmen nachgewiesen wird. Diese Arbeit wird eine Erweiterung der Ergebnisse für allgemeinere Annahmen darstellen.
Wie in der Einleitung bereits erwähnt, werden in dieser Arbeit große Datensätze mit Hilfe von Unterraumeinbettungen in kleinere transformiert, um zu überprüfen, ob diese eingebetteten Datensätze eine gute Approximation des Originaldatensatzes liefern. Genauer formuliert handelt es sich bei der untersuchten Methode um die ε-Unterraumeinbettung nach Clarkson und Woodruff (2013), welche im R-Softwarepaket `RaProR` von Geppert et al. (2015) implementiert ist. Die Anwendbarkeit bezüglich der Bayes-Regression unter Verwendung von uninformativem A-priori-Verteilungen bezüglich der ℓ_2-Abstandsnorm wurde bereits in der Arbeit von Geppert et al. (2015) nachgewiesen. Eine Erweiterung dieser Untersuchung, unter Verwendung von verschiedenen A-priori-Verteilungen, wird den ersten Teil der vorliegenden Arbeit umfassen. Hierfür werden zahlreiche Simulationen untersucht und vorgestellt.
Zudem wird im zweiten Teil der Untersuchung eine Modifizierung mit exponentialverteilten Zufallsvariablen nach Woodruff und Zhang (2013) durchgeführt, womit es möglich ist die Unterraumeinbettungs-Methode nach Clarkson und Woodruff (2013) auch bezüglich ℓ_p-Abstandsnormen für $p \in [1, \infty)$ anwendbar zu machen. Die Untersuchung der Anwendbarkeit dieser Modifizierung ist ein weiterer Bestandteil dieser Arbeit. Hierfür werden simulierte Regressionsdatensätze verwendet, welche zu Beginn von Kapitel 5.1 genauer beschrieben werden. Die Regressionsdatensätze werden den Fall $n \gg d$ umfassen, also dass es wesentlich mehr Beobachtungen n als Einflussvariablen d gibt. Für die simulierten Datensätze, werden zunächst gutmütige Annahmen getroffen. Weitere Untersuchungen werden auch die Fälle mittels ungünstigerer Modellannahmen umfassen. Des Weiteren werden verschiedene Größen der eingebetteten Datensätze untersucht. Eine hohe Anzahl an eingebetteten Daten zieht eine bessere Anpassung und damit eine bessere Approximation der Regressionskoeffizienten mit sich. Dies führt aber auch zu einer längeren Laufzeit der Verfahren. Insgesamt werden zahlreiche Simulationen durchgeführt, welche diverse Modellannahmen und Einstellungen der Unterraumeinbettungen abdecken, um damit die Anwendbarkeit nachzuweisen. Als nächstes werden die dafür

benötigten statistischen Methoden erläutert.

3 Regressionsmethoden

Ziel dieser Arbeit ist es, unter Verwendung verschiedener Abstandsnormen, die Anwendbarkeit von ε-Unterraumeinbettungen bezüglich Bayes-Regression zu untersuchen. Dies geschieht in Kapitel 5 mit Hilfe zahlreicher Simulationen.
Alle dafür benötigten statistischen Regressionsmethoden werden in diesem Kapitel detailiert erläutert.
Zunächst wird eine kurze Einführung in die lineare Regression, sowie die ℓ_p-Regression (3.1) gegeben. Um die Bayes-Regression zu erläutern wird zunächst eine Einführung in die Bayes-Statistik (3.2) allgemein gegeben. Anschließend wird die Idee der Bayes-Regression (3.3) erläutert, sowie die dabei relevanten Markov-Chain-Monte-Carlo-Methoden. Des Weiteren wird die p-verallgemeinerte Normalverteilung nach Subbotin (3.4) eingeführt, welche in dieser Arbeit zur Modellierung der Parameter im Bayes-Modell verwendet wird.

3.1 Lineare Regression und ℓ_p-Regression

Eines der größten Anwendungsgebiete der Statistik ist die Regressionsanalyse. Bei der Regression werden stochastische Zusammenhänge zwischen mindestens zwei Variablen modelliert, um die Eigenschaften einer Zielvariablen durch eine Funktion von Einflussgrößen darzustellen. In dieser Arbeit wird ein einfaches multivariates lineares Regressionsmodell ohne Interzept mit n Beobachtungen und d Einflussvariablen verwendet, welches mit

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{id} + \xi_i, \quad i = 1, \ldots, n$$

(Teschl, Teschl, 2007) dargestellt wird. In Matrizenschreibweise ist dieses Modell mit

$$Y = X\beta + \xi$$

gegeben. Hierbei ist $Y \in \mathbb{R}^n$ der Vektor der Zielvariable und $X \in \mathbb{R}^{n \times d}$ die Designmatrix, welche alle d Einflussgrößen umfasst. Der Wert ξ ist ein nicht beobachtbarer n-dimensionaler Fehlerterm mit Erwartungswertvektor $E(\xi) = 0$ und unbekannter

Kovarianzmatrix $\sigma^2 I_n$, mit I_n der n-dimensionalen Einheitsmatrix. Ziel ist es, den unbekannten Parametervektor $\beta \in \mathbb{R}^d$ zu schätzen und damit die lineare Abhängigkeit der Zielvariable Y in Bezug auf die d Einflussgrößen zu modellieren. Üblich ist diese Schätzung, so dass sich der mittlere quadratische Fehler minimiert. Dafür wird eine Funktion gesucht, für die sich Summe der quadrierten Abstände

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} |\beta_1 x_{i1} + \ldots + \beta_d x_{id} - y_i|^2 = \text{argmin}_{\beta \in \mathbb{R}^d} \|X\beta - Y\|_2^2$$

(Teschl, Teschl, 2007) zwischen dem Schätzwert und dem wahren Wert der Zielvariable minimiert. Dies entspricht der Gauß'schen Methode der Kleinsten-Quadrate. Dabei entspricht $\|.\|_2^2$ der quadrierten euklidischen Norm. Die p-Norm eines Vektors $v \in \mathbb{R}^n$ sei wie folgt definiert:

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}},$$

(Dasgupta et al., 2008) für einen beliebigen Vektor $v \in \mathbb{R}^n$. Die ℓ_p-Regression kann als Erweiterung der Kleinsten-Quadrate-Methode verstanden werden. Im Fall des soeben beschriebenen linearen Modells, soll der unbekannte Parametervektor bezüglich der p-Norm optimiert werden. Für die Definition der ℓ_p-Regression ergibt sich, dass für $p \in [1, \infty)$ der ℓ_p-Schätzer von $\hat{\beta}$ durch Minimierung von $\beta \in \mathbb{R}^d$ zum ℓ_p-Abstand durch die Formel

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^d} \sum_{i=1}^{n} |\beta_1 x_{i1} + \ldots + \beta_d x_{id} - y_i|^p = \text{argmin}_{\beta \in \mathbb{R}^d} \|X\beta - Y\|_p^p$$

(Dasgupta et al., 2008) gegeben ist. Für die Spezialfälle $p = 2$ ergibt sich der bereits erwähnte Fall des Kleinste-Quadrate-Schätzer und im Fall $p = 1$ der Schätzer der kleinsten absoluten Abweichung. Die ℓ_1-Regression ist somit deutlich robuster hinsichtlich Ausreißern als die ℓ_2-Regression (Woodruff, Zhang, 2013). Dies wird in Kapitel 5.3 beim Vergleich verschiedener ℓ_p-Regressionen veranschaulicht.
Ein allgemeine Penalisierung der Regressionsparameter lässt sich im frequentistischen Fall wie folgt

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^d} \|X\beta - Y\|_p^p + \lambda \|\beta\|_q^q,$$

(Hastie et al., 2009) mit Penalisierungsparameter $\lambda > 0$ realisiert werden. Dabei entspricht der Fall $p = 2$ und $q = 1$, bei dieser verallgemeinerten Penalisierung dem bekannten Fall der LASSO-Regression. Durch die Wahl von $p = 2$ und $q = 2$ ergibt sich der Spezialfall der Ridge-Regression (Hastie et al., 2009). In dieser Arbeit wird

die Regression mittels Bayes-Methoden durchgeführt, welche im nächsten Abschnitt berschrieben werden.

3.2 Einführung in die Bayes-Statistik

Bevor die Bayes-Regression beschrieben wird, wird eine allgemeine Kurzeinführung in die Bayes-Statistik gegeben. Die Unterschiede zwischen frequentistischer und bayesianischer Sichtweise liegen in der Interpretation des Wahrscheinlichkeitsbegriffes. In der Frequentistik sind Wahrscheinlichkeiten nur für wiederholbare Ereignisse ermittelbar. Hingegegen kann in der Bayes-Statistik für jedes Ereignis und jede Aussage eine subjektive Wahrscheinlichkeit bestimmt werden. Dies führt zu einer Erweiterung des Wahrscheinlichkeitsbegriffes. Möglich macht dies das Bayes-Theorem

$$p(\theta|M) = \frac{p(\theta, M)}{p(M)} = \frac{p(\theta)p(M|\theta)}{p(M)},$$

(Gelman et al., 2004) wobei $p(M|\theta)$ die Likelihood der Daten M, $p(\theta)$ die A-priori-Dichte und $p(\theta|M)$ die A-posteriori-Dichte von θ gegeben M bezeichnet. Die A-priori-Dichte fasst das gesamte Wissen über die Parameter $\theta = (\theta_1, \ldots, \theta_p)$ zusammen, welches unabhängig von den Daten M vorliegt. Es beschreibt das Wissen über den Parameter vor Sichtung der Daten. Hierfür müssen sowohl Lage als auch Streuung der Parameter bestimmt werden. Die Stärke des A-priori-Wissens kann über die Varianz festgelegt werden. Im Fall des Regressionsmodells gilt, dass die Parameter $\theta = (\beta_1, \ldots, \beta_d, \sigma)$ entsprechen. Die Likelihood $p(M|\theta)$ modelliert die Daten abhängig von den Parametern. Im Allgemeinen gilt bei fester Modellvarianz, je größer die Stichprobengröße ist, umso mehr Gewicht liegt auf den Daten M. Bei wenigen Beobachtungen kommt dem A-priori-Wissen mehr Bedeutung zu. Mit Hilfe des Bayes-Theorems lässt sich die Dichtefunktion für die unbekannten Parameter θ herleiten, nachdem die Daten M beobachtet wurden. Die Konstante

$$p(M) = \int p(M, \theta) d\theta = \int p(M|\theta)p(\theta)d\theta$$

(Gelman et al., 2004) stellt sicher, dass sich die A-posteriori-Verteilung zu 1 integrieren lässt. Damit erhalten alle unbekannten Parameter eine Wahrscheinlichkeitsverteilung. Im Gegensatz zur Frequentistik sind in der Bayes-Statistik die Parameter nicht fest, sondern werden als Zufallsvariablen aufgefasst, deren Verteilung es zu schätzen gilt. Die Wahl der A-priori-Verteilung ist wichtig, aber meistens schwierig. Daher muss genau

überlegt werden, welches Wissen vor Sichtung der Daten vorliegt. Liegt kein Wissen vor, kann dies durch eine uninformative A-priori-Verteilung modelliert werden. In diesem Fall geht die Varianz der Verteilung gegen unendlich.

3.3 Bayes-Regression und Markov-Chain-Monte-Carlo Methoden

Im Fall der Bayes-Regression ist die Zielvariable Y bedingt durch die Matrix der Einflussgrößen X und den unbekannten Parametervektor $\theta = (\beta_1, \ldots, \beta_d, \sigma)$ mit unbekannter Verteilung. Die Herausforderung ist die Schätzung der A-posteriori-Dichte

$$p(\beta|X, Y) \propto p(\beta)p(Y|X, \beta)$$

(Gelman et al., 2004) unter gegebenen A-priori-Wissen $p(\beta)$. Die verschiedenen ℓ_p-Regressionen können im Bayesianischen durch die passende Wahl der Verteilung bezüglich der Likelihood $p(Y|X, \beta)$ realisiert werden. Gesucht wird eine Schätzung für den unbekannten Parameter β, welcher sich aus der realisierten A-posteriori-Verteilung $p(\beta|X, Y)$ ergibt.
Ein Problem in der Berechnung der A-posterori-Verteilung ist, dass diese meist nicht analytisch berechnet werden kann, sondern numerische Verfahren verwendet werden müssen. Hierfür finden in der Bayes-Statistik die Markov-Chain-Monte-Carlo-Methoden Einsatz. Diese liefern ein flexibles Verfahren für statistische Modellierungen, in der die Likelihood punktweise ausgewertet werden kann. Grundidee dieser Methode ist es Werte aus einer groben approximativen A-posteriori-Verteilung zu ziehen und diese solange zu verbessern, bis die Werte aus einer sehr genauen Approximation der A-posteriori-Verteilung entstammen.
Die Umsetzung dieser Idee erfolgt durch die Konstruktion einer Markovkette, welche als stationäre Verteilung die unbekannte A-posteriori-Verteilung besitzt. Eine Markovkette ist eine Sequenz von Zufallsvariablen $\theta = \{\theta^0, \theta^1, \theta^2, \ldots\}, \quad \theta^t \in \mathbb{R}^{(d+1)}$, wobei die Verteilung von θ^t zu Zeitpunkt t, gegeben allen Vorgängerwerten, nur vom letzten Wert θ^{t-1} abhängig ist (Gelman et al., 2004). Eine Möglichkeit eine Markovkette so zu konstruieren, dass sie die A-posteriori-Verteilung $p(\theta|M)$ als stationäre Verteilung hat, ist mittels des Metropolis-Hastings-Algorithmus. Ausgehend von einem Startpunkt θ^0 mit zugehöriger Startverteilung wird in jedem Iterationsschritt ein neuer Kandidat θ^* aus der sogenannten Proposal Distribution $q(\theta^*|\theta^{t-1})$ generiert. Dieser Kandidat wird

in jedem Iterationsschritt mit Akzeptanzwahrscheinlichkeit

$$\alpha(\theta^*|\theta^{t-1}) = \min\left(\frac{p(\theta^*|M)q(\theta^{t-1}|\theta^*)}{p(\theta^{t-1}|M)q(\theta^*|\theta^{t-1})}, 1\right),$$

(Gelman et al., 2004) als neuer Wert der Markovkette akzeptiert. Wird der Kandiddat θ^* abgelehnt, wird der Vorgängerwert θ^{t-1} als neuer Wert der Markovkette $\theta^t = \theta^{t-1}$ gesetzt. Nach hinreichend vielen Iterationsschritten ist die Markovkette gegen die unbekannte A-posteriori-Verteilung konvergiert. Da der stationäre Zustand nicht mehr verlassen wird, sofern er einmal erreicht wurde, können die generierten Werte der Markovkette als Stichprobenwerte (Samples) der A-posteriori-Verteilung verwendet werden um diese zu schätzen.
Die Werte der Markovkette bis die stationäre Verteilung erreicht wird, wird auch Burn-In-Phase gennant und soll für die Schätzungen der A-posteriori entfernt werden. Die Länge der Burn-In-Phase, sowie die Anzahl an Stichprobenwerten müssen vom Anwender selbst festgelegt werden.
In dieser Arbeit wird für die Berechnung der A-posteriori-Verteilung bei der Bayes-Regression die Software *OpenBUGS* (Lunn et al., 2012) verwendet. Die Implementierungen enthalten eine Menge allgemeinerer Algorithmen, die zu der Metropolis-Hastings-Klasse gehören. Speziell in dieser Arbeit erfolgt die Bestimmung der unbekannten Modellparamter mittels dem Hybrid/Hamiltonian Monte-Carlo-Algorithmus. Dieser macht sich die Hamilton Dynamik zu nutze, welche ein Grundkonzept der Physik darstellt. Diese Methode garantiert eine schnellere Konvergenz der Markovkette, mittels Reduzierung der Korrelation zwischen den aufeinanderfolgenden Werten der Markovkette, durch die Einführung einer zusätzlichen Hilfsvariable (Gelman et al., 2004). Die zusätzliche Hilfsvariable, welche im physikalischen Sinne die Momentumvariable bezeichnet, beeinflusst im wesentlichen die Richtung der Markovkette, so dass diese schneller gegen die stationäre Verteilung konvergiert. Ein Metropolis-Hastings-Akzeptanzschritt stoppt die Bewegung der Momentumvariable, sofern diese die stationäre Verteilung erreicht hat. Eine ausführliche Einführung in die Physik und die Methoden der Hamiltonian Monte-Carlo-Methode kann in Neal (2012) vertiefend nachvollzogen werden. Für diese Arbeit sollen die Grundzüge genügen, da die Umsetzung des Algorithmus nicht erheblicher Bestandteil der Betrachtung ist.
Für die Bayes-Regression gilt, dass falls als Verteilung der Likelihood die Normalverteilung gewählt wird, dies dem Fall der ℓ_2-Regression entspricht (Hastie et al., 2009). Für eine Laplaceverteilte Likelihood kann der Fall der ℓ_1-Regression realisiert werden.

Dies führt zu einer robusteren Regressionsschätzung. Zudem kann mit der geeigneten Wahl der A-priori-Verteilung die Verbindung zur Ridge-Regression oder der LASSO-Regression geschaffen werden. Bei der Wahl einer Laplaceverteilten A-priori-Verteilung und einer normalverteilten Likelihood entspricht dies dem frequentistischen Fall der LASSO-Regression (Hastie et al., 2009). Festzuhalten bleibt, dass die A-priori-Verteilung im Fall der Bayes-Regression als Strafterm bzw. Penalisierungsterm verwendet werden kann.
Um die Bayes-Regression bezüglich der verschiedenen ℓ_p-Fälle für allgemeine p zu untersuchen, wird im Folgenden Kapitel (3.4) die p-verallgemeinerte Normalverteilung nach Subbotin (1923) eingeführt. Diese implementiert die Spezialfälle, dass sich für $p = 2$ die Normalverteilung und für $p = 1$ die Laplace-Verteilung ergibt. Dadurch ist es möglich mit Hilfe der verallgemeinerten Normalverteilung die verschiedenen Szenarien der ℓ_p Regression umzusetzen.
Die A-priori-Informationen über den Varianzparameter der Daten σ^2 werden in dieser Arbeit als eher uninformativ, also mit hoher Varianz modelliert.

3.4 p-verallgemeinerte Normalverteilung

In diesem Unterkapitel wird die verallgemeinerte Normalverteilung nach Subbotin (1923) (Kotz et al., 1970) beschrieben. Diese stellt eine Erweiterung der bekannten Normalverteilung dar. Sie besitzt für $p \in (0, \infty)$ die folgende symmetrische Dichte

$$f(x) = \frac{p}{2\alpha\Gamma\left(\frac{1}{p}\right)} \exp\left(-\frac{|x-\mu|^p}{\alpha^p}\right)$$

(Kotz et. al, 1970) mit Erwartungswert bzw. Lokationsparameter $\mu \in \mathbb{R}$ und Streuungsparameter $\alpha \in \mathbb{R}^+$. Die Gammafunktion $\Gamma(.)$ sei hierbei mit für eine positive reelle Zahl r mit $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt$ definiert. Für den Spezialfall $p = 1$ ergibt sich damit die Laplace-Verteilung

$$f(x) = \frac{1}{2\alpha} \exp\left(-\frac{|x-\mu|}{\alpha}\right)$$

und für den Fall $p = 2$ die Normalverteilung

$$f(x) = \frac{1}{\alpha\sqrt{\pi}} \exp\left(-\frac{|x-\mu|^2}{\alpha^2}\right),$$

mit $\Gamma(\frac{1}{2})^2 = \pi$. Die Varianz ist mit

$$\sigma^2 = \frac{\alpha^2 \Gamma\left(\frac{3}{p}\right)}{\Gamma\left(\frac{1}{p}\right)}$$

gegeben. Für einige feste p grafisch anschaulich wird die p-verallgemeinerte Normalverteilung in Abbildung 3.4.1 dargestellt.

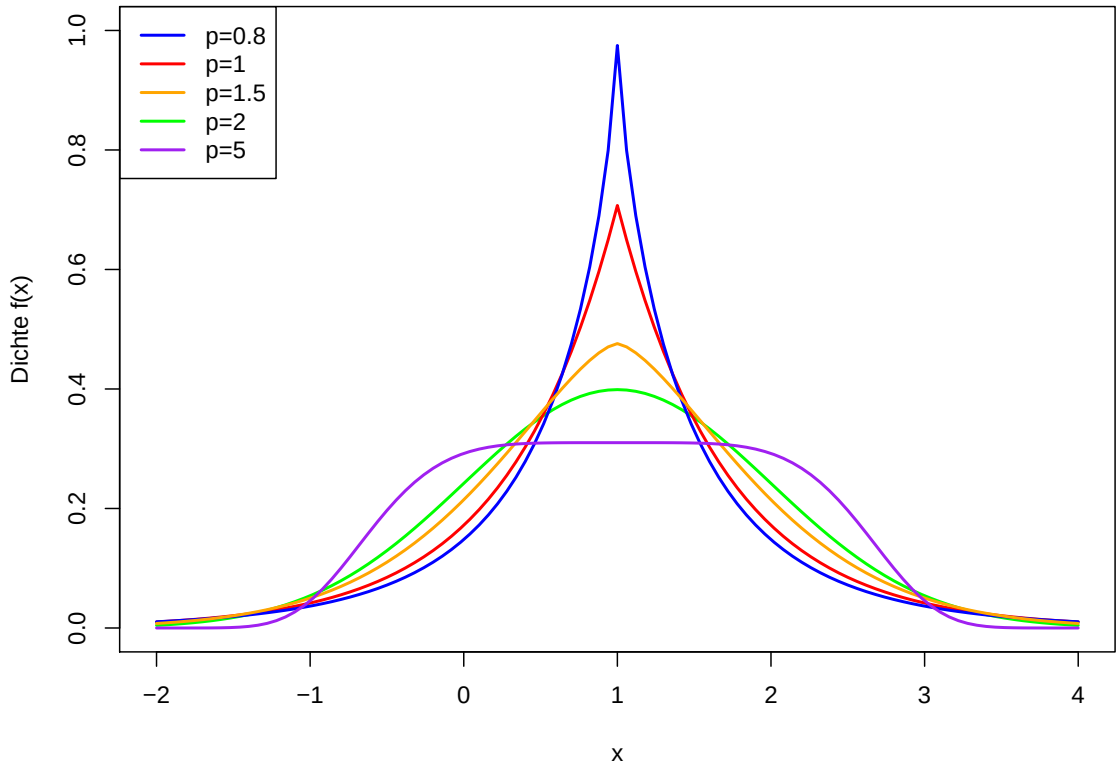

Abbildung 3.4.1: p-verallgemeinerte Normalverteilungen mit Erwartungswert $\mu = 1$ und Varianz $\sigma^2 = 1$ unter Verwendung verschiedener
$p = \{0.8, 1, 1.5, 2, 5\}$.

Der Erwartungswert, sowie die Varianz ist hierbei in allen Fällen mit 1 fest gewählt. In der Abbildung zu erkennen sind die Spezialfälle $p = 2$, welcher die Normalverteilung (grün) ergibt, sowie der Fall $p = 1$, aus welchem die Laplace-Verteilung (rot) resultiert. Des Weiteren ist zu erkennen, dass für steigende p die Verteilung gegen die Gleichverteilung strebt. Für kleiner werdendes p werden die Randbereiche der resultierenden

Verteilung immer breiter. Die p-verallgemeinerte Normalverteilung wird in dieser Arbeit zur Modellierung der A-priori-Verteilungen sowie der Likelihood im Bayes-Modell angewandt. Wird die Likelihood mit einer p-verallgemeinerten Normalverteilung modelliert, entspricht dies der passenden ℓ_p-Regression im bayesianischen Fall. Wie sich die Dichten der p-verallgemeinerten Normalverteilung in ihrer 2-dimensionalen Domäne unterscheiden, wird in Abbildung 3.4.2, anhand der Darstellung von verschiedener Einheitsbälle $B = \{x \in \mathbb{R}^n : \|x\|_p \leq 1\}$ zur jeweiligen p-Norm ersichtlich.

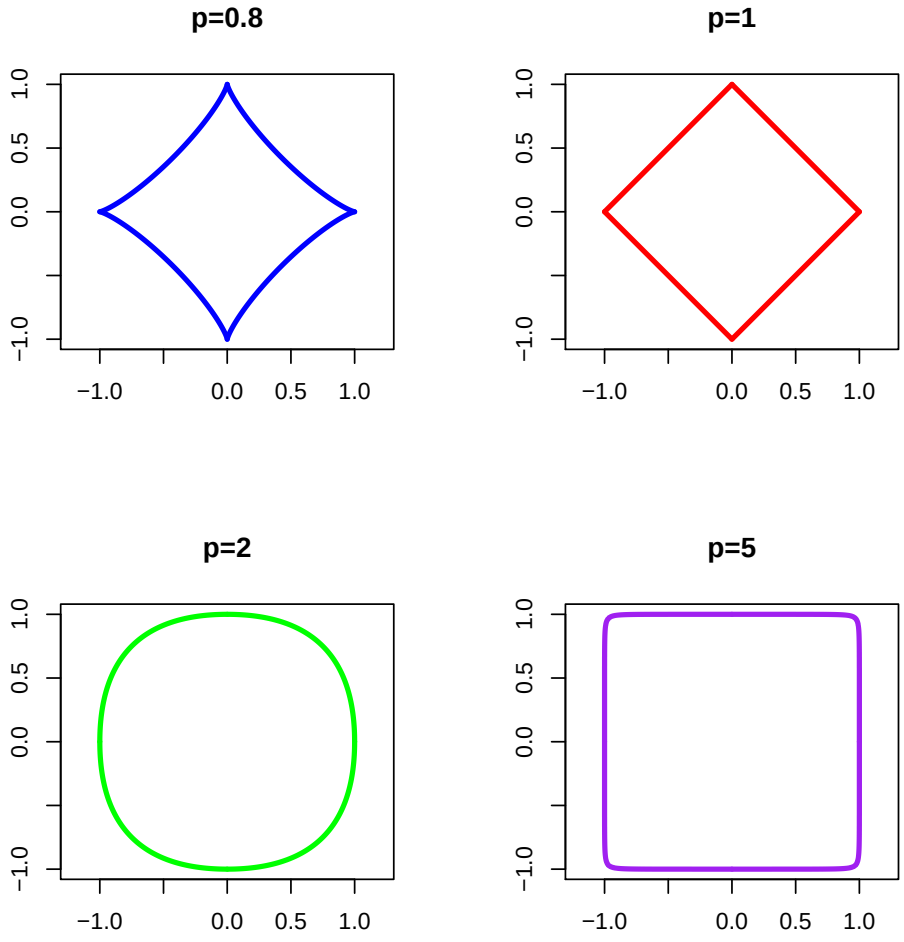

Abbildung 3.4.2: Einheitsbälle der jeweiligen p-Norm für $p = \{0.8, 1, 2, 5\}$.

Im nächsten Kapitel werden die ε-Einbettungen allgemein eingeführt, sowie anschließend die in dieser Arbeit verwendete Einbettung nach Clarkson und Woodruff, sowie die Modifizierung nach Woodruff und Zhang beschrieben.

4 Einbettungsmethoden

Dieses Kapitel wird die ε-Unterraumeinbettungen (4.1) zunächst allgemein beschreiben, sowie die für diese Arbeit verwendete Einbettungsmethode nach Clarkson und Woodruff (4.2). Anschließend wird ein Algorithmus (Kapitel 4.3) aus Woodruff und Zhang (2013) vorgestellt, der für den Fall der ℓ_p-Regression im Fall $p \neq 2$ verwendet wird, um die Unterraumeinbettung für diese Fälle anwendbar zu machen. Abgeschlossen wird das Kapitel von einem Beweis (4.4), welcher erbracht werden muss, um die Güte dieser Modifizierung zu gewährleisten.

4.1 ε-Unterraumeinbettungen

Dieses Kapitel wird eine Einführung in das Konzept von ε-Unterraumeinbettungen liefern. Um große Datensätze für die Regression in kleinere zu transformieren, können ε-Unterraumeinbettungen verwendet werden. Die Untersuchung derer Anwendbarkeit für verschiedene p-Abstandsnormen ist ein Hauptbestandteil dieser Arbeit. Idee hinter den Unterraumeinbettungen ist es, einen großen Datensatz mit Hilfe von zufälligen Projektionen, in einen kleineren Datensatz zu transformieren, welcher algebraisch, bis auf einen kleinen bestimmbaren Fehlerterm ε die selbe Struktur aufweist, wie der gesamte Datensatz. Der Datensatz mit Dimension $n \times d$, $n \gg d$ beschreibt einen d-dimensionalen Untervektorraum des $\mathbb{R}^n$. Sei Π eine zufällige Projektionsmatrix mit Dimension $\mathbb{R}^{n' \times n}$. Mit Hilfe dieser Projektionsmatrix soll der d-dimensionale Untervektorraum des $\mathbb{R}^n$ in die Dimension des $\mathbb{R}^{n'}$ eingebettet werden (Geppert et al., 2015). Die Reduzierung durch zufällige Projektionen kann einfach und effizient implementiert werden.
Der Vorteil von Unterraumeinbettungen ist, dass die Größe der Einbettung nicht von der Anzahl an Beobachtungen n, sondern von der Anzahl an Einflussvariablen d abhängt. Aufgrund dieser Tatsache kann eine Reduktion des Datensatzes, bei Anwendung von Bayes-Regression enorm von Vorteil sein, da die Likelihood schneller bestimmt werden kann.
Für die Unterraumeinbettung wird zusätzlich zur p-Vektornorm (siehe Kapitel 3.1) die p-Matrixnorm benötigt.
Sei dafür $V \in \mathbb{R}^{n \times d}$ eine Matrix. Dann ergibt sich die p-Norm der Matrix V als

$$\|V\|_p = \left(\sum_{i=1}^{d} \|V_i\|_p^p \right)^{\frac{1}{p}} = \left(\sum_{j=1}^{n} \|V^j\|_p^p \right)^{\frac{1}{p}},$$

(Woodruff, Zhang, 2013), wobei $\{1, \ldots, n\}$ die Indexmenge der natürlichen Zahlen bis n beschreibt. Des Weiteren seien V_i der i-te Spalte und V^j die j-te Zeile von V.
Sei eine Matrix $M \in \mathbb{R}^{n \times (d+1)}$ gegeben, welche in dieser Arbeit den Datensatz $M = [Y, X]$ umfassen wird. Dann ist eine ε-Einbettung ($\varepsilon \in (0, 0.5]$) mittels zufälliger Projektionsmatrix $\Pi : \mathbb{R}^n \to \mathbb{R}^{n'}$ gegeben, falls

$$(1 - \varepsilon)\|Mx\|_p \leq \|\Pi Mx\|_p \leq (1 + \varepsilon)\|Mx\|_p, \quad \forall x \in \mathbb{R}^d$$

(Dasgupta et al., 2008) gilt. Die Abbildung Π wird ε-Unterraumeinbettung genannt. Dabei beschreibt $n' = \mathcal{O}\left(\frac{d^2}{\epsilon^2}\right)$ die Anzahl an Zeilen der eingebetteten Daten $\Pi M \in \mathbb{R}^{n' \times (d+1)}$ (Geppert et al., 2015). Der mittels Unterraumeinbettung verkleinerte Datensatz ΠM wird im Folgenden auch als Skizze bezeichnet. Die Wahl von ε beeinflusst zum einen die Größe der Skizze und zum anderen die Genauigkeit der Einbettung, also wie ähnlich sich M und ΠM bezüglich der algebraischen Struktur sind. Ein kleines ε liefert genauere Anpassungen, aber auch eine größere Skizze. Es besteht also ein Trade-off zwischen Anpassungsgüte und Skizzengröße durch das funktionale Verhältnis $n' = \mathcal{O}\left(\frac{d^2}{\epsilon^2}\right)$ (Geppert et al., 2015). Idee und Ziel der Untersuchung ist, ob die spezielle ε-Einbettung nach Woodruff und Clarkson (2013), welche im Folgenden beschrieben wird, bezüglich der Bayes-Regression bis auf eine geringe kontrollierbare Abweichung, die selben Ergebnisse liefert wie die Bayes-Regression ohne Einbettung. Dazu wird überprüft, ob die mittels MCMC-Methoden ermittelte A-posteriori-Verteilung $p(\beta|X, Y)$ mit der A-posteriori-Verteilung $p(\beta|\Pi X, \Pi Y)$ für die eingebetteten Daten übereinstimmt.

4.2 ε-Unterraumeinbettung nach Woodruff und Zhang

Im Folgenden wird eine Methode zur Generierung einer Unterraumeinbettung beschrieben, welche auf der Arbeit von Clarkson und Woodruff (2013) basiert. Diese spezielle Einbettungsmethode wird in dieser Arbeit verwendet. Die Einbettung hat dabei ihre Gültigkeit bezüglich der euklidischen Norm, da diese Methode für ℓ_2-Regressionsprobleme entwickelt ist. Für eine Erweiterung auf ℓ_p-Regression wird zusätzlich eine Modifzierung (4.3) benötigt.
Die ε-Unterraumeinbettung nach Woodruff und Clarkson verwendet eine spärlich besetzte Einbettungsmatrix. Dies garantiert, dass diese Methode den schnellsten Algorithmus für den Fall $n \gg d$ (Clarkson, Woodruff, 2013) liefert. Für den Parameter

n', welcher die Anzahl Zeilen der Skizze beschreibt, sei die Einbettung als lineare Abbildung

$$\Pi = \Phi D : \mathbb{R}^n \to \mathbb{R}^{n'}$$

(Clarkson, Woodruff, 2013) definiert. Für die binäre Matrix $\Phi \in \{0,1\}^{n' \times n}$ sei die zufällige Abbildung

$$h : \{1, \ldots, n\} \to \{1, \ldots, n'\}$$

gegeben, so dass $h(i) \in \{1, \ldots, n'\}$ für jedes $i \in \{1, \ldots, n\}$ diskret gleichverteilt auf $\{1, \ldots, n'\}$ ist (Clarkson, Woodruff, 2013). Dies bedeutet, dass jede Spalte der Matrix $\Phi \in \{0,1\}^{n' \times n}$ mit einem paarweise unabhängig, zufälligen gezogenen Eintrag $\Phi_{(h(i),i)} = 1$ ausgestattet ist. Die restlichen Einträge jeder Spalte sind Null.
Die Matrix $D \in \mathbb{R}^{n \times n}$ ist eine Diagonalmatrix mit zufällig und 4-fach unabhängig gezogenen Einträgen aus der Zweipunktmenge $\{-1, 1\}$ mit jeweils gleicher Wahrscheinlichkeit $\frac{1}{2}$. Dass dabei die eingeschränkten Unabhängigkeitsannahmen genügen, kann in der Arbeit von Geppert et al., (2015) nachvollzogen werden. Die Anzahl an Beobachtungen $n' = \mathcal{O}\left(\frac{d^2}{\epsilon^2}\right)$ der Skizze ist nicht mehr abhängig von der ursprünglichen Anzahl an Beobachtungen n. Im Vergleich zu anderen Einbettungsmethoden ist die Methode nach Clarkson und Woodruff eine sehr einfache und damit schnell umzusetzende. Implementiert wurde diese im R-Softwarepaket `RaProR` von Geppert et al., (2015). Bei der Implementierung der Unterraumeinbettung im Paket `RaProR` wird eine spezielle Methode nach Dietzfelbinger et al. (1997), welche in Geppert et al., (2015) aufgeführt und referenziert ist, verwendet, die es ermöglicht, dass die Einbettungsmatrix Π nicht explizit gespeichert werden muss. Durch die eingeschränkte paarweise bzw. vierfache Unabhängigkeit, muss neben der Skizze $\Pi M \in \mathbb{R}^{n' \times (d+1)}$, lediglich der Seed gespeichert werden, was den Speicherverbrauch auf $\mathcal{O}(log(n))$ statt $\mathcal{O}(n)$ reduziert und damit die Effizienz des Algorithmus garantiert (Geppert et al., 2015).
Für den Fall der ℓ_2-Regression im Bayesianischen Fall, also den Fall $p = 2$ und damit die Normalverteilungsannahme bezüglich der Likelihood $p(Y|X, \beta)$, können mit den bisherigen Methoden bereits erste Ergebnisse ermittelt werden. Die dazugehörigen Simulationen werden in Kapitel 5.2 dargestellt. Für den Fall $p \neq 2$ wird noch eine Modifizierung der Unterraumeinbettung nach Clarkson und Woodruff benötigt, da diese ihr Gültigkeit lediglich zur euklidischen Norm besitzt. Diese Modifizierung der Unterraumeinbettung wird im nächsten Kapitel beschrieben.

4.3 Algorithmus zur Modifizierung von Unterraumeinbettungen nach Woodruff und Zhang für die ℓ_p-Regression

In diesem Unterkapitel wird ein Algorithmus vorgestellt, um die Unterraumeinbettung von Clarkson und Woodruff (2013) so zu modifzieren, dass diese ihre Gültigkeit für Berechnungen bei ℓ_p-Regression ($p \in [1, 2)$) behält, also im Fall dieser Arbeit, dass die Likelihood p-verallgemeinert normalverteilt modelliert wird. Für den Fall $p = 2$ konnte die Optimalität bereits in Clarkson und Woodruff (2013) gezeigt werden.
Für die Gültigkeit bezüglich der ℓ_p-Regression liefert der Algorithmus aus der Arbeit von Woodruff und Zhang (2013) in Kombination mit einer abgewandelten Version des Cauchy-Fast-Regression-Algorithmus aus der Arbeit von Clarkson et al. (2013) die gewünschte Lösung. Für den hier untersuchten Fall $n \gg d$ garantiert die Arbeit von Woodruff und Zhang (2013) den schnellsten Algorithmus für die ℓ_p-Regression für reelle $p \in [1, \infty)$ außer $p = 2$.
Der Algorithmus verwendet die Konstruktion von *gut-konditionierten-Basen* für den ℓ_p-Raum, welche zu Beginn des Kapitel 4.4 definiert sind.
Zunächst soll der in dieser Arbeit angewendete Algorithmus nach Woodruff und Zhang (2013) beschrieben werden, welcher unter Verwendung von Unterraumeinbettungen und exponentialverteilten Zufallsvariablen das ℓ_p-Regressionsproblem löst. Anstatt der üblichen Verwendung von p-stabilen Zufallsvariablen, werden bei Woodruff und Zhang exponentialverteilte Zufallsvariablen verwendet, da diese im Gegensatz zu p-stabilen Zufallsvariablen nicht nur für $p \in [1, 2]$ definiert sind. Zwischen der Verwendung von p-stabilen Zufallsvariablen und der Verwendung des Kehrwertes von exponentialverteilten Zufallsvariablen gibt es einen Zusammenhang, welcher in Woodruff und Zhang (2013) nachvollzogen werden kann, für diese Arbeit aber nicht von Bedeutung ist. Sei die Matrix S eine Realisierung der Unterraumeinbettungsmatrix, im Fall dieser Arbeit die Unterraumeinbettung nach Clarkson und Woodruff (Kapitel 4.2) und $E \in \mathbb{R}^{n \times n}$ eine Diagonalmatrix mit unabhängig identisch verteilten Einträgen

$$1/u_i^{\frac{1}{p}}, \ i = 1, \ldots, n,$$

(Woodruff, Zhang, 2013) wobei $u_i \sim Exp(1)$ eine Realisierung einer exponentialverteilten Zufallsvariable mit Parameter 1 ist. Die für den Algorithmus verwendete Einbettungsmatrix Π ergibt sich dann aus der Multiplikation der beiden eben beschriebenen Matrizen $\Pi = SE$.
Der Algorithmus von Woodruff und Zhang besteht im Wesentlichen aus den folgenden

vier Schritten, die im Anschluss näher erläutert werden:
1. Berechne ΠM mit $M = [X, -Y] \in \mathbb{R}^{n\times(d+1)}$ und $\Pi = SE$
2. Konstruiere eine Matrix $R \in \mathbb{R}^{d\times d}$, so dass gilt, dass MR^{-1} eine gut-konditionierte Matrix ist.
3. Bestimme mit Hilfe der Matrix R eine Samplingmatrix Π', so dass gilt:

$$(1-\varepsilon)\|Mx\|_p \leq \|\Pi' Mx\|_p \leq (1+\varepsilon)\|Mx\|_p, \quad \forall x \in \mathbb{R}^d$$

4. Finde optimale Lösung des Minimierungsproblems

$$\text{argmin}_{\beta\in\mathbb{R}^d} \|\Pi' X\beta - \Pi' Y\|_p.$$

Die Umsetzung der Schritte 2 und 3 wird in dieser Arbeit mit einer leicht abgewandelten Version der Fast-Cauchy-Regression aus Clarkson et al. (2013) durchgeführt. Dieser dient in seiner ursprünglichen Form zur Lösung der ℓ_1-Regression. Er wird im Folgenden für beliebige $p \in [1, 2)$ verallgemeinert. Dabei wird sich auf den Fall $p \in [1, 2)$ beschränkt. Der Fall $p = 2$ wurde bereits in Geppert et al., (2015) untersucht. Für die Fälle $p > 2$ ist aus Woodruff und Zhang bekannt, dass die Dimension der Skizzen für mittelgroße Daten sehr groß werden, so dass sich diese, für eine Betrachtung im Rahmen dieser Arbeit, nicht eignen. Die für diese Arbeit relevanten Schritte werden im Folgenden kurz erläutert und für den eben beschriebenen Algorithmus von Woodruff und Zhang angepasst. Für die Erzeugung der Matrix $R \in \mathbb{R}^{d\times d}$ wird die QR-Zerlegung von ΠM (Π und M aus Schritt 1) berechnet, also $\Pi M = QR$. Die Matrix R wird im Weiteren zur Konstruktion der Matrix $U = MR^{-1}$ verwendet. Der Beweis, dass U auch tatsächlich eine gut-konditionierte Matrix ist, wird zum Abschluss dieses Kapitels geliefert. In Schritt 3 soll eine Samplingmatrix Π' erzeugt werden. Dafür sei für jedes $i \in \{1, \ldots, n\}$ mit

$$\hat{p}_i = \min\left\{1, \frac{\lambda_i}{\sum_{i=1}^n \lambda_i}\right\}$$

(Clarkson et al., 2013) eine Wahrscheinlichkeit für Π' gegeben ist. Die Werte λ_i entsprechen dabei den Leverage-Scores der Matrix U und ergeben sich entsprechend aus den Zeilen von U, mit $\lambda_i = \|U_{(i)}\|_p^p$. Mit Hilfe der Wahrscheinlichkeiten von $\hat{p}_i$ wird die Diagonalmatrix Π' folgendermaßen

$$\Pi'_{ii} = \begin{cases} (1/\hat{p}_i \cdot s)^{1/p} & \text{, mit Wahrscheinlichkeit } \hat{p}_i \\ 0 & \text{, mit Wahrscheinlichkeit } 1-\hat{p}_i \end{cases}$$

(Clarkson et al., 2013) definiert. Die Anzahl s entspricht der Anzahl Zeilen der Skizze. Diese Matrix erfüllt die Bedingungen aus Schritt 3 und mit derer Hilfe kann das Optimierungsproblem aus Schritt 4

$$\text{argmin}_{\beta \in \mathbb{R}^d} \|\Pi' X\beta - \Pi' Y\|_p$$

gelöst werden. In dieser Arbeit wird der Schritt 4 nicht wie angegeben, sondern mittels Verfahren der Bayes-Regression gelöst. Also durch die Bestimmung der A-posteriori $p(\beta|\Pi' X, \Pi' Y)$ mittels beschriebener Markov-Chain-Monte-Carlo-Methoden.
Abschließend wird im folgenden Kapitel noch der Beweis erbracht, dass $U = MR^{-1}$ tatsächlich eine gut-konditionierte Matrix ist.

4.4 Beweis der (α, β, p)-gut-Konditionierung von MR^{-1}

Sei für den Beweis der gut-Konditionierung der Matrix MR^{-1} zunächst die Definition einer (α, β, p)-gut-konditionierten Matrix eingeführt.
Sei $A \in \mathbb{R}^{n \times m}$ eine Matrix mit Rang d. Sei für ein $p \in [1, \infty)$ die Dualnorm q mit $\frac{1}{q} + \frac{1}{p} = 1$ gegeben. Dann gilt weiter $U \in \mathbb{R}^{n \times d}$ ist eine (α, β, p)-gut-konditionierte Basis für den Spaltenraum von A, wenn die beiden Bedingungen

$$(1.) \quad \|U\|_p \leq \alpha$$

$$(2.) \quad \|x\|_q \leq \beta \|Ux\|_p$$

(Dasgupta et al., 2008) erfüllt sind. Sind die Bedingungen erfüllt, gilt dass U eine gut konditionierte Basis ist und für alle $x \in \mathbb{R}^d$, gilt dass $\|x\|_p$ annähernd dem Wert von $\|Ux\|_p$ entspricht.
Eine Ungleichung, welche für den Beweis Verwendung findet, ist aus Woodruff, Zhang, 2013) entnommen und lautet:

$$\forall x : \frac{1}{\mu} \|Mx\|_p \leq \|\Pi Mx\|_2 \leq \mu \|Mx\|_p \quad (i)$$

(Woodruff, Zhang, 2013). Dabei ist μ durch $\mathcal{O}((d \log(d))^{\frac{1}{p}})$ nach oben begrenzt. Durch Umformung ergibt sich als Abschätzung

$$\forall x : \|Mx\|_p \leq \mu \|\Pi Mx\|_2, \quad (ii)$$

welche ebenfalls im Beweis Verwendung findet. Zudem sei $R^{-1} \in \mathbb{R}^{d\times d}$, die obere Dreiecksmatrix aus der QR-Zerlegung. Bezeichne mit $R_i^{-1} \in \mathbb{R}^d$ die i-te Spalte der Matrix R^{-1}. Die Matrix R^{-1} soll im Folgenden in die Summe von d Matrizen mit jeweils selber Dimension wie R^{-1} durch

$$R^{-1} = \sum_{i=1}^{d} R_i^{-1} \cdot e_i^T = \sum_{i=1}^{d} \Big(0, \dots, 0, \underbrace{R_i^{-1}}_{\text{Spalte } i}, 0, \dots, 0 \Big) \quad (iii)$$

aufgespalten werden, mit $e_i^T, \quad i = 1, \dots, d$ den transponierten Einheitsvektoren und $0 \in \mathbb{R}^d$ dem d-dimensionalen Nullvektor.
Es gilt, dass die Matrix $R_i^{-1} \cdot e_i^T$ und der Vektor der i-ten Spalte von R_i^{-1} bezüglich der p-Norm gleich sind.
Im Folgenden wird gezeigt, dass die Matrix MR^{-1} entsprechend $(d\mu, \mu, p)$-gut konditioniert ist. Dazu werden die beiden Bedingungen $\|MR^{-1}\|_p \leq \alpha^*$ und $\|x\|_q \leq \beta^* \|MR^{-1}x\|_p$ aus der Definiton der (α, β, p)-gut-konditionierten Matrix nachgewiesen, mit $\alpha^* = d\mu$ und $\beta^* = \mu$. Alle relevanten Schritte werden dabei numeriert und anschließend ausführlich erläutert.

$$\begin{aligned}
&(1.) \quad \|U\|_p = \|MR^{-1}\|_p \overset{(1)}{=} \Big[\Big(\|MR^{-1}\|_p \Big)^2 \Big]^{\frac{1}{2}} \overset{(2)}{=} \Big[\Big(\|M \textstyle\sum_{i=1}^{d} R_i^{-1} \cdot e_i^T\|_p \Big)^2 \Big]^{\frac{1}{2}} \overset{(3)}{=} \\
&\Big[\Big(\|\textstyle\sum_{i=1}^{d} MR_i^{-1} \cdot e_i^T\|_p \Big)^2 \Big]^{\frac{1}{2}} \overset{(4)}{\leq} \Big[\Big(\textstyle\sum_{i=1}^{d} \|MR_i^{-1} \cdot e_i^T\|_p \Big)^2 \Big]^{\frac{1}{2}} \overset{(5)}{=} \\
&\Big[\Big(\textstyle\sum_{i=1}^{d} \|MR_i^{-1}\|_p \cdot 1 \Big)^2 \Big]^{\frac{1}{2}} \overset{(6)}{\leq} \Big[\underbrace{\Big(\sum_{i=1}^{d} 1^2 \Big)}_{=d} \Big(\textstyle\sum_{i=1}^{d} (\|MR_i^{-1}\|_p)^2 \Big) \Big]^{\frac{1}{2}} = \\
&\sqrt{d} \Big[\textstyle\sum_{i=1}^{d} (\|MR_i^{-1}\|_p)^2 \Big]^{\frac{1}{2}} \overset{(7)}{\leq} \sqrt{d} \Big[\textstyle\sum_{i=1}^{d} (\mu \|\Pi MR_i^{-1}\|_2)^2 \Big]^{\frac{1}{2}} = \\
&\sqrt{d} \Big[\mu^2 \textstyle\sum_{i=1}^{d} \underbrace{\|\Pi MR_i^{-1}\|_2^2}_{=1} \Big]^{\frac{1}{2}} \overset{(8)}{=} \sqrt{d} \Big[\mu^2 \underbrace{\sum_{i=1}^{d} 1}_{=d} \Big]^{\frac{1}{2}} = \sqrt{d}\sqrt{d}\mu = d\mu
\end{aligned}$$

Im ersten Schritt wird lediglich der ganze Term quadriert und radiziert, um in Schritt (6) die Cauchy-Schwarz-Ungleichung anwenden zu können. Schritt (2) beinhaltet die Umformung der Matrix R^{-1} in die Summe $\sum_{i=1}^{d} {R_i}^{-1} \cdot e_i^T$ welche in der der Gleichung (iii) definiert ist. Durch die Linearität kann in Schritt (3) die Matrix M in die Summe gezogen werden. Die Ungleichung (4) entspricht der Dreiecksungleichung $\|\sum x\|_p \leq \sum \|x\|_p$, mit $x = MR_i^{-1} \cdot e_i^T$. In Schritt (5) wird verwendet, dass $MR_i^{-1} e_i^T$ und MR_i^{-1} bezüglich der

p-Norm gleich sind, nämlich dass

$$\|MR_i^{-1} \cdot e_i^T\|_p = \Big(\sum_{j=1}^{n} \| \underbrace{MR_i^{-1} \cdot e_{ij}^T}_{=0 \text{ für } i \neq j} \|_p^p\Big)^{\frac{1}{p}} = \Big(\|MR_i^{-1} e_{ii}\|_p^p\Big)^{\frac{1}{p}} = \|MR_i^{-1}\|_p$$

gilt. Wie bereits erwähnt, wird in Schritt (6) die Cauchy-Schwarz-Ungleichung $(\sum x_i y_i)^2 \leq \sum(x_i)^2 \sum(y_i)^2$ angewandt, indem $x_i = \|MR_i^{-1}\|_p \in \mathbb{R}$ und $y_i = 1$ gewählt wird. In Schritt (7) wird die Ungleichung (ii) aus der Arbeit von Woodruff und Zhang verwendet. Nach weiterer Umformung wird in Schritt (8) ausgenutzt, dass die Spalten von $\Pi MR^{-1} = Q$ aus der QR-Zerlegung orthonormal sind und damit gilt $\|\Pi MR_i^{-1}\|_2^2 = 1$. Daraus resultiert, dass $\alpha = d\mu$ gilt.
Des Weiteren gilt:

$$(2.) \quad \|x\|_q \overset{(1)}{\leq} \|x\|_2 \overset{(2)}{=} \|\Pi MR^{-1}x\|_2 \overset{(3)}{\leq} \mu\|MR^{-1}x\|_p = \mu\|Ux\|_p$$

In der ersten Ungleichung wird lediglich mit $q \geq 2$ die Monotonie der p-Norm ausgenutzt. In Schritt (2) wird die Orthonormalität der d Spalten von $\Pi MR^{-1} = Q$ verwendet. Schritt (3) gilt durch die obere Abschätzung aus Ungleichung (i) (Woodruff, Zhang, 2013). Damit ergibt sich, dass $\beta = \mu$ entspricht, für das wiederum die maximale Abschätzung $\mathcal{O}((d\log(d))^{\frac{1}{p}})$ gilt. Insgesamt folgt, dass $U = MR^{-1}$ entsprechend $(d\mu, \mu, p)$-gut-konditioniert ist.
Aus dieser Tatsache der (α, β, p)- Konditionierung von MR^{-1} resultiert, dass für eine Samplingmatrix $\Pi \in \mathbb{R}^{n' \times n}$ mit Wahrscheinlichkeit 0.99 die Ungleichung

$$(1-\varepsilon)\|Mx\|_p \leq \|\Pi Mx\|_p \leq (1+\varepsilon)\|Mx\|_p, \quad \forall x \in \mathbb{R}^d$$

(Woodruff, Zhang, 2013) erfüllt ist, wobei der Wert n' für den Fall $p \in [1, 2)$ nach oben abgeschätzt wird durch $\mathcal{O}\Big((\alpha\beta)^p d \log\Big(\frac{1}{\varepsilon}\Big)\frac{1}{\varepsilon^2}\Big)$. Damit gilt, dass eine garantierte Güte für die ε-Unterraumeinbettungen für den soeben beschriebenen Algorithmus ermittelt werden kann. Allerdings ist im Fall der $(d\mu, \mu, p)$-gut-Konditionierung die ermittelte Schranke für die Wahl der nötigen Zeilen der Skizze n' mit $\mathcal{O}\Big(d^{p+3}\log(d)^2\log(\frac{1}{\varepsilon})\frac{1}{\varepsilon^2}\Big)$ relativ groß und damit im Fall von Mittelgroßen Daten eher ungeeignet. Ob die Dimension der modifizierten Variante der Unterraumeinbettung tatsächlich so groß gewählt werden muss, um hinreichend gute Resultate bezüglich der Bayes-Regression zu erhalten und wie sich die Resultate für die verschiedenen Werte p verhalten, wird in Kapitel 5.3 empirisch untersucht.

5 Empirische Untersuchung der Anwendbarkeit von ε-Unterraumeinbettungen

In diesem Kapitel wird mit Hilfe zahlreicher Simulationen die Anwendbarkeit der Unterraumeinbettungsmethode nach Clarkson und Woodruff (2013), sowie die Erweiterung nach Woodruff und Zhang (2013), für Bayes-Regression unter Verwendung verschiedener Abstandsnormen untersucht. Umgesetzt wird dies im Folgenden Kapitel durch verschiedenene Verteilungsannahmen bezüglich der Likelihood und der A-priori-Verteilung, sowie durch geeignet modellierte Regressionsdatensätze.
Ziel wird es sein, zu untersuchen, wie sich die eingebetteten Datensätze im Vergleich zu den zugehörigen gesamten Daten verhalten, unter den unterschiedlichsten Modellannahmen bezüglich der Bayes-Regression. Dazu wird überprüft, ob für verschiedene A-priori-Dichten, sowie Modellierungen der Likelihood, die Einbettungsmethode nach Clarkson und Woodruff (2013) eine $1 + \mathcal{O}(\varepsilon)$- Approximation des Originaldatensatzes liefert. Hierzu werden die Resultate der jeweiligen Bayes-Regressionen in Form der Mittelwerte der A-posteriori-Samples für den gesamten Datensatz $M = [Y, X] \in \mathbb{R}^{n\times(d+1)}$, sowie die der eingebetteten Datensätze $\Pi M = [\Pi Y, \Pi X] \in \mathbb{R}^{n'\times(d+1)}$ analysiert und miteinander verglichen. Ein Vergleich dieser Ergebnisse liefert Erkenntnisse über die Praxistauglichkeit der Einbettungsmethode bezüglich der Bayes-Regression.
Da sowohl für die Modellierung der Likelihood, sowie auch für die Modellierung A-priori-Verteilung die verallgemeinerte Normalverteilung aus Kapitel 3.4 Verwendung findet, wird im Folgenden von einer p-verallgemeinerten Normalverteilung für die Modellierung der Likelihood und einer q-verallgemeinerten Normalverteilung für die Wahl der A-priori-Verteilung unterschieden. Zunächst wird in Kapitel 5.2 der Fall einer normalverteilten Likelihood ($p = 2$) und verschiedenen q-verallgemeinerten Normalverteilungen für die A-priori-Verteilung ($q \in \{0.6, 1, 1.2, 1.4, 1.6, 1.8, 2, 3, 5\}$) betrachtet. Dabei werden unterschiedliche A-priori-Erwartungswerte, sowie Varianzen untersucht. Durch die geeignete Wahl der A-priori-Verteilung können diese als Penalisierungsterm für die Parameter β betrachtet und angewendet werden. Zu untersuchen gilt es hierbei, ob sich diese Modellannahmen auf die eingebetteten Daten sowie die gesamten Daten gleichermaßen auswirken.
In Kapitel 5.3 wird die Verteilung der Likelihood $p(Y|X, \beta)$ mit einer p-verallgemeinerter Normalverteilung modelliert (für $p \in \{1, 1.2, 1.4, 1.6, 1.8\}$). Dadurch soll der Fall der ℓ_p-Regression erzeugt und simuliert werden. In diesen Fällen kommt der in Kapitel 5.3 beschriebene Algorithmus zum Einsatz, um die Einbettungsmethode nach Clarkson

und Woodruff bezüglich der ℓ_p-Regression zu modifizieren. Dabei werden zunächst uninformative A-priori-Dichten betrachtet. In Kapitel 5.4 werden diese Modelle um eine informative A-priori-Annahme erweitert und ausgewertet.
In allen Fällen wird untersucht, wie sich die Güte der Anpassung bei verschiedenen Skizzengrößen, sowie verschiedenen Werten für ε verhält.
Alle Simulationen werden mit Hilfe von Markov-Chain-Monte-Carlo-Verfahren (Kapitel 3.3) durchgeführt. Die verwendeten Softwareprogramme sind *OpenBUGS* (Lunn et al., 2012), sowie die Software R (R Core Team), welche mit dem R-Softwarepaket `R2WinBUGS` von Gelman et al. (2005) verknüpft werden. Für die Erstellung der benötigten Skizze für die ε-Einbettung wird das R-Softwarepaket `RaProR` *Random Projections for Bayesian linear regression* von Geppert et al. (2015) verwendet. Bevor die Ergebnisse der verschiedenen Simulationen dargestellt werden, wird zunächst im folgenden Unterkapitel die Simulation der in dieser Arbeit verwendeten Datensätze vorgestellt.

5.1 Simulation des Regressionsdatensatzes

Für die in dieser Arbeit zu untersuchende Situation wird der Fall $n \gg d$, also einer wesentlich größeren Anzahl der Beobachtungen n als Einflussvariablen d betrachtet. Es wird ein Datensatz für ein multivariates lineares Regressionsproblem $Y = X\beta + \xi$ (Kapitel 3.1) modelliert. In diesem Datensatz wird $n = 10000$ und $d = 10$ gewählt, mit Designmatrix $X \in \mathbb{R}^{10000\times 10}$ und $\beta \in \mathbb{R}^{10}$, dem Vektor der unbekannten Parameter β. Die Parameter β_i werden in dieser Arbeit mit einer Gamma(1,1)-Verteilung realisiert, um diese positiv und nahe bei 0 zu modellieren.
Für die Einflussvariablen werden zunächst 10 Zufallsvariablen $\nu_i = 1, \ldots, 10$ aus einer Standardnormalverteilung gezogen, welche dann als Erwartungswerte für die Einflussvariablen dienen. Anschließend werden für jede Einflussgröße 10000 Beobachtungen mit Erwartungswert ν_i und Standardabweichung 10 simuliert, sodass für jeden Wert aus X

$$X_{ij} \sim N(\nu_i, 100) \qquad , i = 1, \ldots, 10, j = 1, \ldots, 10000,$$

gilt. Der jeweilige Eintrag des resultierenden Vektors $X\beta \in \mathbb{R}^{10000}$ dient zur Simulation der Zielvariable Y, welche ebenfalls mit einer Normalverteilung und Standardabweichung $\xi \in \{1, 10\}$ simuliert

$$Y_j \sim N(X\beta_j, \xi^2) \qquad , j = 1, \ldots, 10000,$$

wird. Ziel ist eine möglichst genaue Bestimmung des als unbekannt angenommenen Parametervektors β.
Um die Ergebnisse vergleichen zu können wird im Folgenden eine Realisierung für den Parameter β festgehalten. Die Schätzung mittels frequentistischer Regression für die Regressionskoeffizienten sei durch

$$\hat{\beta} = (0.638, 0.002, 0.016, 0.374, 5.479, 0.499, 1.373, 0.259, 0.047, 0.430)^T$$

gegeben. Zur Reproduzierbarkeit der Ergebnisse werden diese Realisation der Gammaverteilung für die Regressionsparameter in den folgenden Simulationen festgehalten. Zu erkennen ist, dass der Koeffizient für $\hat{\beta}_5$ im Vergleich zu den restlichen geschätzen Regressionsparameter mit 5.479 weiter entfernt von 0 ist und damit von den anderen Werten abweicht. Die Summe der Residuen mit frequentistischer linearer Regression zur euklidischen Norm beträgt dabei $\|X\hat{\beta} - Y\|_2 = 1006.832$. Dieser Wert wird verwendet um eine erste Güte der Unterraumeinbettung zu bestimmen. Es ist zu beachten, dass die Einbettungsmethode eine zufällige Projektion ist und sich damit jedes mal Skizzen mit anderer Güte ergeben. So kann es, wenn auch mit geringer Wahrscheinlichkeit δ passieren, dass die Einbettung bereits zu größeren Abweichungen als der vorgegebene Wert ε führt. Um diese Fälle bereits vorher auszuschließen, werden hierfür zunächst mit Hilfe der mit der frequentistischen linearen Regression die Regressionskoeffizienten $\hat{\beta^G}$ für den Gesamtdatensatz geschätzt und mit den geschätzten Regressionskoeffizieten $\hat{\beta^S}$ des skizzierten Datensatz miteinander bezüglich der euklidischen Norm verglichen. Eine Skizze wird demnach nur betrachtet, wenn

$$\frac{\|X\hat{\beta^S} - Y\|_2}{\|X\hat{\beta^G} - Y\|_2} < 1 + \varepsilon,$$

erfüllt ist. Dies gewährleistet, dass der Datensatz erfolgreich eingebettet wurde. Anhand der erfolgreich eingebetteten Daten lassen sich dann unter den verschiedenen Modellannahmen Aussagen über die Anwendbarkeit treffen. Ein wesentlicher Bestandteil dieses Gütekriteriums ist die Wahl von ε.
Der Wert ε ist die vorgegebene Abweichung bei der ε-Unterraumeinbettung und wird in dieser Arbeit mit 0.1 oder 0.2 gewählt.

5.2 ℓ_2-Regression unter Verwendung von p-verallgemeinert normalverteilten A-priori-Verteilungen

In diesem Kapitel wird die Anwendbarkeit und Güte bei Bayes-Regression mittels eingebetteten Datensätzen und normalverteilter Likelihood untersucht. Bezüglich der A-priori-Dichten werden hierbei verschiedene q-verallgemeinerte Normalverteilungen (Kapitel 3.4) verwendet.
($q \in \{0.6, 1, 1.2, 1.4, 1.6, 1.8, 2, 3, 5\}$). Dabei gilt, dass diese Verteilung für höhere Werte q gegen die Gleichverteilung strebt. Für $q = 1$ entspricht dies dem Fall einer Laplaceverteilten A-priori-Verteilung. Im Fall $q = 2$ entspricht dies dem Fall der Normalverteilung. Die Wahl von q beeinflusst neben der A-priori-Varianz die Stärke und Bedeutsamkeit des A-priori-Wissens. Aufgrund der Tatsache, dass für kleinere Werte q die entsprechende Verteilung breitere Randbereiche besitzt, nimmt dadurch der Einfluss der entsprechenden Verteilung auf die A-posteriori-Resultate ab.
Als Szenario für die zu untersuchenden Simulationen werden zudem verschiedene A-priori-Varianzen ($\sigma^2 = \{0.01, 0.1, 1, 10, 100\}$) und verschiedene A-priori-Erwarungswerte ($\mu = \{1, 10\}$) verwendet. Jede Einstellung wird sowohl mit dem gesamten Datensatz, sowie mit dem skizzierten Datensatz durchgeführt. Aufgrund der Tatsache dass die Likelihood normalverteilt modelliert wird, ergibt sich für jede Simulation der Fall der ℓ_2-Regression im bayesianischen.
Für die Unterraumeinbettung nach Clarkson und Woodruff wird in dieser Arbeit standardmäßig die Einstellung $\varepsilon = 0.1$ oder $\varepsilon = 0.2$ verwendet, woraus resultiert, dass bei $d = 10$ Einflussgrößen, der eingebettete Datensatz für $\varepsilon = 0.1$ lediglich noch $n' = 1024$ Beobachtungen besitzt. Wird $\varepsilon = 0.2$ gewählt verringert sich die Anzahl zu lediglich $n' = 256$. Dabei sei noch einmal darauf verwiesen, dass diese Anzahl unabhängig von der Gesamtdatenmenge n ist.
Bei der Bayes-Regression gilt allgemein, dass umso mehr Datenpunkte zu Verfügung stehen, dass umso mehr Gewicht auf der Likelihood, also den Daten liegt und umso weniger Bedeutung dem A-priori-Wissen und deren Modellierung zu kommt. In der Arbeit von Geppert et al., (2015) wird gezeigt, dass die Likelihood der eingebetteten Daten durch umgewichten der Datenpunkte bei der Einbettungsmethode im Wesentlichen die selbe Wahrscheinlichkeitsdichte aufweist, wie die Likelihood des Gesamtdatensatz. Dies bedeutet, dass der Einfluss der A-priori-Verteilung bezüglich der A-posteriori-Verteilung, sowohl im Fall der eingebetteten Daten als auch im Gesamtdatenfall, gleich sein sollte, obwohl die eingebetteten Daten eine deutlich geringere Anzahl an Beobachtungen auf-

weisen.
Im Folgenden sei das arithmetische Mittel der jeweils 10000 realisierten A-posteriori-Samples bei der Bayes-Regression, der betrachtete Schätzwert $\tilde{\beta}^G \in \mathbb{R}^{10}$ für die Koeffizienten von β im Fall der gesamten Daten und $\tilde{\beta}^S \in \mathbb{R}^{10}$ im Fall von skizzierten Daten. Als Schätzung für den somit tatsächlich beobachteten ℓ_2-Abstand zwischen der Anpassung des Gesamtdatensatzes und der skizzierten Daten wird im Folgenden der Quotient der Residuen

$$\hat{\varepsilon} := \frac{\|X\tilde{\beta}^S - Y\|_2}{\|X\tilde{\beta}^G - Y\|_2} - 1 = \frac{\tilde{\nu}^S}{\tilde{\nu}^G} - 1$$

betrachtet. Dieser Wert gilt im Folgenden als ein Kriterium für die Güte der Unterraumeinbettung. Der Schätzwert $\hat{\varepsilon}$ sollte den vorgegeben Fehlerterm ε nicht überschreiten. Zu beachten ist, dass bei dieser Berechnung der Schätzwert $\tilde{\beta}^S$ mit dem Originaldatensatz X multipliziert wird und nicht mit dem eingebetteten ΠX. Dies geschieht aus dem einfachen Grund, dass die realisierten Koeffizienten $\tilde{\beta}^S$ eine möglichst genaue Schätzung $X\tilde{\beta}^S$ für den wahren Wert Y der Originaldaten liefern sollen. Für die Einstellungen der Markov-Chain-Monte-Carlo-Methoden sei im Folgenden der Paramter für die Standardabweichung der Likelihood als uninformativ modelliert. Dies geschiet mit einer Gleichverteilung von 0 bis 150 (bzw. 0 bis 500 in einigen Fällen der ℓ_p-Regression). Dadurch wird für diesen Parameter eine eher uninformative Annahme modelliert. Für jeden Parameter werden 20000 Iterationen realisiert, wobei die ersten 10000 als Burn-In-Phase aus der Berechnung entfernt werden. Bei den meisten Simulationen würde eine geringere Anzahl an Burn-In Werten genügen, die Wahl von 10000 garantiert aber in allen Fällen eine hinreichend große Burn-In-Phase.
Dies führt zum ersten Teil der Untersuchung mit normalverteilter Likelihood und q-verallgemeinerten A-priori-Dichten bezüglich des Regressionskoeffizienten β. Als erstes werden die resultierenden A-posteriori-Resultate $\tilde{\theta} = \{\tilde{\beta}_1, \ldots, \tilde{\beta}_d, \tilde{\sigma}\}$ der Gesamtdaten und der skizzierten Daten betrachtet. Dabei sind die Regressionskoeffizienten β alle mit A-priori Erwartungswert $\mu = 1$ und Varianz $\sigma^2 = 1$ modelliert. Es wird der Datensatz mit Standardabweichung $\xi = 10$ betrachtet. Die ermittelten Schätzwerte der Regressionkoeffizienten für die Fälle $q \in \{0.6, 1, 1.2, 1.4, 1.6, 1.8, 2, 3, 5\}$ im Fall der Gesamtdaten können der Tabelle 5.2.1 entnommen werden. Die zugehörigen Ergebnisse für die eingebetteten Daten mittels Unterraumeinbettung und $\varepsilon = 0.1$ sind entsprechend in Tabelle 5.2.2 aufgeführt.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.638	0.638	0.638	0.638	0.638	0.638	0.638	0.638	0.638
$\tilde{\beta}_2$	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
$\tilde{\beta}_3$	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016
$\tilde{\beta}_4$	0.373	0.373	0.373	0.373	0.373	0.373	0.373	0.373	0.373
$\tilde{\beta}_5$	5.480	5.480	5.480	5.480	5.480	5.480	5.479	5.479	5.468
$\tilde{\beta}_6$	0.499	0.499	0.499	0.499	0.499	0.499	0.498	0.498	0.499
$\tilde{\beta}_7$	1.372	1.372	1.372	1.372	1.372	1.372	1.372	1.372	1.372
$\tilde{\beta}_8$	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260
$\tilde{\beta}_9$	0.049	0.049	0.049	0.049	0.049	0.048	0.049	0.048	0.048
$\tilde{\beta}_{10}$	0.430	0.430	0.430	0.430	0.430	0.430	0.430	0.430	0.430
$\tilde{\sigma}^G$	10.074	10.074	10.074	10.075	10.074	10.074	10.074	10.074	10.075
$\tilde{\nu}^G$	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.9

Tabelle 5.2.1 : Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 1 mittels gesamten Daten.

Anhand der Ergebnisse ist zu erkennen, dass die Resultate für die Bayes-Regression mit dem vollen Datensatz sehr nahe an den zu Anfang des Kapitel genannten frequentistischen Schätzwerten $\hat{\beta}$ liegen und beide Regressionen im wesentlichen zu den gleichen Ergebnissen führen. Auch die geschätzte Standardabweichung $\tilde{\sigma}^G$ stimmt mit der Modellierung des Datensatz überein. Des Weiteren ist zu erkennen, dass die resultierenden A-posteriori-Ergebnisse sich im Fall der gesamten Daten für die Wahl von q nicht unterscheiden. Dies lässt darauf schließen, dass die A-priori-Verteilungen in diesem Modell mit einer Varinz $\sigma^2 = 1$ nicht informativ genug sind um die Ergebnisse zu beeinflussen. Es ist zu beachten dass der Fall von sehr großen Datensätzen betrachtet wird. Der Einfluss der A-priori-Verteilung ist daher eher gering. Bestätigt wird dies durch die Fälle in dem die A-priori Varianz mit $\sigma^2 = \{10, 100\}$ modelliert wird. Bei der Verwendung einer höheren A-priori-Varianz liefern die Regressionen für alle q die selben Ergebnisse und gute Resultate durch Verwendung der ε-Unterraumeinbettung. So liefern für eine A-priori-Varianz $\sigma^2 = 10$ alle Modelle die selben geschätzten Werte $\hat{\varepsilon} = 0.005$ (Tabelle A.7). Die Ergebnisse werden im Folgenden nicht näher betrachtet. In diesen Fällen kann generell von einer Anwendbarkeit der Unterraumeinbettung bezüglich Bayes-Regression gesprochen werden.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.618	0.619	0.617	0.618	0.617	0.618	0.616	0.617	0.622
$\tilde{\beta}_2$	-0.041	-0.039	-0.038	-0.040	-0.041	-0.041	-0.041	-0.039	-0.035
$\tilde{\beta}_3$	-0.014	-0.010	-0.013	-0.012	-0.015	-0.013	-0.014	-0.011	-0.003
$\tilde{\beta}_4$	0.341	0.339	0.339	0.341	0.340	0.340	0.338	0.338	0.338
$\tilde{\beta}_5$	5.494	5.491	5.491	5.492	5.494	5.492	5.490	5.482	5.390
$\tilde{\beta}_6$	0.431	0.435	0.432	0.431	0.431	0.432	0.430	0.432	0.433
$\tilde{\beta}_7$	1.361	1.362	1.364	1.364	1.362	1.363	1.366	1.364	1.368
$\tilde{\beta}_8$	0.291	0.291	0.291	0.290	0.292	0.292	0.290	0.291	0.285
$\tilde{\beta}_9$	0.049	0.049	0.049	0.048	0.047	0.047	0.050	0.049	0.051
$\tilde{\beta}_{10}$	0.435	0.433	0.433	0.433	0.434	0.433	0.432	0.431	0.434
$\tilde{\sigma}^S$	31.02	31.03	31.04	31.02	31.03	31.04	31.03	31.03	31.21
$\tilde{\nu}^S$	1011.8	1011.4	1011.6	1011.6	1011.9	1011.7	1011.8	1011.5	1014.8
$\hat{\varepsilon}$	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.008

Tabelle 5.2.2 : Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 1 mittels skizzierten Daten für $\varepsilon = 0.1$.

Die Schätzungen der Regressionkoeffizienten mittels skizzierten Daten unter Verwendung von $\varepsilon = 0.1$ führen anhand der Werte $\hat{\varepsilon}$ erkennbar für alle q zu hinreichend genauen Ergebnissen. Erwähnenswert ist, dass der Schätzwert für die Standardabweichung $\tilde{\sigma}^S = 31.03$ am Beispiel $q = 1$ im eingebetteten Datensatz höher ist als die geschätzte Standardabweichung im Originaldatensatz $\hat{\sigma}^G = 10.074$. Der Grund für die höhere Standardabweichung liegt an dem Skalierungswert $\sqrt{\frac{n}{n'}} = \sqrt{\frac{10000}{1024}} = 3.125$, um den die Standardabweichung im skizzierten Datensatz steigt. Im Fall $\varepsilon = 0.2$ ergibt sich dieser Wert zu $\sqrt{\frac{10000}{256}} = 6.25$. Dadurch ergeben sich auch größere Varianzen bezüglich der A-posteriori-Samples. Dieser Effekt wird anhand Abbildung 5.2.1 veranschaulicht. In dieser werden Boxplots der jeweils 10000 A-posteriori-Samples in den Fällen Gesamtdatensatz, sowie den skizzierten Datensätzen für $\varepsilon = \{0.1, 0.2\}$ repräsentativ für die Parameter $\tilde{\beta}_1$ (links) und $\tilde{\beta}_5$ (rechts) dargestellt. In beiden Fällen werden die Werte betrachtet, welche mit einer Laplaceverteilte A-priori ermittelt werden. Die Standardabweichungen der A-posteriori-Samples im Fall $\tilde{\beta}_1$ liegen bei 0.010 für die Gesamtdaten und 0.030 bei der Einbettung mit $\varepsilon = 0.1$ und 0.063 für die Einstellung $\varepsilon = 0.2$. Dabei zeigt sich

deutlich der Effekt der höheren Streuung für die Wahl eines größeren ε. Wie genau die Schätzungen der Regressionskoeffizienten in den eingebetteten Datensätzen sind unterliegt dabei der Beschränkung des gewählten ε und der daraus resultierenden Größe der Skizze.

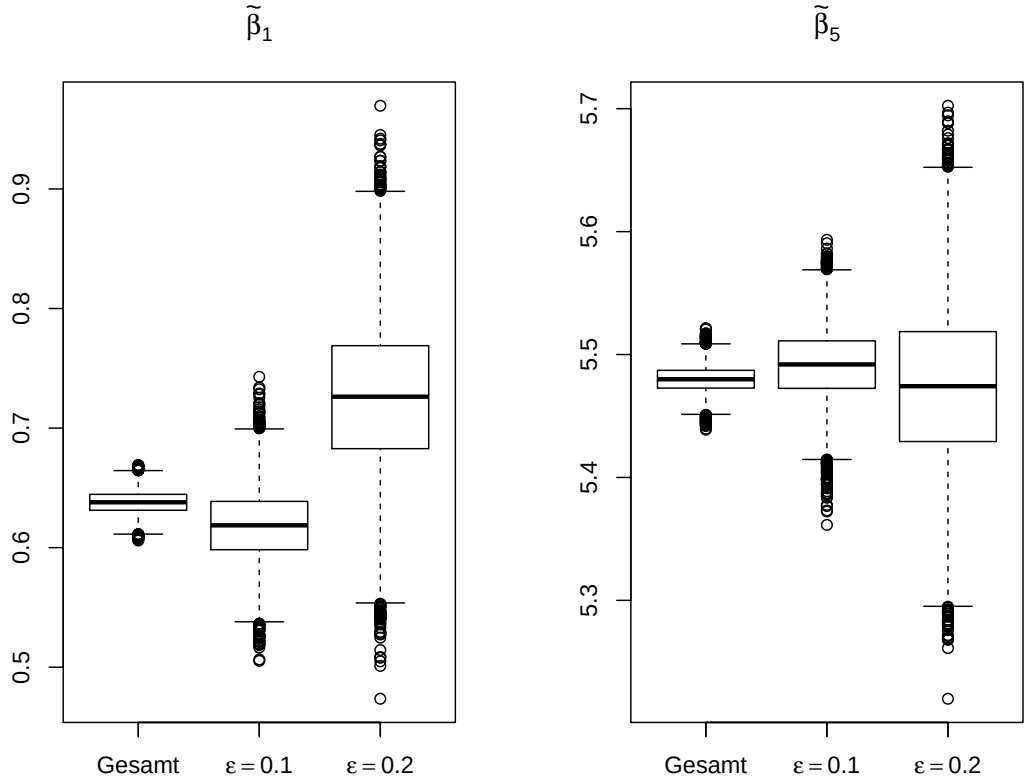

Abbildung 5.2.1: Boxplots der 10000 resultierenden A-posteriori-Samples für $\tilde{\beta}_1$ (links) und $\tilde{\beta}_5$ (rechts) aus der Bayes-Regression mit laplaceverteilter A-priori mit Erwartungswert 1 und Varianz 1 für den Gesamtdatensatz, sowie zwei Unterraumeinbettungen nach Clarkson und Woodruff mit $\varepsilon = \{0.1, 0.2\}$.

Die Mittelwerte der A-posteriori Samples für die Regressionskoeffizienten, sowie der Standardabweichung im Fall $\varepsilon = 0.2$ sind der Tabelle (A.1) im Anhang zu entnehmen. Es zeigen sich dabei im Grunde die Ergebnisse bestätigt, dass in dieser Modellwahl die Wahl der A-priori-Verteilung keinen wesentlichen Einfluss auf die A-posteriori-Resultate haben. Aus diesem Grund wird im Folgenden die A-priori-Varianz reduziert, was ein stärkeres A-priori-Wissen darstellen soll. Dabei wird die Wahl des A-priori Erwartungs-

wert für alle Parameter bei 1 beibehalten. Dies bedeutet gerade für den Parameter β_5, dass eine ungünstige A-priori-Annahme getroffen wird. In Tabelle 5.2.3 sind erneut die A-posteriori-Mittelwerte der Bayes-Regressionen mit den gesamten Daten für die verschiedenen Verteilungen mit A-priori-Erwartungswert 1 und A-priori-Varianz $\sigma^2 = 0.1$ gegeben.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.638	0.638	0.638	0.638	0.638	0.638	0.638	0.638	0.637
$\tilde{\beta}_2$	0.003	0.003	0.004	0.003	0.004	0.004	0.004	0.005	0.038
$\tilde{\beta}_3$	0.016	0.017	0.017	0.017	0.017	0.017	0.017	0.017	0.081
$\tilde{\beta}_4$	0.373	0.373	0.374	0.374	0.373	0.374	0.374	0.374	0.363
$\tilde{\beta}_5$	5.480	5.479	5.479	5.479	5.478	5.477	5.475	5.436	3.560
$\tilde{\beta}_6$	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.488
$\tilde{\beta}_7$	1.371	1.371	1.372	1.371	1.371	1.372	1.371	1.372	1.367
$\tilde{\beta}_8$	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.264
$\tilde{\beta}_9$	0.049	0.049	0.049	0.049	0.049	0.049	0.049	0.050	0.083
$\tilde{\beta}_{10}$	0.430	0.430	0.430	0.430	0.430	0.430	0.430	0.430	0.436
$\tilde{\sigma}^G$	10.07	10.07	10.07	10.07	10.07	10.07	10.07	10.08	21.61
$\tilde{\nu}^G$	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1007.8	2160.4

Tabelle 5.2.3 : Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 0.1 mittels gesamten Daten.

Für die Fälle $q \leq 3$ zeigen sich erneut keine Unterschiede in den A-posteriori-Resultaten. Für dem extremeren Fall $q = 5$ werden bereits schlechtere Anpassungen ausgemacht. Wie in Abbildung 3.4 zu erkennen, besitzt die q-verallgemeinerte Normalverteilung für $q = 5$ kaum noch Randbereiche. Die Wahl einer niedrigen Varianz führt zusätzlich, dazu dass die A-priori-Annahme an Bedeutung gewinnt und stärkeren Einfluss auf die resultierende A-posteriori-Verteilung haben sollte. Die Regressionsparameter sollten dadurch alle näher an dem A-priori-Erwartungswert liegen. In Tabelle 5.2.4 sind die Resultate mittels Unterraumeinbettung nach Clarkson und Woodruff für $\varepsilon = 0.1$ und in Tabelle 5.2.5 mit $\varepsilon = 0.2$ dargestellt.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.619	0.620	0.620	0.620	0.623	0.621	0.624	0.636	0.752
$\tilde{\beta}_2$	-0.037	-0.035	-0.033	-0.031	-0.028	-0.031	-0.024	0.015	0.380
$\tilde{\beta}_3$	-0.011	-0.009	-0.008	-0.008	-0.009	-0.004	-0.001	0.037	0.381
$\tilde{\beta}_4$	0.342	0.344	0.344	0.343	0.341	0.341	0.343	0.343	0.433
$\tilde{\beta}_5$	5.491	5.492	5.487	5.482	5.476	5.466	5.450	5.093	2.318
$\tilde{\beta}_6$	0.436	0.438	0.436	0.435	0.438	0.436	0.440	0.448	0.553
$\tilde{\beta}_7$	1.361	1.357	1.359	1.362	1.358	1.366	1.362	1.381	1.471
$\tilde{\beta}_8$	0.294	0.295	0.294	0.295	0.296	0.294	0.295	0.280	0.331
$\tilde{\beta}_9$	0.051	0.048	0.052	0.055	0.057	0.060	0.060	0.095	0.398
$\tilde{\beta}_{10}$	0.435	0.437	0.437	0.438	0.438	0.435	0.437	0.442	0.524
$\tilde{\sigma}^S$	31.03	31.04	31.03	31.03	31.04	31.04	31.07	33.59	106.4
$\tilde{\nu}^S$	1011.2	1011.0	1010.9	1010.9	1010.8	1010.9	1010.7	1081	3366
$\hat{\varepsilon}$	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.0073	0.558

Tabelle 5.2.4 : Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 0.1 mittels skizzierten Daten für $\varepsilon = 0.1$.

Dabei liefern die Ergebnisse für die Einbettung mit vorgegeben Fehlerterm $\varepsilon = 0.1$, wodurch der skizzierte Datensatz noch $n' = 1024$ Beobachtungen enthält, relativ stabile Resultate, trotz der sehr starken und im Fall von β_5 ungünstigen A-priori-Anahme. Die Resultate führen wieder nur im Fall $q = 5$ zu einer schlechteren Anpassung. Dies gilt sowohl im Fall mit den gesamten Daten als im Fall der Unterraumeinbettung. Der Grund dafür sind wohl aufgrund numerischer Probleme auszumachen. Wird der Exponentialterm der verallgemeinerten Normalverteilung für $q = 5$ betrachtet, mit Beispielsweise $x = 5.5$ ergibt sich mit Erwartungswert 1 und Varianz 0.1 der Exponentialterm $\exp(-\frac{|5.5-1|^5}{\alpha^5})$ mit $\alpha = \sqrt{\frac{0.1\Gamma(1/5)}{\Gamma(3/5)}} = 0.18$ insgesamt zu $\exp = (\frac{4.5^5}{0.18^5}) = \exp(23814.97)$, was mit der verwendeten Software nicht ohne weiteres darstellbar ist. Dadurch können die verwendeten Markov-Chain-Monte-Carlo-Methoden beispielsweise den Paramter $\tilde{\beta}_5$ nicht hinreichend gut modellieren. Die verallgemeinerte Normalverteilung eignet sich für hohe q in diesen Fällen nicht zur Modellierung. Dabei gelten die Resultate lediglich für den hier erzeugten Fall einer ungünstigen A-priori-Annahme.
Ein wenig anders verhalten sich die Resultate für den Fall $\varepsilon = 0.2$. In diesem Fall

ist ein Einfluss schon bei einer normalverteilten A-priori-Verteilung ($q = 2$) zu erkennen im Vergleich zu einer laplaceverteilten A-priori-Verteilung ($q = 1$). Für die Werte $\tilde{\nu} = \|X\tilde{\beta} - Y\|_2$ ist ebenfalls bereits ein Anstieg zu verzeichnen, was gleichbedeutend mit einer jeweils immer schlechteren Anpassung für die entsprechende Wahl q ist.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.736	0.731	0.729	0.728	0.725	0.716	0.714	0.645	0.630
$\tilde{\beta}_2$	-0.001	-0.001	0.009	0.010	0.013	0.020	0.029	0.318	0.489
$\tilde{\beta}_3$	-0.014	-0.009	-0.002	0.004	0.010	0.024	0.040	0.482	0.626
$\tilde{\beta}_4$	0.417	0.419	0.413	0.423	0.414	0.412	0.409	0.411	0.502
$\tilde{\beta}_5$	5.481	5.464	5.452	5.435	5.404	5.356	5.283	2.917	1.890
$\tilde{\beta}_6$	0.557	0.549	0.549	0.554	0.553	0.556	0.558	0.714	0.764
$\tilde{\beta}_7$	1.250	1.259	1.259	1.260	1.260	1.258	1.268	1.193	1.159
$\tilde{\beta}_8$	0.173	0.174	0.176	0.176	0.180	0.176	0.177	0.318	0.469
$\tilde{\beta}_9$	0.222	0.228	0.229	0.233	0.238	0.239	0.247	0.540	0.655
$\tilde{\beta}_{10}$	0.420	0.423	0.419	0.420	0.421	0.423	0.425	0.504	0.580
$\tilde{\sigma}^S$	66.15	66.10	66.09	66.25	66.37	66.67	67.67	178.5	242.06
$\tilde{\nu}^S$	1041.2	1038.6	1040.9	1039.9	1043.1	1047.4	1060.0	2583.6	3593
$\hat{\varepsilon}$	0.033	0.033	0.032	0.034	0.036	0.040	0.052	1.83	0.79

Tabelle 5.2.5 : Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 0.1 mittels skizzierten Daten für $\varepsilon = 0.2$.

Im Vergleich mit den Bayes-Regressionen für die gesamten Daten, liefern lediglich die Fälle $q = \{3, 5\}$ bezüglich der A-priori-Verteilung deutlich andere Resultate für die eingebetten Daten mit $\varepsilon = 0.2$.
Im nächsten Schritt wird die Varianz der A-priori-Verteilung in allen Fällen erneut reduziert und mit $\sigma^2 = 0.01$ modelliert. Diesmal werden die Resultate der gesamten Daten (Tabelle 5.2.6) und die dazugehörigen Resulate mit Hilfe von Skizzen mit $\varepsilon = 0.2$ verglichen.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.639	0.639	0.640	0.640	0.640	0.641	0.641	0.676	0.777
$\tilde{\beta}_2$	0.004	0.005	0.005	0.006	0.007	0.010	0.013	0.238	0.670
$\tilde{\beta}_3$	0.017	0.018	0.018	0.019	0.021	0.023	0.027	0.267	0.680
$\tilde{\beta}_4$	0.374	0.374	0.375	0.376	0.377	0.378	0.379	0.470	0.716
$\tilde{\beta}_5$	5.480	5.479	5.477	5.474	5.468	5.457	5.434	3.309	1.517
$\tilde{\beta}_6$	0.500	0.500	0.500	0.501	0.501	0.503	0.504	0.563	0.741
$\tilde{\beta}_7$	1.371	1.371	1.370	1.370	1.369	1.369	1.368	1.316	1.215
$\tilde{\beta}_8$	0.260	0.261	0.262	0.262	0.264	0.265	0.266	0.394	0.696
$\tilde{\beta}_9$	0.049	0.050	0.050	0.051	0.053	0.055	0.058	0.266	0.675
$\tilde{\beta}_{10}$	0.430	0.431	0.432	0.432	0.432	0.434	0.435	0.524	0.731
$\tilde{\sigma}^G$	10.08	10.08	10.08	10.07	10.08	10.08	10.09	24.26	42.79
$\tilde{\nu}^G$	1006.8	1006.8	1006.8	1006.9	1006.9	1007.2	1008.1	2424.7	4278.5

Tabelle 5.2.6 : Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 0.01 mittels gesamten Daten.

Während sich im Fall der gesamten Daten die Auswirkung der sehr starken A-priori-Annahme nur sehr geringfügig auswirken, ist im Fall der eingebetteten Daten der Effekt der A-priori-Wahl für verschiedene q-verallgemeinerte Normalverteilungen ersichtlicht. So wird schon ab einer Wahl von $q = 1.6$ eine deutlich schlechtere Anpassung der Regressionskoeffizienten erzielt, als im Fall der gesamten Daten. Dabei führt auch eine höhere Wahl der Burn-In-Phase zu den selben Ergebnissen. Wie bereits erwähnt, ergeben sich extrem kleine Randbereiche für diese niedrigen Varianzen der A-priori-Verteilungen. Wie in der Arbeit von Geppert et al. (2015) gezeigt wurde, besitzen die Likelihood der gesamten Daten und Likelihood der skizzierten Daten die im wesentlichen gleiche Form. Diese können sich aber um einen geringen Faktor verschieben. Dadurch kann es durch die genannten numerischen Probleme dazu führen, dass in den Fällen mit einer sehr kleinen A-priori-Varianz, auch eine geringe Verschiebung der Likelihood zu unterschiedlichen Ergebnissen führen kann.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.758	0.758	0.756	0.749	0.872	0.906	0.938	0.920	0.925
$\tilde{\beta}_2$	0.004	0.034	0.072	0.127	0.785	0.869	0.883	0.898	0.904
$\tilde{\beta}_3$	0.007	0.052	0.098	0.183	0.882	0.933	0.905	0.954	0.952
$\tilde{\beta}_4$	0.432	0.462	0.475	0.497	0.823	0.877	0.861	0.898	0.909
$\tilde{\beta}_5$	5.471	5.417	5.349	5.179	2.542	1.715	1.492	1.259	1.191
$\tilde{\beta}_6$	0.562	0.600	0.620	0.658	0.946	0.963	0.969	0.970	0.971
$\tilde{\beta}_7$	1.215	1.197	1.180	1.160	1.020	1.016	1.007	1.011	1.021
$\tilde{\beta}_8$	0.188	0.219	0.244	0.289	0.792	0.858	0.874	0.901	0.900
$\tilde{\beta}_9$	0.240	0.273	0.299	0.344	0.891	0.939	0.971	0.953	0.956
$\tilde{\beta}_{10}$	0.433	0.470	0.491	0.523	0.870	0.909	0.879	0.928	0.930
$\tilde{\sigma}^S$	66.504	67.354	68.66	72.9	223.1	271.3	283.7	298.323	302.5
$\tilde{\nu}^S$	1050.9	1066.3	1088.3	1157.4	3555.7	4315.7	4513.8	4738.1	4801.9
$\hat{\varepsilon}$	0.044	0.059	0.080	0.149	2.531	3.285	3.477	0.954	0.122

Tabelle 5.2.7 : Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 0.01 mittels skizzierten Daten für $\varepsilon = 0.2$.

Zu erkennen in Tabelle 5.2.7, ist dabei der Effekt, dass die resultierenden $\tilde{\beta}_i$ in allen Fällen für höhere Werte q näher an den A-priori-Erwartungswert 1 gezogen werden. Insgesamt ist noch einmal zu verdeutlichen, dass dies ein sehr unwahrscheinlich und konstruierter Fall ist. Eine Modellierung einer A-priori-Verteilung mit einer so geringen Varianz, wobei der Erwartungswert falsch gewählt wird ist eher unwahrscheinlich. Es soll dabei lediglich auf die sich möglicherweise ergebenen Probleme aufmerksam gemacht werden. Aus statististischer Sicht kann die Wahl einer geeigneter A-priori-Verteilung problematisch sein. Zum anderen können numerischen Probleme enstehen die aus Sicht der Informatik zu lösen gilt.
Werden die Erwartungswerte für die Parameter β passend gewählt, also näher an den wahren Werten, ergeben sich auch bessere Resultate. Als Fazit ergibt sich von daher, dass die Verwendung von ε-Unterraumeinbettungen auch mit informativen A-priori-Verteilungen gegeben ist. Das der erkennbare Effekt durch die Penalisierung der Regressionsparameter ausbleibt liegt vermutlich an der zu hohen Anzahl an Beobachtungen. Dieser Effekt tritt lediglich bei einer sehr ungünstigen Modellierung der A-priori-Verteilung auf, was

eher auf numerische Probleme zurückzuführen ist. Die Wahl des Erwartungswertes 1 für alle Regressionskoeffizienten mit zusätzlich einer sehr geringen Varianz passt nicht mit der tatsächlichen Modellierung der Daten überein. Die Resultate für die Wahl des A-priori-Erwartungswert $\mu = 10$ zeigen die selben Ergebnisse. Die Parameter liegen für eine sehr kleine Varianz und hohe q näher an dem A-priori Erwartungswert 10. Die genauen Resultate sind den Tabellen (Tabellen A.3-A.5) im Anhang zu entnehmen. Für größere Varianzen und/oder kleine Werte für q können mit Hilfe von Einbettungsmethoden hinreichend gute Ergebnisse auch bei sehr ungünstigen A-priori-Annahmen erzielt werden. Für die Modellierung des Datensatzes mit Standardabweichung $\xi = 1$ zeigen sich ebenfalls die Ergebnisse bestätigt, so dass diese hier nicht näher betrachetet werden. Insgesamt gilt, dass im Fall dieses gutmütigen Regressionsmodells, die Einbettungsmethoden nach Clarkson und Woodruff zu guten Resultaten führen. Lediglich bei der Wahl von sehr ungünstigen A-priori-Verteilungen können sich die erzielten Resultate unterscheiden.

5.3 ℓ_p-Regression für $p \in [1, 2)$ unter Verwendung von uninformativen A-priori-Dichten

Als nächstes werden die ε-Einbettungsmethoden unter Verwendung verschiedener p-verallgemeinerter Normalverteilungen für die Modellierung der Likelihood untersucht. Dies entspricht den entsprechenden Fällen der ℓ_p-Regression im bayesianischen. Um die Gültigkeit der ε-Unterraumeinbettung nach Clarkson und Woodruff (2013) zu gewährleisten, wird in diesem Fall der in Kapitel (4.3) beschriebene Algorithmus nach Zhang und Woodruff (2013) zur Modifizierung der Unterraumeinbettung angewendet. In Kapitel 4.4 konnte die Gültigkeit dieser Modifizierung bereits theoretisch, für eine gewisse Schranke, bewiesen werden. Aufgrund der Tatsache, dass sich diese Schranke mit einer Mindestanzahl an benötigten Zeilen n' der Skizze mit $\mathcal{O}\Big(d^{p+3}\log(d)^2\log(\frac{1}{\varepsilon})\frac{1}{\varepsilon^2}\Big)$ im Fall $d = 10$, als größer als die Anzahl Beobachtungen der Originaldaten n erweist, wird in diesem Kapitel untersucht, ob auch eine kleinere Skizzengrößen zu hinreichend guten Ergebnissen führen können. In diesem Kapitel werden die A-priori-Verteilungen zunächst als uninformativ angenommen, was im Fall der Implementierung in *OpenBUGS* bedeutet, dass die Dichte durch eine Konstante dargestellt wird und damit ein uneigentliche Dichte ist. Es wird zunächst wieder der Datensatz betrachtet, welcher in Kapitel 5.1 beschrieben ist. Dabei werden nun die Fälle betrachtet, dass die Likelihood p-verallgemeinert-normalverteilt modelliert ist mit $p = \{1, 1.2, 1.4, 1.6, 1.8\}$.
Sei für die Güte der Anpassung wieder $\hat{\varepsilon} = \frac{\|X\tilde{\beta}^S - Y\|_p}{\|X\tilde{\beta}^G - Y\|_p} - 1$ der geschätzte Wert der tatsächlich realisierten Abweichung zwischen Bayes-Regression mittels Gesamtdatensatz und skizzierten Daten, wobei $\|.\|_p$ diesmal der p-Vektornorm entspricht. Für die modifizierte Variante der Unterraumeinbettung (Kapitel 4.3) nach Woodruff und Zhang können, im Gegensatz zur regulären Variante, die Dimension der Skizze n', sowie der vorgegebene maximale Fehler der Unterraumeinbettung nach Clarkson und Woodruff, unabhängig voneinander gewählt werden. Dabei gilt nun, dass der Wert ε unbekannt ist und der Abschätzung $\mathcal{O}(d\log d)^{\frac{1}{p}}$ unterliegt. Sei zur besseren Unterscheidung im weiteren Kapitel der Parameter τ der vorgegebene Fehlerterm für die Skizze nach Clarkson und Woodruff. Dieser wird weiterhin in dieser Arbeit mit $\tau = \{0.1, 0.2\}$ gewählt. Da die vorgegebene Abweichung τ lediglich eine Rolle in den Konditionierungsschritten 1 und 2 des Algorithmus einimmt, wird vermutet das die Wahl dieses Parameter nun eine untergeordnete Rolle spielt.
Eine Untersuchung, wie sich die verschiedenen Einstellungskombinationen verhalten wird im Laufe des Kapitels analysiert. Als erstes wird die Einstellung $n' = 1024$ und $\tau = 0.1$

wie im vorherigen Kapitel verwendet. Die zugehörigen Resultate der Bayes-Regression mit den gesamten Daten, sowie der modifizierten Variante der Unterraumeinbettungen sind in Tabelle 5.3.1 dargestellt.

Original	$p = 1$	$p = 1.2$	$p = 1.4$	$p = 1.6$	$p = 1.8$
$\tilde{\beta}_1$	0.638	0.640	0.640	0.640	0.0639
$\tilde{\beta}_2$	-0.016	-0.011	-0.008	-0.003	0.000
$\tilde{\beta}_3$	0.025	0.021	0.020	0.018	0.018
$\tilde{\beta}_4$	0.365	0.368	0.370	0.372	0.373
$\tilde{\beta}_5$	5.476	5.477	5.478	5.478	5.478
$\tilde{\beta}_6$	0.497	0.497	0.498	0.499	0.499
$\tilde{\beta}_7$	1.363	1.365	0.1366	1.369	1.370
$\tilde{\beta}_8$	0.267	0.267	0.265	0.263	0.261
$\tilde{\beta}_9$	0.044	0.046	0.046	0.048	0.048
$\tilde{\beta}_{10}$	0.423	0.426	0.427	0.428	0.429
$\tilde{\sigma}^G$	11.413	10.747	10.388	10.198	10.104
$\|X\tilde{\beta}^G - Y\|_p$	80645.59	18317.63	6418.853	295.429	1618.05
Skizze					
$\tilde{\beta}_1$	0.667	0.664	0.628	0.592	0.695
$\tilde{\beta}_2$	-0.010	0.047	0.035	0.060	-0.017
$\tilde{\beta}_3$	0.016	0.028	0.034	-0.046	0.008
$\tilde{\beta}_4$	0.361	0.386	0.335	0.365	0.431
$\tilde{\beta}_5$	5.528	5.460	5.455	5.480	5.446
$\tilde{\beta}_6$	0.503	0.490	0.485	0.511	0.504
$\tilde{\beta}_7$	1.376	1.299	1.373	1.350	1.361
$\tilde{\beta}_8$	0.310	0.283	0.321	0.289	0.250
$\tilde{\beta}_9$	-0.013	-0.002	0.012	-0.010	0.096
$\tilde{\beta}_{10}$	0.474	0.464	0.416	0.434	0.423
$\tilde{\sigma}^S$	114.699	73.124	54.640	42.304	35.558
$\|X\tilde{\beta}^S - Y\|_p$	81078.86	18429.56	6445.84	2968.03	1626.54
$\hat{\varepsilon}$	0.005	0.006	0.004	0.007	0.005

Tabelle 5.3.1: Resultierende Schätzungen für die Regressionsparameter $\tilde{\beta}^G$ und $\tilde{\beta}^S$ für die verschiedenen l_p-Regressionen bei uninformativer A-priori-Verteilung unter Verwendung des Gesamten Datensatzes und der modifizierten Skizze nach Woodruff und Zhang mit $n' = 1024$ und τ=0.1.

Dabei ergibt sich für die Wahl von p der entsprechende ℓ_p-Regressionsfall. Zu erkennen ist, dass die Mittelwerte der resultierenden A-posteriori-Stichprobenwerte für die unterschiedlichen l_p-Regressionen die im wesentlich selben Schätzungen liefern. Lediglich der Parameter der Standardabweichung $\tilde{\sigma}^G$ unterscheidet sich. Dieser liegt im Fall $p = 1$ mit $\tilde{\sigma}^G = 11.413$ etwas höher als im Fall $p = 1.8$ mit $\tilde{\sigma}^G = 10.104$. Da in den simulierten Regressionsdaten die Zielvariable Y normalverteilt wurde, passen die Verteilungsannahmen der Likelihood für p nahe 2 besser und die Varianz reduziert sich. Um die Ergebnisse vergleichbar zu gestalten, werden aber dennoch in allen Fällen die selben Datensätze verwendet. Bei der Unterraumeinbettung ist die geschätzte Standardabweichung $\tilde{\sigma}^S$ mit $\tilde{\sigma}^S = 114.699$ für $p = 1$ wesentlich höher als $\tilde{\sigma}^S = 35.558$ im Fall $p = 1.8$. Für die Skizzen ergibt sich nun, dass die Standardabweichung der Skizze sich um den Faktor $\left(\frac{n}{n'}\right)^{1/p}$ erhöht. In diesem Fall für $n' = 1024$ wäre eine Erhöhung der Standardabweichung von $\{97.656, 66.796, 50.925, 41.549, 35.468\}$ zu erwarten. Dies stimmt ungefähr mit den resultierten Werten aus Tabelle 5.3.1 überein. Alle Regressionen liefern hinreichend genaue Anpassungen zu der jeweiligen p-Norm anhand der Werte $\hat{\varepsilon}$ zu erkennen. Diese liegen mit Werten zwischen $[0.004, 0.007]$ alle weit unter dem Wert 0.1, so dass in diesen Fällen von einer guten Schätzung mittels Skizze ausgegangen werden kann.
Wie sich die Standardabweichung und die Schätzungen für die tatsächlich beobachtete Anpassung $\hat{\varepsilon}$ für verschiedene Skizzengrößen verhält, wird in den Tabellen 5.3.2 und 5.3.3 ersichtlich. In Tabelle 5.3.2 sind die jeweilig geschätzten Standardabweichungen $\tilde{\sigma}^S$ dargestellt unter Verwendung von verschiedenen Skizzengrößen $n' \in [256, 2048]$ und $\tau = 0.1$ fest. Die Skizzengrößen über $n' = 2048$, also über ein fünftel des Gesamtdatensatzes zu betrachten wird in dieser Arbeit nicht verfolgt, da das Ziel der Einbettungsmethoden eine wesentliche Reduktion des Datensatzes liefern soll.

n'	$p = 1$	$p = 1.2$	$p = 1.4$	$p = 1.6$	$p = 1.8$
128	494.593	394.495	148.163	142.612	106.558
256	424.461	231.378	139.735	110.378	75.340
384	308.222	169.184	107.804	78.353	59.870
512	207.727	126.020	90.771	68.563	52.769
640	171.879	105.577	71.082	58.380	47.509
768	148.403	89.992	62.967	51.209	43.548
896	128.396	83.696	58.847	46.363	39.616
1024	114.699	73.124	54.640	42.304	35.558
1152	102.698	65.882	48.541	38.938	33.418
1280	91.699	60.745	45.246	37.160	31.578
1408	81.506	56.516	42.974	35.571	30.377
1536	75.152	53.499	40.245	33.668	28.233
1664	68.367	47.689	37.845	32.204	28.264
1792	62.911	43.652	35.206	30.126	26.552
1920	59.584	43.688	33.700	29.023	25.933
2048	55.607	40.014	32.726	27.648	24.353

Tabelle 5.3.2: Mittelwerte der A-posteriori-Samples für den Parameter σ unter Verwendung verschiedener p-verallgemeinerter Normalverteilungen bezüglich der Likelihood und verschiedenen Skizzengrößen n' bezüglich der modifzierten Variante der Einbettung nach Woodruff und Zhang mit $\tau = 0.1$ fest.

Des Weiteren werden die passenden Schätzwerte für die tatsächlich beobachtete Abweichung zwischen Gesamtdaten und Skizze $\hat{\varepsilon}$ betrachtet werden. Diese Werte sind in Tabelle 5.3.3 dargestellt. Es zeigt sich Insgesamt, dass mittels der modifizierten Variante der Unterraumeinbettung in allen Fällen gute Resultate in Form von niedrigen Abweichungen zur p-Norm ermittelt werden. Die Schätzungen $\hat{\varepsilon}$ werden zwar für größer werdende n' immer kleiner, was dafür spricht, dass sich die ermittelten Regressionskoeffizienten zwischen Schätzung mit Gesamtdaten und Unterraumeinbettung immer ähnlicher werden, allerdings lassen sich keine Unterschiede bezüglich der Wahl von p feststellen.

n'	$p = 1$	$p = 1.2$	$p = 1.4$	$p = 1.6$	$p = 1.8$
128	0.042	0.024	0.053	0.033	0.037
256	0.019	0.020	0.030	0.022	0.019
384	0.028	0.021	0.017	0.005	0.014
512	0.013	0.011	0.005	0.005	0.012
640	0.007	0.006	0.009	0.007	0.005
768	0.012	0.007	0.005	0.005	0.008
896	0.006	0.010	0.012	0.005	0.006
1024	0.005	0.006	0.004	0.007	0.005
1152	0.002	0.002	0.005	0.005	0.002
1280	0.012	0.003	0.004	0.003	0.001
1408	0.003	0.003	0.004	0.003	0.009
1536	0.002	0.006	0.004	0.004	0.003
1664	0.002	0.002	0.005	0.001	0.001
1792	0.004	0.005	0.005	0.002	0.004
1920	0.003	0.004	0.003	0.004	0.002
2048	0.004	0.001	0.004	0.002	0.004

Tabelle 5.3.3: Geschätzte tatsächlich beobachtete Abweichung $\hat{\varepsilon}$ zwischen der Bayes-Regression mittels gesamten Daten und modifizierter Unterraumeinbettung für verschiedene l_p-Regressionen mit uninformativer A-priori-Verteilung und verschiedenen Skizzengrößen mit $\tau = 0.1$

Die Schätzwerte scheinen Insgesamt zu den selben Ergebnissen zu führen, unterliegen aber durch die zufälligen Projektionen einer relativ hohen Streuung. Dennoch, liegen für sämtliche Skizzengrößen die resultierenden $\hat{\varepsilon} < 0.1$ unter der Schranke 0.1. Für die kleinere Wahl von $\tau = 0.2$ ergeben sich die Werte aus Tabelle 5.3.4.

n'	$p = 1$	$p = 1.2$	$p = 1.4$	$p = 1.6$	$p = 1.8$
128	0.040	0.064	0.023	0.023	0.055
256	0.017	0.058	0.024	0.016	0.015
384	0.025	0.017	0.010	0.008	0.015
512	0.015	0.017	0.005	0.004	0.016
640	0.016	0.009	0.005	0.005	0.009
768	0.007	0.004	0.004	0.005	0.011
896	0.004	0.005	0.009	0.009	0.010
1024	0.004	0.010	0.002	0.004	0.010
1152	0.006	0.001	0.003	0.006	0.008
1280	0.009	0.004	0.004	0.002	0.006
1408	0.003	0.004	0.006	0.004	0.007
1536	0.004	0.002	0.003	0.004	0.003
1664	0.004	0.002	0.004	0.001	0.003
1792	0.002	0.006	0.002	0.004	0.006
1920	0.002	0.002	0.004	0.001	0.003
2048	0.003	0.002	0.003	0.001	0.005

Tabelle 5.3.4: Geschätzte tatsächlich beobachtete Abweichung $\hat{\varepsilon}$ zwischen der Bayes-Regression mittels gesamten Daten und modifizierter Unterraumeinbettung für verschiedene l_p-Regressionsfälle mit uninformativer A-priori-Verteilung und verschiedenen Skizzengrößen mit $\tau = 0.2$

Dabei wird beim Vergleich von Verwendung $\tau = 0.1$ zu $\tau = 0.2$ deutlich, dass diese Wahl wohl keinen erheblichen Effekt auf die Güte der Anpassung besitzt. Um diese Vermutung zu unterlegen, werden die Resultate für die Einstellungskombinationen $n' = \{256, 1024\}$ und $\tau = \{0.1, 0.2\}$ für verschiedene Skizzen wiederholt. Zunächst wird die Skizzengröße $n' = 1024$ betrachtet. Dafür werden für jede ℓ_p-Regression die Ergebnisse 45 mal für verschiedene Skizzen nach der modifizierten Variante nach Woodruff und Zhang wiederholt und die resultierenden Schätzungen $\hat{\varepsilon}$ analysiert. Dafür werden alle resultierenden Schätzungen $\hat{\varepsilon}$ in Form von Boxplots in Abbildung 5.3.1 dargestellt.

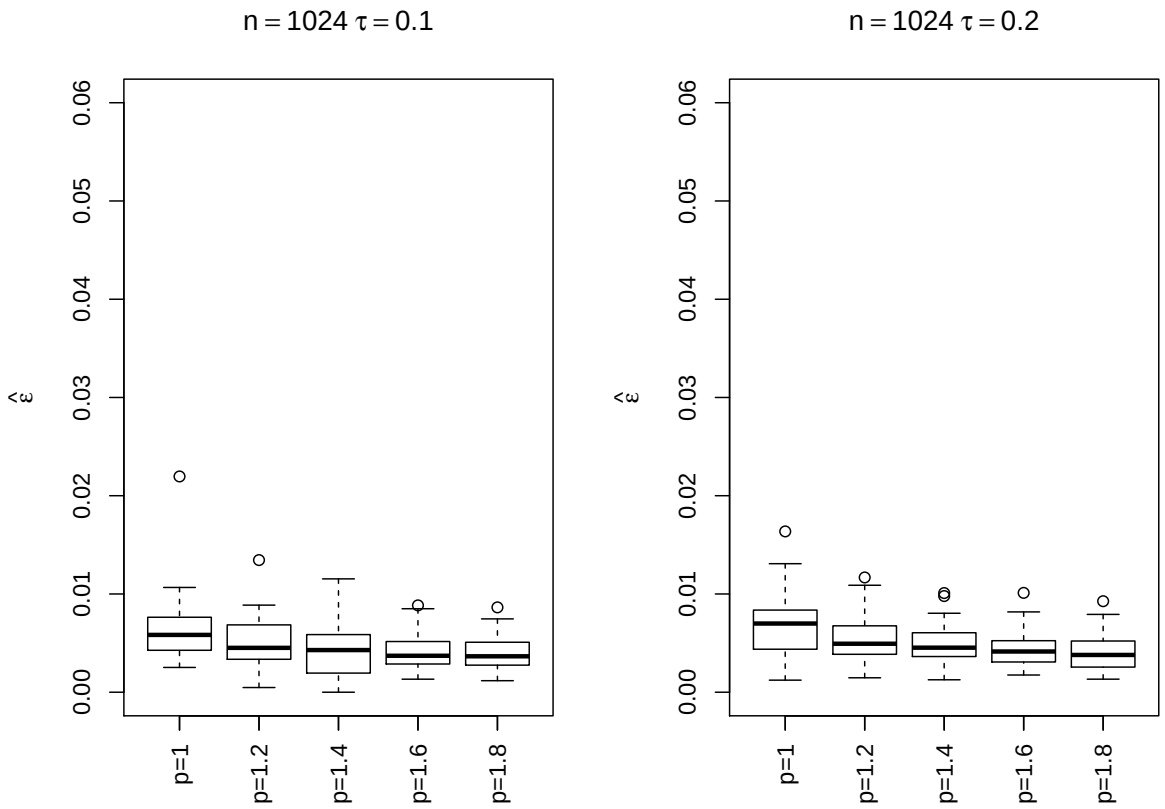

Abbildung 5.3.1 : Boxplots der geschätzten Werte ε für 45 wiederholt erzeugte Unterraumeinbettungen nach der Modifizierung von Woodruff und Zhang, mit Einstellung $n' = 1024$ und $\tau = \{0.1, 0.2\}$, im Fall von p-verallgemeinert normalverteilter Likelihood mit $p \in [1, 2)$.

Dabei scheint sich der Effekt bestätigt, dass die Wahl von τ bei diesen Einstellungen kein wesentlichen Einfluss auf die Güte hat. Zudem scheinen die Anpassungen in dieser Modellwahl für größere p bessere Ergebnisse zu liefern.
Die zugehörigen Kennzahlen der Boxplots sind der Tabelle 5.3.5 zu entnehmen.

$\tau = 0.1$	$p = 1$	$p = 1.2$	$p = 1.4$	$p = 1.6$	$p = 1.8$
Median	0.0058	0.0045	0.0043	0.0037	0.0037
Arith.Mittel	0.0064	0.0050	0.0041	0.0041	0.0040
Maximum	0.0220	0.0134	0.0115	0.0088	0.0086
$\tau = 0.2$					
Median	0.0070	0.0049	0.0045	0.0041	0.0038
Arith.Mittel	0.0069	0.0053	0.0050	0.0043	0.0042
Maximum	0.0164	0.0117	0.0101	0.0101	0.0093

Tabelle 5.3.5: Statistische Kennzahlen der geschätzten Werte $\hat{\varepsilon}$ bei verschiedenen l_p-Regressionen unter Verwendung von 45 Unterraumeinbettungen mit $n' = 1024$ und τ=$\{0.1,0.2\}$.

Der Maximalwert liegt dabei bei 0.022 für $p = 1$ und $\tau = 0.1$. In wesentlichen liefern alle Unterraumeinbettungen nach der modifizierten Variante für alle Regressionen mit $n' = 1024$ hinreichend gute Resultate in Form von kleinen $\hat{\varepsilon}$.
Als nächstes werden die Unterraumeinbettungen für kleinere Skizzengrößen $n' = 256$ untersucht. Hierfür wurden wie im eben beschriebenen Fall sämtliche ℓ_p-Regressionen mit verschiedenen Unterraumeinbettungen nach Woodruff und Zhang wiederholt und die resultierenden $\hat{\varepsilon}$ betrachet. Dafür seien zunächst wieder die resultierenden $\hat{\varepsilon}$ in Form von Boxplots in Abbildung 5.3.2 dargestellt.

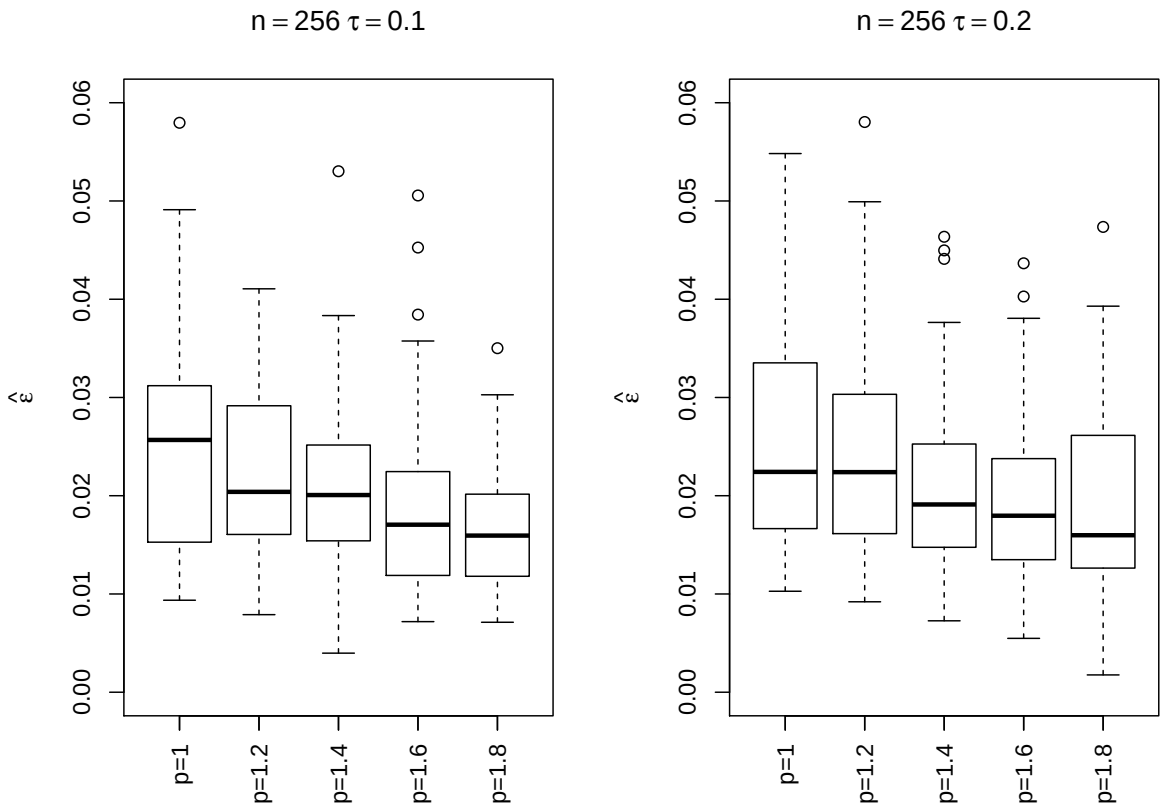

Abbildung 5.3.2 : Boxplots der geschätzten Werte ε für 50 wiederholt erzeugte Unterraumeinbettungen nach der Modifizierung von Woodruff und Zhang, mit Einstellung $n' = 256$ und $\tau = \{0.1, 0.2\}$, im Fall von p-verallgemeinert normalverteilter Likelihood mit $p \in [1, 2)$.

Diesmal sind die Schätzungen $\hat{\varepsilon}$, durch die kleineren Skizzengrößen und damit resultierende ungenauere Anpassungen, etwas höher als im Fall von $n' = 1024$. Dennoch zeigt sich bestätigt, dass die Wahl von τ bei der Unterraumeinbettung nach Clarkson und Woodruff in der Modifizierten Variante keinen erkennbaren Einfluss auf die Güte der Anpassung nimmt. Bestätigt wird dies durch einige ausgewählte Statistische Kennzahlen in Tabelle 5.3.6.

$\tau = 0.1$	$p = 1$	$p = 1.2$	$p = 1.4$	$p = 1.6$	$p = 1.8$
Median	0.0257	0.0204	0.0201	0.0170	0.0159
Arith.Mittel	0.0251	0.0228	0.0210	0.0193	0.0165
Maximum	0.0580	0.0411	0.0530	0.0506	0.0350
$\tau = 0.2$					
Median	0.0224	0.0224	0.0191	0.0180	0.0160
Arith.Mittel	0.0272	0.0251	0.0214	0.0196	0.0191
Maximum	0.0548	0.0580	0.0464	0.0437	0.0474

Tabelle 5.3.6: Statistische Kennzahlen der geschätzten Werte ε bei den die verschiedenen l_p-Regressionen mit 50 verschiedenen Unterraumeinbettungen für $n' = 256$ und $\tau = \{0.1, 0.2\}$.

Insgesamt werden in diesem gutmütigem Regressionsmodell für die modifizierte Unterraumeinbettung nach Woodruff und Zhang bei sämtlichen Einstellungen zufriedenstellende Ergebnisse, in Form von kleinen Schätzungen $\hat{\varepsilon}$, ermittelt. Als nächstes wird der Effekt von eingebauten Ausreisser in den Datensatz betrachtet. Dies soll den Effekt der robusteren Regression im Fall ℓ_1 verdeutlichen. Hierfür werden in die Variable X_5 10 Ausreißer eingebaut, so dass der Paramater β_5 an Einfluss verliert. Dies wird in Abbildung 5.3.3 dargestellt.

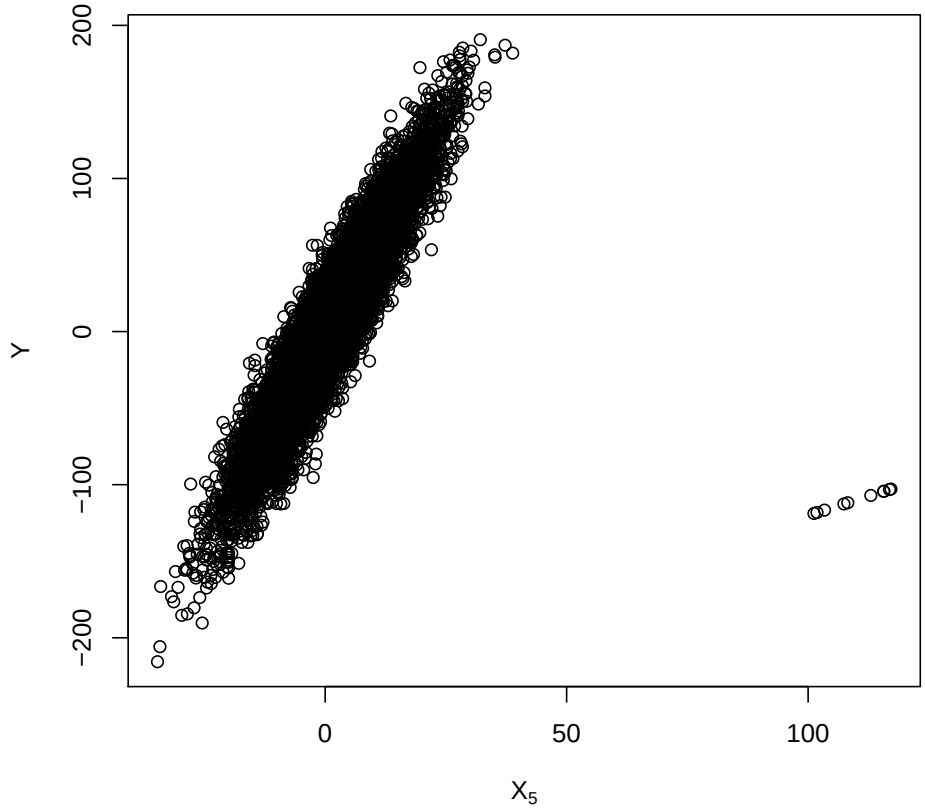

Abbildung 5.3.3: Streudiagramm der Variablen Y und X_5 mit 10 orthogonal zur Punktewolke eingebauten Ausreißern in der Variable X_5.

Zu erwarten ist, dass die Ausreißer auf die ℓ_1-Regression keinen Effekt ausüben, im Fall $\ell_{1.8}$ sich der geschätzte Parameter $\tilde{\beta}_5$ jedoch verkleinert. Anhand der Tabelle 5.3.6 zu erkennen, in der die A-posteriori-Resultate dargestellt sind, dass dieser Effekt in der Variable $\tilde{\beta}_5$ erkennbar ist.

Original	$p = 1$	$p = 1.2$	$p = 1.4$	$p = 1.6$	$p = 1.8$
$\tilde{\beta}_1$	0.638	0.639	0.638	0.639	0.638
$\tilde{\beta}_2$	-0.017	-0.012	-0.009	-0.005	-0.003
$\tilde{\beta}_3$	0.026	0.023	0.023	0.025	0.033
$\tilde{\beta}_4$	0.366	0.370	0.371	0.370	0.363
$\tilde{\beta}_5$	5.461	5.442	5.403	5.313	5.125
$\tilde{\beta}_6$	0.497	0.499	0.501	0.505	0.513
$\tilde{\beta}_7$	1.362	1.363	1.361	1.358	1.347
$\tilde{\beta}_8$	0.267	0.266	0.268	0.269	0.272
$\tilde{\beta}_9$	0.045	0.046	0.049	0.052	0.058
$\tilde{\beta}_{10}$	0.422	0.425	0.425	0.426	0.426
$\tilde{\sigma}^G$	12.42	12.54	13.63	15.86	19.33
$\tilde{\nu}^G$	87738.74	21379.69	8421.07	4582.3	3096.0
Skizze					
$\tilde{\beta}_1$	0.664	0.641	0.665	0.569	0.627
$\tilde{\beta}_2$	-0.071	-0.012	-0.031	0.043	0.013
$\tilde{\beta}_3$	-0.051	0.002	-0.053	0.038	0.028
$\tilde{\beta}_4$	0.395	0.392	0.400	0.384	0.354
$\tilde{\beta}_5$	5.493	5.407	5.441	5.403	5.369
$\tilde{\beta}_6$	0.546	0.444	0.536	0.478	0.537
$\tilde{\beta}_7$	1.389	1.317	1.367	1.353	1.355
$\tilde{\beta}_8$	0.186	0.292	0.299	0.254	0.331
$\tilde{\beta}_9$	0.088	0.129	0.050	0.052	0.076
$\tilde{\beta}_{10}$	0.502	0.472	0.417	0.390	0.391
$\tilde{\sigma}^S$	120.72	81.04	62.87	53.26	48.49
$\tilde{\nu}^S$	88763.14	21521.41	8452.65	4602.2	3127.47
$\hat{\varepsilon}$	0.012	0.006	0.004	0.004	0.010

Tabelle 5.3.7: Resultierende Schätzungen für die Regressionsparameter für die verschiedenen l_p-Regressionen bei uninformativer A-priori-Verteilung unter Verwendung des Originaldatensatz mit Ausreißern und den eingebetteten Datensätzen mit Ausreißern.

Dabei zeigt sich, dass bis auf gewisse Schwankungen in der Anpassung der Unterraumeinbettung, der Effekt den die Ausreisser ausüben, auch in den Unterraumeinbettungen

beibehalten wird. Dabei zeigten sich allerdings bei weiteren Untersuchungen, dass die Ausreisser dabei nicht beliebig groß gewählt werden dürfen, um gute Resultate zu resultieren. In der Arbeit von Geppert et al., (2015) wird gezeigt, dass die Güte der Anpassung von der Varianz der Daten abhängt. Sofern die Varianz der Daten durch Ausreißer steigt und sich die Daten nichtmehr adäquat für eine lineare Regression eignen, kann die algebraische Struktur nicht korrekt dargestellt werden.
Im hier betrachteten Fall ist dies jedoch nicht gegeben und es können mit der modifizierten Variante der Unterraumeinbettung nach Woodruff und Zhang hinreichend gute Ergebnisse ermittelt werden.
Die Ergebnisse der empirischen Untersuchung bestätigen die theoretischen Annahmen aus Kapitel 4.4. Dadurch zeigt sich, dass die Unterraumeinbettungen auch für robuste Bayes-Regression eignen und zugänglich gemacht werden konnten. Zudem konnte gezeigt werden, dass die theoretische Schranke für die Wahl der Skizzengröße um die Gültigkeit der Einbettung zu gewährleisten, in der Praxis nicht erreicht werden muss, um gute Resultate zu erzielen. Auch kleinere Skizzengrößen führen in den betrachteten Fällen zu hinreichend guten Ergebnissen. Abschließend werden die Fälle der ℓ_p-Regression noch um informative A-priori-Verteilungen erweitert.

5.4 ℓ_p-Regression für $p \in [1, 2)$ unter Verwendung informativer A-priori-Dichten

Abschliessend werden die soeben beschrieben ℓ_p-Regressionsfälle mit informativen A-priori-Verteilungen untersucht. Es zeigten sich bei der Untersuchung die Ergebnisse der empirischen Untersuchung aus Kapitel 5.2. und 5.3 bestätigt. Die Verwendung von A-priori-Verteilungen ist für die Regressionsparameter ist möglich, wirkt sich aber im Fall der großen Datensätze in den hier untersuchten Regressionen nicht deutlich auf die Ergebnisse aus. Repräsentativ werden die Ergebnisse für die ℓ_1-Regression und der $\ell_{1.8}$-Regression unter Verwendung verschiedener q-verallgemeinerter A-priori-Verteilungen dargestellt. Dabei werden für die modifizierte Variante der Unterraumeinbettung nach Woodruff und Zhang die Skizzengröße $n' = 1024$, sowie die Einstellung $\tau = 0.1$ angewendet. Für die q-verallgemeinerten A-priori-Verteilungen werden Erwartungswert $\mu = 1$ und Varianz $\sigma^2 = 1$ gewählt. Die Fälle $q < 2$ werden aus der Betrachtung entfernt, da sich diese wie in Kapitel 5.2 dargestellt, eher nicht zur Modellierung eignen. Die Ergebnisse der ℓ_1-Regression sind der Tabelle 5.4.1 zu entnehmen.

ℓ_1 Original	$q = 1$	$q = 1.2$	$q = 1.4$	$q = 1.6$	$q = 1.8$	$q = 2$
$\tilde{\beta}_1$	0.639	0.639	0.639	0.638	0.639	0.638
$\tilde{\beta}_2$	-0.016	-0.016	-0.016	-0.016	-0.016	-0.016
$\tilde{\beta}_3$	0.025	0.025	0.024	0.024	0.024	0.025
$\tilde{\beta}_4$	0.365	0.365	0.365	0.365	0.365	0.365
$\tilde{\beta}_5$	5.476	5.476	5.476	5.476	5.476	5.476
$\tilde{\beta}_6$	0.497	0.497	0.497	0.497	0.497	0.497
$\tilde{\beta}_7$	1.363	1.363	1.363	1.363	1.363	1.363
$\tilde{\beta}_8$	0.267	0.268	0.267	0.268	0.267	0.268
$\tilde{\beta}_9$	0.044	0.044	0.044	0.044	0.044	0.044
$\tilde{\beta}_{10}$	0.423	0.424	0.424	0.423	0.423	0.423
$\tilde{\sigma}^G$	11.411	11.413	11.413	11.413	11.412	11.413
$\tilde{\nu}^G$	80645.56	80645.51	80645.53	80645.54	80645.57	80645.55
Skizze						
$\tilde{\beta}_1$	0.552	0.551	0.551	0.551	0.551	0.551
$\tilde{\beta}_2$	0.001	0.001	0.001	0.001	0.001	0.001
$\tilde{\beta}_3$	0.083	0.083	0.083	0.083	0.083	0.083
$\tilde{\beta}_4$	0.344	0.343	0.343	0.343	0.343	0.343
$\tilde{\beta}_5$	5.451	5.450	5.450	5.449	5.449	5.448
$\tilde{\beta}_6$	0.567	0.566	0.566	0.566	0.566	0.566
$\tilde{\beta}_7$	1.404	1.405	1.405	1.405	1.406	1.406
$\tilde{\beta}_8$	0.268	0.268	0.268	0.268	0.268	0.268
$\tilde{\beta}_9$	0.062	0.063	0.062	0.062	0.062	0.062
$\tilde{\beta}_{10}$	0.489	0.489	0.489	0.489	0.488	0.488
$\tilde{\sigma}^S$	109.736	109.765	109.782	109.795	109.770	109.742
$\tilde{\nu}^S$	81578.67	81583.48	81582.41	81584.87	81583.82	81588.48
$\hat{\varepsilon}$	0.0116	0.0116	0.0116	0.0116	0.0116	0.0117

Tabelle 5.4.1: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei ℓ_1-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 1, für die Gesamtdaten, sowie die modifizierte Unterraumeinbettung nach Woodruff und Zhang mit $\tau = 0.1$ und $n' = 1024$.

Dabei werden zwar unabhängig von der Wahl von q die selben guten Schätzungen

ermittelt. Der Einfluss von der Wahl der A-priori-Verteilung ist allerdings nicht direkt ersichtlich. Jediglich die Werte von $\tilde{\beta}_5$ lassen erahnen, dass die Werte für eine hohe Wahl von q an den A-priori-Erwartungswert gezogen werden. Dieser Effekt wurde in Kapitel 5.2 bereits erläutert. Die Ergebnisse zeigen sich auch in den Modellen der $\ell_{1.8}$-Regression, wie in Tabelle 5.4.2 dargestellt.
Die modifizierte Variante der Unterraumeinbettung nach Woodruff und Zhang zeigt auch in den Fällen widersprüchlicher A-priori-Verteilung und Verteilung der Likelihood keine Auffälligkeiten. Die Ergebnisse aus den vorherigen Kapiteln scheinen sich damit zu bestätigen. Es lassen sich mit der in dieser Arbeit vorgestellten Modifizierung gute Resultate hinsichtlich den untersuchten Fälle ermitteln. Dadurch ist die Verwendung von Unterraumeinbettungen bei der Bayes-Regression für die verschiedenen hier betrachteten Abstandsnormen gegeben. Die Anwendung von informativen A-priori-Verteilungen zeigte sich in den Fällen von sehr großen Datenmengen als eher schwierig. Um einen erkennbaren Einfluss zu nehmen mussten die A-priori-Varianzen sehr gering gewählt werden was wiederum numerische Schwierigkeiten mit sich bringen kann. Für uninformatives A-priori-Wissen und unter milderen Modellannahmen konnte die Anwendbarkeit von verschiedenen Abstandsnormen gezeigt werden.

$\ell_{1.8}$ Original	$q = 1$	$q = 1.2$	$q = 1.4$	$q = 1.6$	$q = 1.8$	$q = 2$
$\tilde{\beta}_1$	0.639	0.639	0.639	0.639	0.639	0.639
$\tilde{\beta}_2$	0.001	0.000	0.000	0.000	0.000	0.000
$\tilde{\beta}_3$	0.017	0.017	0.017	0.017	0.017	0.017
$\tilde{\beta}_4$	0.373	0.373	0.373	0.373	0.373	0.373
$\tilde{\beta}_5$	5.479	5.479	5.479	5.479	5.479	5.478
$\tilde{\beta}_6$	0.499	0.499	0.499	0.499	0.499	0.499
$\tilde{\beta}_7$	1.370	1.370	1.371	1.371	1.371	1.371
$\tilde{\beta}_8$	0.262	0.262	0.262	0.262	0.262	0.262
$\tilde{\beta}_9$	0.049	0.048	0.048	0.048	0.048	0.048
$\tilde{\beta}_{10}$	0.429	0.429	0.429	0.429	0.429	0.429
$\tilde{\sigma}^G$	10.106	10.105	10.105	10.105	10.105	10.105
$\tilde{\nu}^G$	1618.05	1618.05	1618.05	1618.05	1618.05	1618.05
Skizze						
$\tilde{\beta}_1$	0.607	0.605	0.605	0.604	0.606	0.605
$\tilde{\beta}_2$	0.030	0.030	0.032	0.030	0.030	0.032
$\tilde{\beta}_3$	0.022	0.019	0.021	0.020	0.021	0.018
$\tilde{\beta}_4$	0.362	0.364	363	0.365	0.363	0.364
$\tilde{\beta}_5$	5.510	5.512	5.510	5.509	5.509	5.510
$\tilde{\beta}_6$	0.523	0.525	0.524	0.523	0.523	0.524
$\tilde{\beta}_7$	1.342	1.344	1.344	1.345	1.345	1.345
$\tilde{\beta}_8$	0.207	0.207	0.206	0.204	0.205	0.206
$\tilde{\beta}_9$	0.089	0.090	0.088	0.089	0.089	0.089
$\tilde{\beta}_{10}$	0.436	0.434	0.434	0.433	0.434	0.434
$\tilde{\sigma}^S$	35.913	35.907	35.916	35.913	35.908	35.908
$\tilde{\nu}^S$	1625.348	1625.796	1625.426	1625.598	1625.413	1625.543
$\hat{\varepsilon}$	0.045	0.0047	0.0045	0.0046	0.0045	0.0046

Tabelle 5.4.2: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei $\ell_{1.8}$-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 1, für die Gesamtdaten, sowie die modifizierte Unterraumeinbettung nach Woodruff und Zhang mit $\tau = 0.1$ und $n' = 1024$.

6 Zusammenfassung und Ausblick

Ziel dieser Arbeit ist die Untersuchung der Anwendbarkeit von ε-Unterraumeinbettungen bei Bayes-Regression bezüglich verschiedenen Abstandsnormen.
Die Verwendung von ε-Unterraumeinbettungen führt im Fall von hochdimensionalen Datensätzen zu einer Reduktion der Daten, so dass bis auf einen kleinen kontrollierbaren Fehler, die Information der Regressionkoeffizienten beibehalten wird. Dafür wird auf den Ergebnissen von Geppert et al. (2015) aufgebaut. Ausgehend von der Anwendbarkeit bei linearer Bayes-Regression im ℓ_2-Fall mit uninformativen A-priori-Annahmen in Geppert et al., werden diese Ergebnisse in dieser Arbeit für die Fälle der ℓ_p ($p \in [1, 2)$) erweitert. Der erste Teil der Untersuchung hat, mit Hilfe verschiedener Simulationen und Modellannahmen gezeigt, dass auch die Wahl von informativen A-priori-Verteilungen aus der Klasse der q-verallgemeinerter Normalverteilungen bei Bayes-Regression im ℓ_2-Fall zu hinreichend guten Ergebnissen führen. Die Verwendung von ε-Unterraumeinbettungen kann in diesen Fällen bedingt gezeigt werden. Des Weiteren wurde auf numerische Probleme aufmerksam gemacht, die bei der ungünstigen Modellierung der A-priori-Verteilung auftreten können. Der zweite Teil der Untersuchung beschäftigt sich mit der Erweiterung auf die Fälle der ℓ_p-Regressionen $p \in [1, 2)$. Hierfür wird in Kapitel (4.4) eine theoretische untere Schranke für die benötigte Anzahl an Zeilen der Skizze ermittelt, für welche die Güte der Einbettungsmethode nach Woodruff und Zhang (2013) gegeben ist. Dass in der Praxis auch für wesentlich kleinere Skizzengrößen, als die vorgegebene Schranke, hinreichend gute Ergebnissen erzielt werden können, wird zudem anhand von Simulationen in Kapitel 5.3 unterlegt. Dabei wird der Fall von uninformativem A-priori-Verteilungen betrachet. In Kapitel 5.4 werden diese Ergebnisse für informative A-priori-Verteilungen aus der Klasse der verallgemeinerten Normalverteilungen erweitert.
Für weitere Verallgemeinerungen gibt es noch eine Vielzahl an zu untersuchenden Fällen. Eine theoretische Verallgemeinerung der Ergebnisse für beliebige A-priori-Dichten wäre wünschenswert. Zudem könnten die Untersuchungen auf die Klasse der generalisierten Modelle ausgeweitet werden. So ist eine Untersuchung hinsichtlich logistischer Regression oder der Poisson-Regression denkbar.
Des Weiteren könnten weitere algorithmische Ansätze und Methoden untersucht werden. In der Methodik von Cohen und Peng (2015) wird ebenfalls ein Samplingalgorithmus für allgemeinere p-Normen unter Verwendung von approximativen *Lewis-weights* statt Leverage-Scores verwendet. So könnte eine empirische Vergleichsstudie mit dem Ansatz

nach Cohen und Peng (2015) und der in dieser Arbeit verwendeten Modifizierung nach Woodruff und Zhang (2013), der für den ℓ_p-Fall verwendet wurde, interessant sein. Insgesamt gibt es noch zahlreiche Möglichkeiten an interessanten Erweiterungen und Verallgemeinerungen in diesem Forschungsgebiet.

Literaturverzeichnis

Clarkson, K. L., Drineas, P., Magdon-Ismail, M., Mahoney, M. W., Meng, X., Woodruff, D. P. (2013): *The Fast Cauchy Transform and Faster Robust Linear Regression*, In: Proceedings of SODA, pp. 466-477.

Clarkson, K. L., Woodruff, D. P. (2013): *Low Rank Approximation and Regression in Input Sparsity Time*, In: Proceedings of STOC, pp. 81-90.

Cohen, M. B., Peng, R. (2015): ℓ_p *Row Sampling by Lewis Weights*, In: Proceedings of STOC, pp. 183-192.

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., Mahoney, M. W. (2008): *Sampling Algorithms and Coresets for* ℓ_p *Regression*, In: Proceedings of SODA, pp. 932-941.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, B., Vehtari, A. und Rubin, D. B. (2004): *Bayesian Data Analysis* Texts in Statistical Science, 2. Aufl., Hall/CRC, Boca Raton.

Gelman, A., Ligges, U., Sturtz, S., (2005): R2WinBUGS: *A Package for Running WinBUGS from R*, R-package, Journal of Statistical Software, 12(3), 1-16.

Geppert, L.,Ickstadt, K., Munteanu, A., Quedenfeld, J. und Sohler, C. (2015): *Random Projections for Bayesian Regression*, Statistics and Computing (Online First), Springer.

Geppert, L.,Ickstadt, K., Munteanu, A., Quedenfeld, J. und Sohler, C. (2015): *RaProR*: *Random Projections for Bayesian Regression*, R-package, Version 1.0, http://ls2-www.cs.uni-dortmund.de/projekte/RaProrR/.

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

Kotz, S., Balakrishnan, N., Johnson, N. L.(1970): *Continuous univariate distributions*, Band 1, 2. Aufl., Wiley Series in Probability and Statistics.

Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. (2000) OpenBUGS – *a Bayesian modelling framework: concepts, structure, and extensibility.* Statistics and Computing, 10:325–337, Version 3.2.3 rev1012.

Neal, R. M. (2012): *MCMC using Hamiltonian dynamics*, Version 1, veröffentlicht als Kapitel 5 im Handbook of Markov Chain Monte Carlo, 2011.

R Core Team (2015): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Wien, Version R 3.2.2.

Teschl, G., Teschl, S. (2007): *Mathematik für Informatiker*, Analysis und Statistik, 2. Aufl., Band 2, Springer Verlag Berlin Heidelberg.

Woodruff, D. P., Zhang, Q. (2013): *Subspace Embeddings and ℓ_p-Regression Using Exponential Random Variables*, In: Proceedings of COLT pp.546.567.

Anhang

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.731	0.726	0.723	0.719	0.720	0.721	0.721	0.719	0.703
$\tilde{\beta}_2$	-0.010	-0.009	-0.009	-0.004	-0.008	-0.010	-0.011	-0.012	-0.004
$\tilde{\beta}_3$	-0.021	-0.026	-0.025	-0.026	-0.025	-0.026	-0.028	-0.024	0.016
$\tilde{\beta}_4$	0.404	0.406	0.405	0.404	0.403	0.400	0.402	0.398	0.356
$\tilde{\beta}_5$	5.475	5.473	5.476	5.474	5.470	5.469	5.464	5.423	5.061
$\tilde{\beta}_6$	0.539	0.535	0.532	0.530	0.530	0.531	0.531	0.531	0.544
$\tilde{\beta}_7$	1.272	1.274	1.281	1.280	1.281	1.281	1.283	1.288	1.287
$\tilde{\beta}_8$	0.159	0.157	0.156	0.157	0.155	0.154	0.153	0.149	0.127
$\tilde{\beta}_9$	0.216	0.219	0.212	0.213	0.212	0.210	0.212	0.217	0.242
$\tilde{\beta}_{10}$	0.409	0.410	0.408	0.409	0.407	0.405	0.406	0.401	0.386
$\tilde{\sigma}^S$	66.08	66.08	66.02	66.04	65.99	65.99	66.00	66.172	71.207
$\tilde{\nu}^S$	1037.2	1037.3	1035.3	1035.1	1035.1	1035.0	1035.5	1037.4	1121.6

A.1: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 1 mittels skizzierten Daten für $\varepsilon = 0.2$.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.626	0.630	0.634	0.638	0.645	0.653	0.678	0.912	0.921
$\tilde{\beta}_2$	-0.035	-0.024	-0.014	-0.003	0.020	0.051	0.128	0.811	0.846
$\tilde{\beta}_3$	-0.008	0.001	0.007	0.019	0.038	0.066	0.133	0.810	0.839
$\tilde{\beta}_4$	0.344	0.349	0.355	0.360	0.367	0.381	0.406	0.798	0.842
$\tilde{\beta}_5$	5.489	5.478	5.465	5.437	5.376	5.265	4.969	1.558	1.286
$\tilde{\beta}_6$	0.439	0.448	0.453	0.460	0.471	0.484	0.521	0.864	0.882
$\tilde{\beta}_7$	1.356	1.350	1.347	1.345	1.347	1.343	1.340	1.165	1.140
$\tilde{\beta}_8$	0.295	0.302	0.308	0.311	0.316	0.326	0.343	0.766	0.820
$\tilde{\beta}_9$	0.054	0.060	0.065	0.076	0.097	0.118	0.178	0.799	0.838
$\tilde{\beta_{10}}$	0.438	0.445	0.451	0.456	0.463	0.475	0.502	0.838	0.864
$\tilde{\sigma}^S$	31.04	31.05	31.09	31.18	31.43	32.22	36.13	136.37	144.67
$\tilde{\nu}^S$	1010.8	1010.1	1010.1	1010.9	1016.5	1037.6	1154.6	4346.5	4625.7

A.2: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 1 mittels skizzierten Daten für $\varepsilon = 0.1$.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.638	0.638	0.638	0.638	0.638	0.638	0.639	0.644	4.619
$\tilde{\beta}_2$	0.003	0.003	0.003	0.003	0.004	0.004	0.004	0.010	4.505
$\tilde{\beta}_3$	0.016	0.016	0.016	0.016	0.017	0.017	0.017	0.023	4.437
$\tilde{\beta}_4$	0.373	0.373	0.373	0.373	0.374	0.374	0.374	0.380	4.583
$\tilde{\beta}_5$	5.480	5.480	5.480	5.480	5.480	5.480	5.480	5.482	6.380
$\tilde{\beta}_6$	0.499	0.499	0.499	0.498	0.499	0.500	0.499	0.505	4.617
$\tilde{\beta}_7$	1.372	1.372	1.372	1.372	1.373	1.373	1.373	1.377	4.791
$\tilde{\beta}_8$	0.260	0.260	0.260	0.260	0.260	0.260	0.261	0.266	4.468
$\tilde{\beta}_9$	0.048	0.049	0.049	0.049	0.049	0.049	0.050	0.055	4.468
$\tilde{\beta}_{10}$	0.430	0.430	0.430	0.430	0.430	0.430	0.431	0.436	4.557
$\tilde{\sigma}^G$	10.08	10.08	10.07	10.07	10.07	10.07	10.07	10.08	126.160
$\tilde{\nu}^G$	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1007	12613.3

A.3: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 10 und Varianz 1 mittels gesamten Daten.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.617	0.619	0.619	0.618	0.619	0.622	0.623	0.674	4.889
$\tilde{\beta}_2$	-0.038	-0.038	-0.039	-0.039	-0.037	-0.036	-0.031	0.034	4.929
$\tilde{\beta}_3$	-0.014	-0.016	-0.014	-0.012	-0.013	-0.009	-0.005	0.053	4.794
$\tilde{\beta}_4$	0.339	0.339	0.339	0.340	0.341	0.344	0.347	0.391	4.773
$\tilde{\beta}_5$	5.494	5.495	5.496	5.497	5.498	5.500	5.497	5.489	5.989
$\tilde{\beta}_6$	0.432	0.431	0.432	0.433	0.431	0.437	0.439	0.498	4.988
$\tilde{\beta}_7$	1.364	1.364	1.366	1.367	1.371	1.367	1.372	1.415	5.102
$\tilde{\beta}_8$	0.290	0.292	0.293	0.293	0.296	0.298	0.299	0.349	4.822
$\tilde{\beta}_9$	0.049	0.049	0.050	0.050	0.055	0.051	0.056	0.111	4.750
$\tilde{\beta}_{10}$	0.433	0.435	0.434	0.434	0.436	0.438	0.440	0.489	4.851
$\tilde{\sigma}^S$	31.02	31.03	31.03	31.04	31.037	31.03	31.05	31.58	149.98
$\tilde{\nu}^S$	1011.6	1011.8	1011.7	1011.5	1011.8	1011.2	1010.8	1017.6	13547.3

A.4: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 10 und Varianz 1 mittels skizzierten Daten für $\varepsilon = 0.1$.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.722	0.725	0.728	0.730	0.731	0.738	0.748	1.433	4.614
$\tilde{\beta}_2$	-0.012	-0.009	-0.006	-0.004	0.006	0.012	0.018	1.002	4.653
$\tilde{\beta}_3$	-0.032	-0.025	-0.023	-0.019	-0.011	-0.002	0.012	1.205	4.840
$\tilde{\beta}_4$	0.407	0.408	0.409	0.413	0.419	0.435	0.446	1.608	4.970
$\tilde{\beta}_5$	5.489	5.489	5.489	5.495	5.497	5.503	5.503	5.726	6.480
$\tilde{\beta}_6$	0.530	0.533	0.536	0.539	0.545	0.554	0.568	1.553	4.859
$\tilde{\beta}_7$	1.288	1.291	1.292	1.297	1.301	1.301	1.315	2.012	4.882
$\tilde{\beta}_8$	0.153	0.155	0.161	0.163	0.169	0.178	0.190	1.127	4.635
$\tilde{\beta}_9$	0.211	0.211	0.216	0.215	0.220	0.231	0.240	1.201	4.718
$\tilde{\beta}_{10}$	0.411	0.407	0.411	0.414	0.423	0.434	0.445	1.550	4.912
$\tilde{\sigma}^S$	66.04	66.00	66.03	66.08	66.11	66.21	66.47	149.43	149.9
$\tilde{\nu}^S$	1035.0	1034.7	1035.0	1034.6	1034.8	1037.7	1040.2	3245.4	13316.6

A.5: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 10 und Varianz 1 mittels skizzierten Daten für $\varepsilon = 0.2$.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.638	0.638	0.638	0.638	0.638	0.638	0.638	0.638	0.638
$\tilde{\beta}_2$	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
$\tilde{\beta}_3$	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016	0.016
$\tilde{\beta}_4$	0.373	0.373	0.373	0.373	0.373	0.373	0.373	0.373	0.373
$\tilde{\beta}_5$	5.480	5.480	5.480	5.480	5.480	5.480	5.480	5.480	5.480
$\tilde{\beta}_6$	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499
$\tilde{\beta}_7$	1.372	1.372	1.372	1.372	1.372	1.372	1.372	1.372	1.372
$\tilde{\beta}_8$	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260
$\tilde{\beta}_9$	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048
$\tilde{\beta}_{10}$	0.430	0.430	0.430	0.430	0.430	0.430	0.430	0.430	0.430
$\tilde{\sigma}^G$	10.074	10.073	10.074	10.074	10.075	10.075	10.073	10.074	10.074
$\tilde{\nu}^G$	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8	1006.8

A.6: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 10 mittels gesamten Daten.

q	0.6	1	1.2	1.4	1.6	1.8	2	3	5
$\tilde{\beta}_1$	0.619	0.619	0.616	0.616	0.617	0.616	0.616	0.617	0.618
$\tilde{\beta}_2$	-0.040	-0.039	-0.040	-0.040	-0.040	-0.041	-0.041	-0.042	-0.039
$\tilde{\beta}_3$	-0.013	-0.018	-0.017	-0.017	-0.015	-0.016	-0.017	-0.016	-0.014
$\tilde{\beta}_4$	0.339	0.338	0.339	0.340	0.339	0.340	0.340	0.338	0.338
$\tilde{\beta}_5$	5.492	5.493	5.496	5.496	5.494	5.496	5.496	5.496	5.493
$\tilde{\beta}_6$	0.433	0.429	0.429	0.429	0.431	0.428	0.429	0.429	0.431
$\tilde{\beta}_7$	1.364	1.364	1.363	1.363	1.363	1.364	1.363	1.364	1.363
$\tilde{\beta}_8$	0.291	0.292	0.293	0.293	0.290	0.292	0.292	0.292	0.290
$\tilde{\beta}_9$	0.049	0.050	0.047	0.047	0.049	0.046	0.046	0.046	0.048
$\tilde{\beta}_{10}$	0.433	0.435	0.433	0.433	0.432	0.434	0.434	0.433	0.433
$\tilde{\sigma}^S$	31.034	31.025	31.031	31.030	31.029	31.027	31.028	31.026	31.024
$\tilde{\nu}^G$	1011.6	1012.1	1012.1	1012.1	1011.8	1012.1	1012.1	1012.2	1011.7
$\hat{\varepsilon}$	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005

A.7: Resultierende Mittelwerte der 10000 A-posteriori-Samples bei Bayes-Regression mit entsprechender q-verallgemeinerter Normalverteilungsannahme bezüglich der A-priori-Verteilung mit Erwartungswert 1 und Varianz 10 mittels skizzierten Daten für $\varepsilon = 0.1$.

Eidesstattliche Versicherung

_____ _____

Name, Vorname Matr.-Nr.

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

_____

_____

_____

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 Unterschrift

*Nichtzutreffendes bitte streichen

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 Unterschrift