Albert-Ludwigs-Universität Freiburg

Bachelorarbeit

Verzerrung der Inferenz bei Verwendung gemischter Modelle in latenten Repräsentationen

Autor:
Yannick Bantel

Professor:
Prof. Dr. Harald Binder

Betreuer:
Clemens Schächter

Abgabedatum:
26. Juni 2024

universität freiburg

Inhaltsverzeichnis

1 Zusammenfassung i

Einleitung **iii**
0.1 Motivation iii

Theoretische Grundlagen **v**
1 Einführung in Variational Autoencoder (VAE) v
1.1 Strucktur des VAEs vi
1.2 Training VAE vii
2 Grundlagen Gemischte Modelle x
3 Likelihood Inferenz und Verzerrung xii
3.1 Likelihood Berechnung gemischter Modelle xii
3.2 Likelihood-Ratio-Test xv

Empirische Ergebnisse **xvii**
1 gemischte Modelle auf simulierten Daten xvii
2 gemischte Modelle in latenten Repräsentationen xix
2.1 Der Datensatz xx
2.2 Gemischtes Modell auf latenter Datenwolke xxi
3 Modellierungstechniken xxii
4 Analysemethoden xxii
4.1 Post-Selection-Inferenz xxii
5 Experimente und Ergebnisse xxii
5.1 Methodik xxii
5.2 Experimentelles Design xxiii
5.3 Durchführung xxiii
5.4 Analyse der Ergebnisse xxiii
5.5 Gemischtes Modell auf latenter Datenwolke mit separatem Training xxiv
5.6 Gemischtes Modell auf latenter Datenwolke mit gleichzeitigem Training xxiv

Diskussion und Fazit **xxv**
1 Interpretation der Ergebnisse xxv
2 Vergleich mit bestehenden Arbeiten xxv
3 Limitationen und Herausforderungen xxv

Fazit **xxvii**

Anhang **xxix**

Appendix **xxxi**
1 Supporting Data xxxi
2 Some Code Listings xxxi

1 Zusammenfassung

Eine kurze Zusammenfassung der Arbeit auf Deutsch.

Abstract

A brief abstract of this thesis in English.

Einleitung

In der modernen Datenanalyse spielen gemischte Modelle eine zentrale Rolle, da sie es ermöglichen, sowohl feste als auch zufällige Effekte zu berücksichtigen. Dies macht sie besonders in den Bereichen der Biostatistik, der Sozialwissenschaften und der ökonomischen Modellierung populär.
Im Rahmen dieser Arbeit werden gemischte Modelle auf die Analyse medizinischer Daten angewendet. Mit dem Aufkommen von Big Data und komplexen Datenstrukturen hat sich der Fokus zunehmend auf die effiziente und genaue Extraktion von Informationen aus großen und oft unübersichtlichen Datensätzen verlagert.
In diesem Zusammenhang gewinnen latente Repräsentationen an Bedeutung, da sie es ermöglichen, inhärente Strukturen innerhalb der Daten zu identifizieren und zu nutzen, um tiefere Einblicke zu gewinnen.

Die Integration von gemischten Modellen in latente Repräsentationen birgt jedoch das Risiko einer Verzerrung der Inferenzergebnisse, was die Genauigkeit und Zuverlässigkeit der aus den Daten gezogenen Schlussfolgerungen erheblich beeinträchtigen kann.
Die vorliegende Arbeit widmet sich der Untersuchung von Verzerrungen, die bei der Anwendung gemischter Modelle auf latente Repräsentationen auftreten können. Das Ziel dieser Arbeit ist es, die Mechanismen zu verstehen, die zu diesen Verzerrungen führen, sowie Methoden zu entwickeln, um ihre Auswirkungen zu minimieren.

Das Problem der Verzerrung ist von besonderer Relevanz, da eine fehlerhafte Inferenz zu Fehlentscheidungen führen kann, die in praktischen Anwendungen schwerwiegende Konsequenzen haben können.
Die Arbeit zielt darauf ab, durch eine sorgfältige Analyse und Bewertung von gemischten Modellansätzen in Verbindung mit latenten Repräsentationen einen Beitrag zur Verbesserung der Modellgenauigkeit und der Zuverlässigkeit von Inferenzschlüssen zu leisten.

Die Arbeit ist in mehrere Teile gegliedert, die zunächst die theoretischen Grundlagen von gemischten Modellen und latenten Repräsentationen behandeln. Im Anschluss erfolgt eine Diskussion der Methoden zur Messung und Korrektur von Verzerrungen.
Im Anschluss werden die zuvor theoretisch erörterten Konzepte anhand von empirischen Studien praktisch angewendet und evaluiert. Auf Basis der gewonnenen Erkenntnisse werden abschließend Empfehlungen für die Anwendung dieser Techniken in Forschung und Praxis gegeben.

0.1 Motivation

Theoretische Grundlagen

Im Vorfeld der Erörterung der Methodik dieser Arbeit ist eine theoretische Aufarbeitung der behandelten Themen unabdingbar. Dieses Kapitel widmet sich den theoretischen Grundlagen, die für das Verständnis und die Analyse von Verzerrungen in der Inferenz erforderlich sind, wenn gemischte Modelle in latenten Repräsentationen zum Einsatz kommen.

Das vorliegende Kapitel beginnt mit einer detaillierten Einführung in Variational Autoencoder (VAE). VAEs sind generative Modelle, die es ermöglichen, hochdimensionale Daten in niedrigdimensionale latente Repräsentationen zu überführen. Die Modelle reduzieren die Komplexität der Daten und sind in der Lage, sowohl die zugrunde liegende Struktur der Daten zu erfassen als auch neue Daten zu generieren, die ähnliche Merkmale wie die Trainingsdaten aufweisen. Die Architektur und das Training von VAEs werden detailliert beschrieben, um ein fundiertes Verständnis ihrer Funktionsweise zu vermitteln.

Im zweiten Teil des Kapitels erfolgt eine Behandlung von gemischten Modellen. Die Modelle kombinieren feste und zufällige Effekte, um die Variabilität in den Daten zu erfassen. Ihre Anwendung ist insbesondere bei der Analyse longitudinaler und Cluster-Daten von Vorteil, wie sie in den Bereichen Medizin, Sozialwissenschaften und Ökonomie häufig auftreten. Die Grundlagen gemischter Modelle, einschließlich der Annahmen und mathematischen Formulierungen, werden ausführlich erörtert.

Ein zentrales Element der theoretischen Grundlagen ist die Likelihood-Inferenz. Im Folgenden wird die Schätzung der Parameter von gemischten Modellen unter Verwendung der Maximum-Likelihood-Methode erörtert. Es wird insbesondere dargelegt, wie die Likelihood-Funktion zur Schätzung der festen und zufälligen Effekte maximiert wird und wie der Likelihood-Ratio-Test (LRT) zur Evaluierung der Modelle zum Einsatz kommt. Der Likelihood-Ratio-Test (LRT) erlaubt die Bestimmung der Signifikanz zusätzlicher Parameter sowie die Identifikation potenzieller Verzerrungen in der Inferenz.

In diesem Kapitel wird die theoretische Basis für die nachfolgenden empirischen Untersuchungen dargelegt. Der vorliegende Abschnitt bietet eine umfassende Darstellung der relevanten Methoden und Konzepte, welche erforderlich sind, um die Verzerrung der Inferenz bei der Verwendung gemischter Modelle in latenten Repräsentationen zu verstehen und zu analysieren.

1 Einführung in Variational Autoencoder (VAE)

Die Anwendung der gemischten Modelle auf einer latenten Repräsentation erfolgt,wie bereits erwähnt, mittels Variational Autoencoder (VAE). Variational Autoencoder sind für die Modellierung latenter Repräsentationen von großem Interesse, da sie hochdimensionale Datensätze mit Hilfe ihres Encoders im latenten Raum niedrigdimensional darstellen können. Dies reduziert die Komplexität der Modellierung und ermöglicht es, gemischte Modelle effizienter und genauer zu betreiben. Sie sind generative Modelle, welche versuchen die zugrunde liegende Struktur der Inputdaten x im latenten Raum zu modellieren. Im Gegensatz zu herkömmlichen Autoencodern ist der VAE in der Lage, nicht nur den Eingabedatensatz zu rekonstruieren, sondern auch neue Inhalte, die ähnliche Merkmale wie die Trainingsdaten aufweisen, zu generieren. Dies wird durch die verbesserte Repräsentation ermöglicht (vgl. [**bigdata-insider-vae**]). Insbesondere wird der latente Raum nicht wie bei normalen Autoencoder durch feste Punkte modellieret, wie

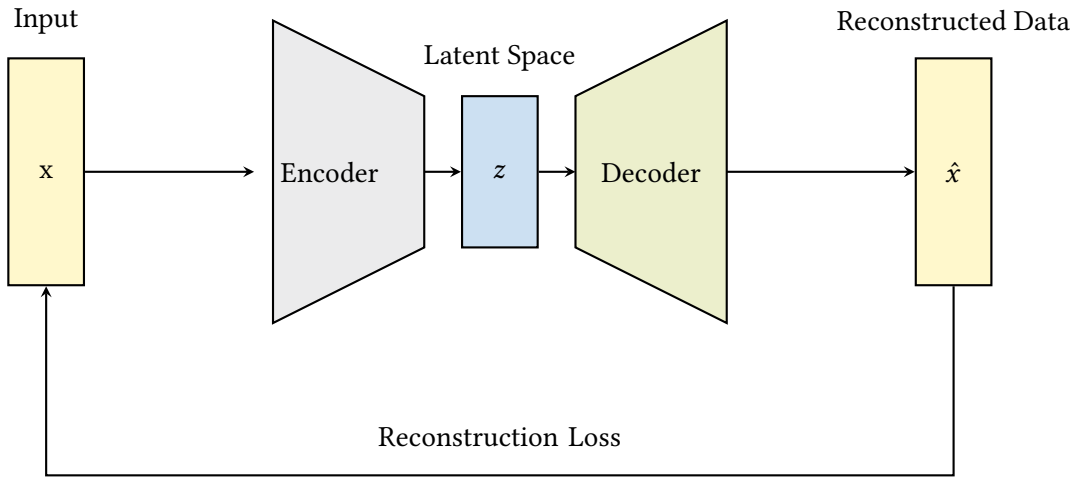

Abbildung 1 Aufbau eines herkömmlichen Autoencoders

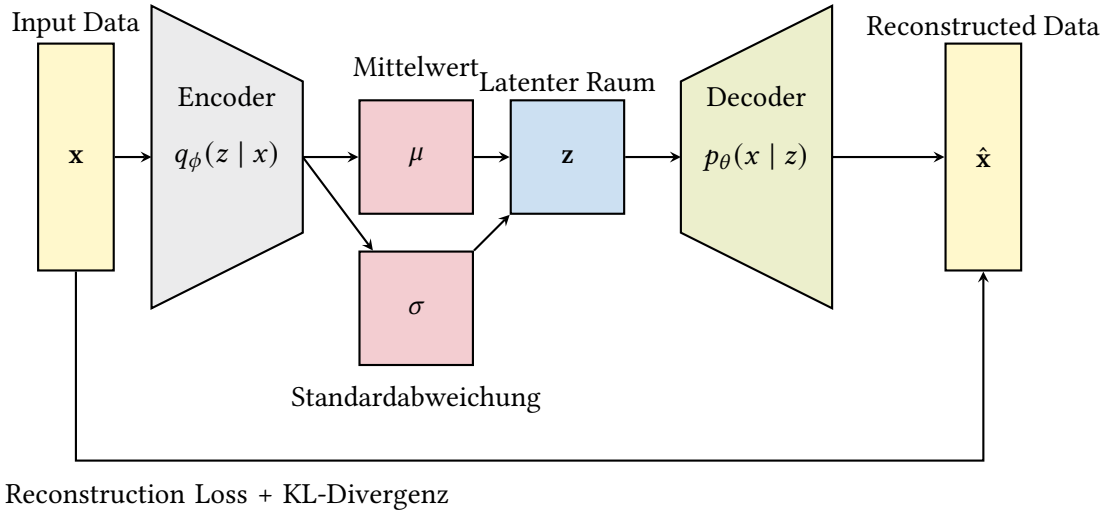

Abbildung 2 Darstellung der Architektur eines Variational Autoencoders (VAE)

es in Abbildung 2 dargestellt ist, sondern in der Erweiterung VAE durch eine Wahrscheinlichkeitsverteilung (Normalverteilung).

1.1 Strucktur des VAEs

Die Architektur eines VAE basiert auf zwei neuronalen Netzwerken (einem Encoder Modell und einem Decoder Modell). Der Encoder transformiert die Inputdaten x in eine niedrigdimensionale latente Repräsentation z. Die latenten Variablen werden als Verteilung in Form eines Mittelwerts μ und einer Standardabweichung σ kodiert. Der Decoder versucht aus den latenten Daten die Originaldaten so genau wie möglich zu rekonstruieren. Beide Modelle bestehen jeweils aus mehreren neuronalen Schichten, die jeweils die Transformation durchführen und lernen die wesentlichen Merkmale der Eingabedaten zu extrahieren und eine komprimierte Version dieser Daten zu erzeugen.

Latenter Raum
Variablen, die man nicht direkt messen kann, demnach nicht Teil des erhaltenen Datensatzes sind, bezeichnet man als latente Variablen. Sie werden erst mithilfe der gegebenen Daten erschlossen und ergeben im Verbund den latenten Raum.
Im VAE werden die latenten Variablen z aus der priori-Verteilung $p_\theta(x)$ gezogen, welche eine multiva-

riate Normalverteilung $p_\theta(x) = \mathcal{N}(z; 0, I)$ ist. Die latenten Daten werden aus den Inputdaten durch den Encoder gezogen, welcher die posteriori Verteilung durch eine variable Verteilung $q_\phi(z, x)$ approximiert. Der Encoder erlernt somit zwei Vektoren, nämlich den Mittelwert μ und die Standardabweichung σ^2 der Normalverteilung $q_\phi(z, x) = \mathcal{N}(z; \mu(x), \sigma^2(x))$.
Der Decoder versucht aus den latenten Variablen die Inputdaten x durch die likelihood-Verteilung $p_\theta(x|z)$ zu rekonstruieren. Die Wahrscheinlichkeit, dass die beobachteten Daten aus den latenten Repräsentationen generiert wurden, wird durch dieses Modell modelliert. Auch hier wird typischerweise eine Normalverteilung angenommen, sofern die Daten reellwertig sind. Im Falle binärer Daten wird die Verteilung als Bernoulli-Verteilung modelliert.
Für weiterführende Details wird auf die Publikation [**Auto-EncodingVariationalBayes**] verwiesen.

1.2 Training VAE

Das Training eines Variational Autoencoder basiert auf den Prinzipien der Variationsinferenz, einer Methode zur Approximation komplexer posterior Verteilungen. Die Berechnung der posterior Verteilung ist besonders bei komplexen Modellen mit Schwierigkeiten verbunden. Infolgedessen wird bei der Variationsinferenz eine einfachere Verteilung verwendet, um die wahre posteriore Verteilung zu approximieren, wodurch Berechnungen effizienter und skalierbarer werden. Im Kontext eines VAE wird die wahre posterior Verteilung $p(z|x)$ der latenten Variablen in Abhängigkeit der Input-Daten, durch eine einfachere Verteilung $q_\phi(x)$ approximiert. Das Ziel ist es die Parameter dieser Verteilung so zu optimieren, dass $q_\phi(x)$ so nah wie möglich an $p(z|x)$ liegt. Diese Annäherung wird durch die Maximierung des Evidence Lower Bound (ELBO) erreicht, der eine untere Schranke der Datenloglikelihood darstellt.
Zur Effizienten Berechnung wird der Reparametrisierungs Trick verwendet, welcher die Anwendung von Gradientenverfahren zur Optimierung der Modellparameter ermöglicht.

In diesem Abschnitt werden die wesentlichen mathematischen Aspekte hinter dem Training des VAE erläutert und beschrieben.

Das Training eines Variational Autoencoders (VAE) umfasst mehrere Schritte, deren Ziel es ist, die Parameter des Modells so anzupassen, dass der Evidence Lower Bound (ELBO) maximiert wird. Der Prozess lässt sich in drei Hauptkomponenten unterteilen: die Definition des ELBO, die Anwendung des Reparameterization Trick und die Optimierung des Modells mittels stochastischer Gradientenverfahren.

Der ELBO lässt sich aus dem Rekonstruktionsverlust (Reconstruction Loss) und der Kullback-Leibler-Divergenz zusammensetzten. Die Kullback-Leibler-Divergenz quantifiziert die Differenz zwischen der approximierten posterior Verteilung und der prior Verteilung und ist folgendermaßen definiert:

Definition 1.1 (Kullback-Leibler-Divergenz (KL-Divergenz))**.**
Sei $q_\phi(z|x)$ die approximierte Posteriori-Gauß-Verteilung und $p(z)$ die Priori-Gauß-Verteilung. Dann ist die KL-Divergenz definiert als

$$D_{KL}(q_\phi(z|x)||p(z)) = \int q_\phi(z|x) \log\left(\frac{q_\phi(z|x)}{p(z)}\right) dz$$

Der Rekonstruktionsverlust misst, wie gut der Decoder die Input Daten rekonstruiert hat. Er wird als log-Likelihood der rekonstruierten Daten angegeben. Aus der sich auch der ELBO herleiten lässt(vgl.

[**IntroductiontoVAEs**]).

$$\log p_\theta(x) = \log p_\theta(x) * \overbrace{\int q_\phi(z|x)dz}^{=1} \tag{1}$$
$$= \int \log p_\theta(x) q_\phi(z|x)dz \tag{2}$$
$$= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x)] \tag{3}$$
$$= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{p_\theta(z|x)}\right]\right] \tag{4}$$
$$= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)q_\phi(z|x)}{q_\phi(z|x)p_\theta(z|x)}\right]\right] \tag{5}$$
$$= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]\right]}_{\substack{=\mathcal{L}_{\theta,\phi}(x) \\ \text{(ELBO)}}} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{q_\phi(z|x)}{p_\theta(z|x)}\right]\right]}_{=D_{KL}(q_\phi(z|x)||p_\theta(z|x))} \tag{6}$$

Der zweite Term in Gleichung 6 ist nach Definition die nicht negative Kullback-Leibler-Divergenz (KL-Divergenz) zwischen $q_\phi(z|x)$ und $p_\theta(z|x)$) und der erste Term in Gleichung 6 stellt den Evidence Lower Bound (ELBO) dar.
Dieser wird wir folgt definiert:

Definition 1.2 (Evidence Lower Bound (ELBO) für VAEs)**.**
Sei $q_\phi(z|x)$ das Encoder Modell und $p_\theta(x,z)$ das Decoder Modell. Der ELBO ist definiert durch

$$\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x,z) - \log q_\phi(z|x)]$$

Es ist leicht zu sehen, dass durch Umstellen der Gleichung 6 sich die ELBO aus dem Rekonstruktionsverlust und der KL-Divergenz ergibt (Vgl. Gleichung 7). Insbesondere ist zu erkennen, dass der ELBO, dadurch dass die KL-Divergenz nicht negativ ist, eine untere Schranke für die log-Likelihood der Daten darstellt.

$$\mathcal{L}_{\theta,\phi}(x) = \log p_\theta(x) - D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \tag{7}$$
$$\leq \log p_\theta(x) \tag{8}$$

Alternativ kann dies mit der Jensenschen Ungleichung (vgl. [**JensenscheUngleichung**]) hergeleitet werden:

$$\log p_\theta(x) = \log \int p_\theta(x,z)dz \tag{9}$$
$$= \log \int p_\theta(x,z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz \tag{10}$$
$$= \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \tag{11}$$
$$\overset{Jensen}{\geq} \mathbb{E}_{q_\phi(z|x)}\left[\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathcal{L}_{\theta,\phi}(x) \tag{12}$$

Es ist sofort ersichtlich, dass die log-likelihood $p_\theta(x)$ durch Maximieren der ELBO bzgl. θ und ϕ selbst maximiert wird und somit die Qualität unseres generatives Modells verbessert wird. Gleichzeitig minimiert sich dadurch die KL-Differenz zwischen der approximativen Verteilung $q_\phi(z|x)$ und den wahren posterior

Verteilung $p_\theta(z|x)$. Durch die Maximierung der ELBO wird also die Approximation $q_\phi(z|x)$ an die posterior Verteilung optimiert.

Die Maximierung der ELBO kann durch stochastische Gradientenverfahren wie Stochastic Gradient Descent (SGD) oder andere fortschrittliche Verfahren erfolgen. Die Berechnung der Gradienten des ELBOs bzgl. θ stellt keine Probleme dar. Mit dem für solche Methoden üblichen Monte-Carlo Schätzer (Gleichung 16)lassen sich die Gradienten bzgl. θ einfach berechnen, wie man in den folgenden Gleichungen sehen kann.

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{\phi,\theta}(x) &= \nabla_\theta \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x,z) - \log q_\phi(z|x)] && (13)\\
&= \mathbb{E}_{q_\phi(z|x)}[\nabla_\theta(\log p_\theta(x,z) - \log q_\phi(z|x))] && (14)\\
&\approx \nabla_\theta(\log p_\theta(x,z) - \log q_\phi(z|x)) && (15)\\
&=\approx \nabla_\theta \log p_\theta(x,z) && (16)
\end{aligned}$$

Der Monte-Carlo-Schätzer für Gradienten ist eine gängige Methode und kann wie folgt definiert werden (vgl. [**MonteCarloEstimation**]):

$$\nabla_\phi \mathbb{E}_{q_\phi(z)}[f(z)] = \mathbb{E}_{q_\phi(z)}[f(z)\nabla_{q_\phi(z)} \log q_\phi(z)] \approx \frac{1}{L}\sum_{l=1}^{L} f(z)\nabla_{q_\phi(z^{(l)})} \log q_\phi(z^{(l)}) \tag{17}$$

wobei $z^{(l)} \sim q_\phi(z|x)$ ist.
Allerdings ist die Berechnung der Gradienten von $\mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z) - \log q_\phi(z|x)\right]$ bezüglich des Variationsparameters ϕ problematisch, da der Erwartungswert des ELBO bzgl. $q_\phi(z|x)$ genommen wird und $q_\phi(z|x)$ eine Funktion vonϕ ist.

$$\begin{aligned}
\nabla_\phi \mathcal{L}_{\theta,\phi}(x) &= \nabla_\phi \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x,z) - \log q_\phi(z|x)\right] && (18)\\
&\neq \mathbb{E}_{q_\phi(z|x)}\left[\nabla_\phi(\log p_\theta(x,z) - \log q_\phi(z|x))\right] && (19)
\end{aligned}$$

Zur Lösung dieses Problems wird der sogenannte Reparameterization-Trick eingesetzt, welcher die Zufallsvariable transformiert um die Gradienten-Berechnung zu vereinfachen.

Reparametrisierungs Trick

Der Reparameterisierungs-Trick ist eine Methode zur Vereinfachung der Gradientenberechnung in Variational-Autoencodern. Er ermöglicht eine effizientere Berechnung der Gradienten der Evidence Lower Bound und somit eine effizientere Optimierung dessen. Der Reparameterization Trick transformiert die Zufallsvariable z in eine andere von z unabhängige deterministische Funktion von einer in eine von z unabhängigen Hilfsvariablen ϵ. Sei also die latente Variable z, die aus $q_\phi(z|x)$ gezogen wurde, gegeben. Sie wird nun als deterministische Funktion einer Hilfsvariablen ϵ unabhängig von ϕ ausgedrückt. Die Transformation sieht dann wie folgt aus:

$$z = g(\epsilon, x, \phi)$$

Dabei ist $g(\epsilon, x, \phi)$ eine differenzierbare Funktion und ϵ eine Zufallsvariable mit einer bekannten Verteilung (z.B. $\epsilon \sim \mathcal{N}(0, I)$).
Im Falle einer Gaußverteilung $z \sim \mathcal{N}(\mu, \sigma^2)$ könnte die Umparametrisierung wie folgt aussehen

$$z = \mu + \sigma \odot \epsilon \quad \text{mit}\, \epsilon \sim N(0, I).$$

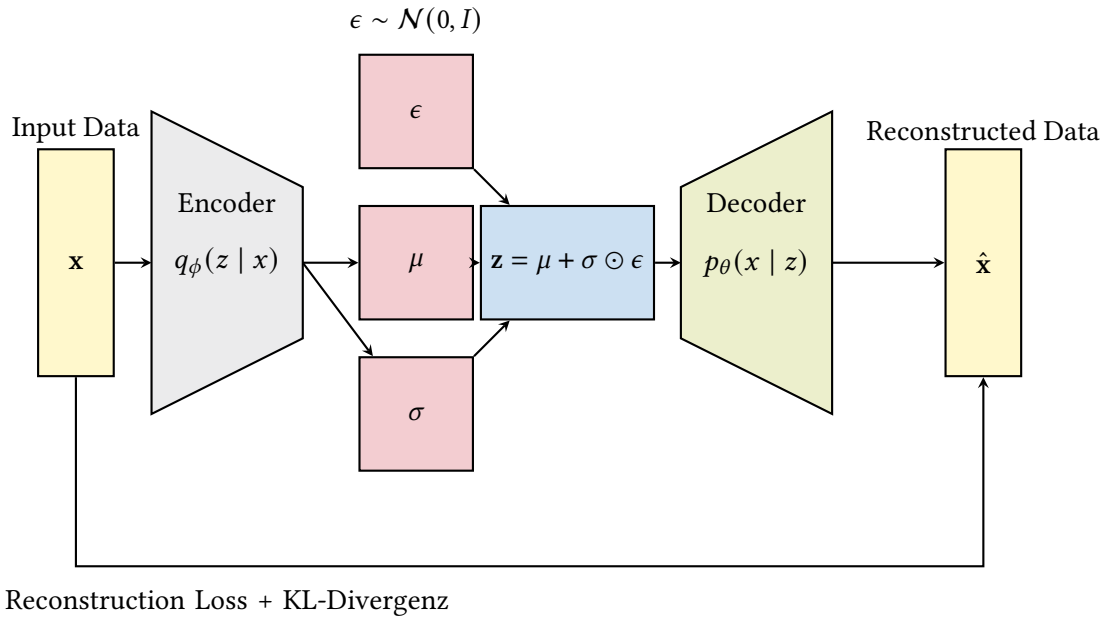

Abbildung 3 Architektur eines VAEs mit Reparameterization Trick

Dabei sind μ und σ die Inferenzparameter ϕ. Durch die Umparametrisierung können die Gradienten bezüglich ϕ effizient berechnet werden, da der Erwartungswert über $q_\phi(z|x)$ sich nun als Erwartungswert über $p(\epsilon)$ ersetzen lässt (vgl. [**MonteCarloEstimation**]).

$$\nabla_\phi \mathbb{E}_{q_\phi(z)}[f(z)] = \nabla_\phi \int q_\phi(z) f(z) dz \tag{20}$$

$$= \nabla_\phi \int p(\epsilon) f(g_(\epsilon, x, \phi)) d\epsilon \tag{21}$$

$$= \nabla_\phi \mathbb{E}_{p(\epsilon)}[f(g(\epsilon, x, \phi))] \tag{22}$$

$$= \mathbb{E}_{p(\epsilon)}[\nabla_\phi f(g(\epsilon, x, \phi))] \tag{23}$$

Der Reparameterisierungstrick bietet somit eine effiziente und flexible Methode zur Berechnung von Gradienten in Modellen mit latenten Variablen und ermöglicht die Anwendung leistungsstarker Optimierungsmethoden wie SGD auf komplexe probabilistische Modelle. Wie der Reparametrisierungstrick in einem VAE aussieht ist in Abbildung 3 veranschaulicht.
Der Gradient der ELBO kann nun geschrieben werden als

$$\nabla_\phi \mathcal{L}_{\theta,\phi}(x) = \mathbb{E}_{p(\epsilon)}\left[\nabla_\phi (\log p_\theta(x, z) - \log q_\phi(z|x))\right]$$

mit $z = g(\epsilon, x, \phi)$. Für die Optimierung des ELBO werden stochastische Gradientenverfahren verwendet. Die Wahl fällt im Rahmen dieser Arbeit auf das Verfahren Stochastic Gradient Descent (SGD).

2 Grundlagen Gemischte Modelle

Ein gemischtes Modell stellt ein statistisches Verfahren zur Datenanalyse dar, welches sowohl feste als auch zufällige Effekte (fixed and random effects) modelliert. Gemischte Modelle finden insbesonder bei der Analyse von longitudinalen und cluster-spezifischen Daten Anwendung, welche aus zeitlich wiederholten Beobachtungen (y_{it}, x_{it}), $t = 1, ..., T_i$ für jedes Individuum $i = 1, ..., n$ bestehen. Die Variable y kennzeichnet dabei eine Antwortvariable, während x ein Vektor von Kovariablen darstellt. Ein Beispiel für einen solchen Datensatz ist ein medizinischer Datensatz,

$$(y_i, x_i) = (y_{i1}, ..., y_{iT_i}, x_{i1}, ..., x_{iT_i})$$

bei dem y_{ij} eine Beobachtung an Individuum i zum Zeitpunkt t_{ij} bezeichnet und T_i ist die Anzahl an Beobachtungen.

Zur Einführung der gemischten Modelle folgen wir den Notationen in [**fahrmeir-2001-multivariate**] und [**fahrmeir-2011-regression**]. Longitudinal und cluster-spezifische Daten weisen zwei Ebenen auf. Im Folgenden betrachten wir das Beispiel des medizinischen Datensatzes. Die erste Ebene bezieht sich dabei auf die Daten innerhalb einer Gruppe oder eines Individuums. In diesem Fall umfasst die erste Ebene den Patienten als Individuum mit seinen unterschiedlichen Werten für die Tests entlang der Zeitreihe T_i . Auf der allgemeineren zweiten Ebene erfolgt eine Betrachtung aller Patienten.

Im Rahmen eines gemischten Modells wird auf der ersten Ebene angenommen, dass die Antwortvariablen linear von den unbekannten bevölkerungsspezifischen festen Effekten β und den unbekannten cluster-spezifischen zufälligen Effekten b_i abhängen.
Die folgende Gleichung beschreibt das Modell:

$$y_{it} = x_{it}^t \beta + z_{it}^t b_i + \epsilon_{it} \tag{24}$$

Innerhalb des Modells werden die Designvektoren z_{it} und w_{it} als unabhängige Variablen definiert, wobei z_{it} beispielsweise die Testwerte in einem medizinischen Datensatz repräsentiert.Die Zufallsvariable ϵ_{it} hingegen ist unkorreliert und folgt einer normalverteilten Wahrscheinlichkeitsdichte mit Erwartungswert $\mathbb{E}(\epsilon_{it}) = 0$ und Varianz $Var(\epsilon_{it}) = \sigma^2$. Der Ausdruck a^t bezeichnet den transponierten Vektor, bzw. die transponierte Matrix a.

Betrachtet man nun die zweite Ebene, so werden die zufälligen Effekte b_i zwischen den verschiedenen Individuen gemäß einer Mischverteilung mit $\mathbb{E}(b_i) = 0$ unabhängig variieren. Es wird angenommen, dass die zufälligen Effekte b_i unabhängig und identisch normalverteilt sind,

$$b_i \sim \mathcal{N}(0, Q) \tag{25}$$

mit der (q x q) Kovarianzmatrix $Cov(b_i) = Q > 0$, welche symmetrisch und positiv semi-definit ist. Eine ausführliche Beschreibung findet sich in [**pinheiro2000**] (Kapitel 2.2.1).
Aufgrund dieser Überlegungen lässt sich nun das Model 24 in einer allgemeineren Form beschreiben:

Definition 2.1 (Lineares gemischtes Modell für Longitudinal- oder Clusterdaten)**.**
Seien $X_i = (x_{i1}, ..., z_{iT_i})$ und $Z_i = (z_{i1}, ..., z_{iT_i})$ bekannte Designmatrizen für die festen und zufälligen Effekte. Seien β ein p-dimensionaler Vektor von festen Effekten und b_i ein q-dimensionaler Vektor von zufälligen Effekten und sei $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{iT_i})$ der normalverteilte Fehlervektor.

Ein lineares gemischtes Modell für den T_i-dimensionalen Antwortvektor der i-ten Gruppe wird durch

$$y_i = X_i * \beta + Z_i * b_i + \epsilon_i$$

$$b_i \sim \mathcal{N}(0, Q), \epsilon_i \sim \mathcal{N}(0, R = \sigma_\epsilon^2 I)$$

definiert.

Die Daten der zufälligen und festen Effekte werden in einer Designmatrix (Datenmatrix) gespeichert.Die Parametervektoren β (für die festen Effekte) und b_i (für die zufälligen Effekte) initialisieren den Einfluss der Daten auf den Antwortvektor. Um auch für immer auftretende Messfehler oder unerwartete Einflüsse gewappnet zu sein, wird ein zufälliges Rauschen ϵ hinzugefügt.

Aufgrund des normalverteilten Fehlervektors können nun auch ein marginales Modell als multivariates heteroskedastisches lineares Regressionsmodell definiert werden. Dieses Modell ist für die Berechnung der Likelihood-Inferenz von entscheidender Bedeutung.

Definition 2.2 (Marginales gemischtes Modell)**.**
Seien die Annahmen von 2.1 gegeben. Das marginale gemischte Modell ist definiert als

$$y_i = X_i\beta + \epsilon_i^*,$$

mit dem multivariaten Fehlervektor $\epsilon_i^* = (\epsilon_{i1}^*, ..., \epsilon_{iT_i}^*)$ *mit* $\epsilon_{it}^* = z_{it}^T b_i + \epsilon_i$*. Die* ϵ_{it}^* *sind dabei unabhängig und identisch verteilt (i.i.d.),*

$$\epsilon_i^* \sim \mathcal{N}(0, V_i), \quad mit \quad V_i = \sigma_\epsilon^2 I + Z_i Q Z_i^t \tag{26}$$

Die einzelnen Cluster/Gruppen können zu einem einzigen allgemeinen linearen gemischten Modell zusammengefasst werden.

Definition 2.3 (Allgemeines lineares gemischtes Modell)**.**
Ein lineares gemischtes Modell ist definiert durch

$$y = X\beta + Zb + \epsilon$$

mit

$$\begin{pmatrix} b \\ \epsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q & 0 \\ 0 & R = \sigma_\epsilon^2 I \end{pmatrix}\right)$$

gegeben. Dabei sind X*, bzw* Z *die Designmatrizen der festen, bzw zufälligen Effekte,* β *und* b *die Parametervektoren der festen und der zufälligen Effekten und* ϵ *der Fehlervektor.*

In Konsequenz dessen lässt sich das marginale Modell verallgemeinern zu:

$$y = X\beta + \epsilon^* \tag{27}$$

wobei $\epsilon^* = Zb + \epsilon$ ist mit $\epsilon^* \sim \mathcal{N}(0, V)$ und $V = R + ZQZ^t$.

3 Likelihood Inferenz und Verzerrung

Um die Verzerrung der Inferenz messen zu können, ist es zunächst erforderlich, die Theorie zur Likelihood-Inferenz von gemischten Modellen einzuführen. Dies umfasst sowohl die Schätzung der Parameter der zufälligen Effekte b_i als auch die Schätzung der Parameter β, σ_ϵ und Q. Um die Verzerrung zu quantifizieren, werden wir ein vollständiges gemischtes Modell mit einem reduzierten Modell ohne einen festen Effekt vergleichen. Dazu wird üblicherweise der sogenannte Likelihood-Ratio-Test (LRT) verwendet. Wie dieser Test genau funktioniert und wie der LRT durchgeführt wird, werden wir später erläutern. Zuvor benötigen wir noch etwas Theorie zur Likelihood-Berechnung.

3.1 Likelihood Berechnung gemischter Modelle

Im Folgenden wird die Schätzung der unbekannten Parameter erörtert. Der Vorliegende Ansatz basiert auf den Ausführungen von [**fahrmeir-2011-regression**].

Die Berechnung der Schätzer erfolgt mittels Maximum-Likelihood-Methode. Als Alternative kann die restringierte ML-Methode heran gezogen werden, die jedoch nicht für den Likelihood-Ratio-Test geeignet ist. Daher erfolgt die Berechnung der Parameter mittels der ML-Methode.
Die Schätzung der Parameter in einem gemischten Modell ist jedoch mit gewissen Schwierigkeiten verbunden. Neben dem β sind auch b_i, Q und σ_ϵ unbekannt. Daher ist es erforderlich, sowohl die festen und zufälligen Effekte als auch die unbekannten Parameter in Q und σ_ϵ, die wir als θ bezeichnen, zu schätzen. Dies bedingt eine geschachtelte Schätzung.

Im Folgenden wird zunächst angenommen, dass die Kovarianzen R, bzw. σ_ϵ, und Q bekannt sind. In diesem Zusammenhang ist auch V gemäß 27 bekannt. Für die Schätzung von β, ausgehend vom marginalen Modell, bietet sich

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y \tag{28}$$

an. Dieser Kleinste-Quadrate-Schätzer für β ergibt sich aus dem verallgemeinertem Kleinste-Quadrate-Kriterium (vgl. [**KQ-SchÃďtzer**]), welches folgenden Term

$$(y - X\beta)^t V^{-1} (y - X\beta)$$

bezüglich β minimiert. Siehe hierzu auch [**fahrmeir-2011-regression**] (Kap. 3).
Der KQ-Schätzer ist gleichzeitig der log-Likelihood Schätzer unter der Normalverteilungsannahme.
Die log-Likelihood für β aus dem marginalen Modell sieht folgendermaßen aus:

$$l(\beta) = -0.5 * (log(|V|) + (y - X\beta)^t V^{-1} (y - X\beta) + N * log(2\pi)).$$

Ableiten nach β ergibt den KQ-Schätzer aus 28.

$$\frac{d}{d\beta} l(\beta) = X^t V^{-1} (y - X\beta) \stackrel{!}{=} 0 \Rightarrow \hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y$$

Siehe hierzu auch [**fahrmeir-2011-regression**] (Kap. 3).

Gemäß dem Gauß-Markov-Theorem stellt $\hat{\beta}$ den besten linearen erwartungstreuen Schätzer (BLUE, best linear unbiased estimator) für die fixen Effekte dar. Zur Ermittlung des Schätzers ist lediglich eine Schätzung der Parameter in V sowie der Einsatz des Schätzers $\hat{V}$ von V in $\hat{\beta}$ erforderlich.
Für den Schätzer von b verwenden wir den bedingten Erwartungswert $E(b|y)$ von b, gegeben die Daten y, welcher unter der Normalverteilungsannahme der beste Schätzer ist (vgl. [**fahrmeir-2011-regression**] Kap. 6.3.1).
Betrachtet man nun die gemeinsame Verteilung von b und y, welche folgendermaßen dargestellt wird:

$$\begin{pmatrix} y \\ b \end{pmatrix} \sim \mathbf{N}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} V & ZQ \\ QZ^t & Q \end{pmatrix} \right)$$

In Anbetracht dessen erhalten wir $E(b|y) = QZ^t V^{-1} (y - X\beta)$.
Ersetzt man nun β durch den Schätzer $\hat{\beta}$ erhält man den Schätzer für die zufälligen Effekte

$$\hat{b} = \hat{Q} Z^t \hat{V}^{-1} (y - X\hat{\beta}).$$

Der Schätzer $\hat{b}$ ist der beste lineare unverzerrte Schätzer (BLUP, best linear unbiased prediction)

Definition 3.1 (Schätzer für feste und zufällige Effekte)**.**
Sei $y = X\beta + Zb + \epsilon$ ein lineares gemischtes Modell und $y = X\beta + \epsilon^*$ das zugehörige Marginale nach 27. Dann ist

$$\hat{\beta} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} y$$

ein Schätzer für die festen Effekte und

$$\hat{b} = \hat{Q} Z^t \hat{V}^{-1} (y - X\hat{\beta})$$

ein Schätzer für die zufälligen Effekte.

Wie bereits erwähnt, soll der Parametervektor θ alle unbekannten Parameter in $V = V(\theta), Q = Q(\theta)$ und $\sigma_\epsilon = \sigma_\epsilon(\theta)$ enthalten. Anhand des Schätzers $\hat{\theta}$ lassen sich der Kovarianzschätzer sowie die Schätzer der festen und zufälligen Effekte berechnen. Die ML-Methode für θ basiert auf dem marginalen Modell

$$y \sim \mathcal{N}(X\beta, V(\theta)).$$

Die Log-Likelihood von β und θ ist gegeben durch

$$l(\beta, \theta) = -\frac{1}{2}(log(|V|) + (y - X\beta)^t V^{-1}(y - X\beta)).$$

Maximieren von $l(\beta, \theta)$ bezüglich β für festes θ ergibt

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y.$$

Setzt man nun $\hat{\theta}$ in $l(\beta, \theta)$ ein, so erhält man die Profil-Log-Wahrscheinlichkeit

$$l(\theta)_p = -\frac{1}{2}(log(|V|) + (y - X\hat{\beta})^t V^{-1}(y - X\hat{\beta})).$$

Folglich erhält man den ML-Schätzer $\hat{\theta}_{ML}$ durch Maximierung von $l(\theta)_p$.

Definition 3.2 (Kovarianz-Schätzer)**.**
Sei $y = X\beta + Zb + \epsilon$ ein lineares gemischtes Modell mit

$$\begin{pmatrix} b \\ \epsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Q & 0 \\ 0 & R \end{pmatrix}\right)$$

und sei θ der unbekannte Parametervektor von Q,R und $V = Var(y)$.
Dann ist $\hat{\theta}_{ML}$ der ML-Schätzer für θ, den man durch maximieren von

$$l(\theta)_p = -\frac{1}{2}(log(|V|) + (y - X\hat{\beta})^t V^{-1}(y - X\hat{\beta}))$$

erhält. Dabei ist

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y$$

Mit dem Schätzer $\hat{V}$ lassen sich die Schätzer der festen und zufälligen Effekte nun berechnen.
Um die Verzerrung der Inferenz messen zu können, müssen wir die log-Likelihood Werte berechnen können, um diese in den Likelihood-Ratio-Test einzusetzen. Der log-Likelihood Wert eines gemischten Modell ergibt sich aus der Maximum-Likelihood (ML)-Methode und ist folgendermaßen definiert:

Definition 3.3 (log-Likelihood Wert für ein gemischtes Modell)**.**
Sei $r = y - X(X^t V^{-1} X)^{-1} X^t V^{-1} y$ und p der Rang von X

$$l_{ML}(Q, R) = -0.5 * (log(|V|) + r' V^{-1} r + N * log(2\pi))$$

$$l_{REML}(Q, R) = -0.5 * (log(|V|) + X' V^{-1} X + r' V^{-1} r + (N - p) * log(2\pi))$$

$l_{REML}(Q, R)$ ist die eingeschränkte log-Likelihood, der sich aus der Methode "Restricted Maximim Likelihoodërgibt und entspricht im Wesentlichen der normalen log-Likelihood mit Ausnahme einer Differenz. Bei der "Restricted Maximim Likelihood"werden im Gegensatz zu der Methode "Maximum Likelihood"die Freiheitsgrade, die für die Schätzung fester Effekte bei der Schätzung von Varianzkomponenten verwendet werden, berücksichtigt. Im Gegensatz zum ursprünglichen Datenvektor basiert die eingeschränkte Maximum-Likelihood-Methode auf linearen Kombinationen der Beobachtungen, die so gewählt sind, dass diese Kombinationen invariant zu den Werten der festen Effektparametern sind.
Diese linearen Kombinationen sind äquivalent zu den Residuen, die nach der Anpassung durch normale kleinste Quadrate (gewichtet bei Angabe einer Regressionsgewichtung) lediglich den festen Effektanteil des Modells berechnen. Das Verfahren führt somit eine Maximierung in einem eingeschränkten Vektorraum durch.

3.2 Likelihood-Ratio-Test

Die Berechnung der Likelihood-Ratio-Test-Statistik (LRT-Statistik) ist relativ einfach, sofern die Theorie der ML-Methode vergegenwärtigt wird. Zur Erinnerung: Der Vergleich eines reduziertes Modells mit dem vollständigen Modell dient der Evaluierung des Einflusses einer Störgröße. Zur Durchführung dieser Analyse dient der Likelihood-Ratio-Test. Er ermöglicht den Vergleich eines einfacheren Modells (Nullmodell) mit einem komplexeren Modell (alternatives Modell), indem er die Likelihoods, bzw. die log-Likelihoods, der beiden Modelle vergleicht. Dies ist zum Beispiel nützlich um den Einfluss eines zusätzlichen Parameters zu beurteilen.
Der Likelihood-Ratio-Test wird wie folgt definiert:

Definition 3.4 (Likelihood-Ratio-Test (LRT))**.**
Sei $\mathbf{L}_{full}$ *der Likelihood-Wert des vollständigen Modells sowie* $\mathbf{L}_{red}$ *der Likelihood-Wert des reduzierten Modells. Es sei i die Anzahl der Freiheitsgrade.*
Dann ist die LRT Statistik gegeben durch

$$LRT = 2(\log \mathbf{L}_{full} - \log \mathbf{L}_{red})$$

Sofern die Größen $\mathbf{L}_{full}$ und $\mathbf{L}_{red}$ gemäß der Definition initialisiert sind, gilt $\mathbf{L}_{full} > \mathbf{L}_{red}$. Insbesondere gilt $\log(\mathbf{L}_{full}) > \log(\mathbf{L}_{red})$. Sofern die Log-Likelihood-Werte der Modelle bereits als $\mathbf{L}_{full}$ und $\mathbf{L}_{red}$ gegeben sind, lässt sich die LRT-Statistik durch $2(\mathbf{L}_{full} - \mathbf{L}_{red})$ berechnen.

Die Teststatistik des Likelihood-Ratio-Tests ergibt sich letztendlich, indem wir die LRT-Werte als Histogramm darstellen, und folgt einer χ^2-Verteilung. Eine Chi-Quadrat-Verteilung mit k Freiheitsgraden ist folgendermaßen definiert:

Definition 3.5 (χ^2-Verteilung)**.**
Sei $X_1, X_2, \ldots, X_k$ *eine Folge von unabhängigen standardnormalverteilten Zufallsvariablen, also* $X_i \sim N(0,1)$ *für* $i = 1, \ldots, k$*. Dann ist die Zufallsvariable*

$$Y = \sum_{i=1}^{k} X_i^2$$

Chi-Quadrat-verteilt mit k Freiheitsgraden. Wir schreiben:

$$Y \sim \chi^2(k)$$

Die Wahrscheinlichkeitsdichtefunktion der χ^2*-Verteilung mit k Freiheitsgraden ist gegeben durch:*

$$f(x;k) = \begin{cases} \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} & x > 0, \\ 0 & x \leq 0, \end{cases}$$

wobei $\Gamma(\cdot)$ *die Gamma-Funktion ist.*

Die χ^2-Verteilung bietet einen Vergleichswert für die Interpretation der Ergebnisse. Somit können wir feststellen, wie signifikant der Einfluss des zusätzlichen Parameters ist und ob die Anwendung gemischter Modelle im latenten Raum die Inferenz verzerrt. Um dies anschaulich darzustellen legt man das Histogramm der Teststatistik unter die χ^2 Verteilung. Dies erleichtert die Analyse, ob die Inferenz verzerrt ist. Die χ^2-Verteilungen sind in Abbildung 4 veranschaulicht.

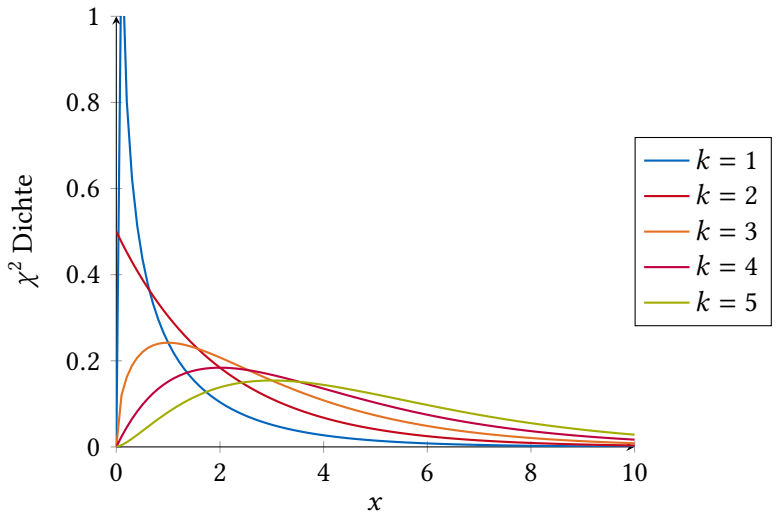

Abbildung 4 Chi-Quadrat-Verteilung für verschiedene Freiheitsgrade k.

Empirische Ergebnisse

Die theoretischen Grundlagen für die Ergebnisse dieser Arbeit sind nun gegeben und es kann mit der Analyse fortgefahren werden. Um am Ende die Verzerrung der Inferenz zu messen, fällt die Wahl im Rahmen dieser Arbeit auf einen Likelihood-Ratio-Test. Um einen Vergleichswert zu haben, schaffen wir zuerst ein Szenario, in dem eine χ^2-Verteilung in der LRT-Statistik erwartbar ist.

1 gemischte Modelle auf simulierten Daten

Da im Rahmen der späteren Analyse ein komplexer longitudinaler medizinischer Datensatz verwendet wird, erfolgt für das einfachere Szenario die Wahl eines Simulationsdesigns für einen einfachen longitudinalen medizinischen Datensatz, welchem wir dann eine Variable hinzufügen, die keinen Einfluss auf die Testergebnisse haben soll. Im Folgenden wird ein Simulationsdesign für eine Studie präsentiert, welche die Herzgesundheit von Patienten über einen Zeitraum von zehn Jahren analysiert. Die Gewichtung der verschiedenen Parameter auf den sogenannten „Health-Score" ist unterschiedlich.

Simulationsdesign
Im Rahmen einer zehnjährigen Studie wurden 500 Patienten im Alter zwischen 30 und 60 Jahren auf verschiedene Parameter untersucht, die einen Einfluss auf die Herzgesundheit haben. Die Simulationen für jeden Parameter basieren auf einer Normalverteilung und umfassen Daten über den Zeitraum von zehn Jahren. Die in Tabelle 1 dargestellten Einflussfaktoren sind als feste Parameter für die Herzgesundheit zu betrachten. In der Berechnung des Health-Scores wird insbesondere berücksichtigt, dass es zu zufälligen Einflussfaktoren kommen kann, die die Herzgesundheit betreffen. Daher wurde in die Berechnung ein zufälliger Interzept und eine zufällige Steigung integriert. Es wird eine Normalverteilung für den random_intercept $\sim \mathcal{N}(0, 2)$ und den random_slope $\sim \mathcal{N}(0, 0.1)$ angenommen. Zu Beginn der Studie wird jedem Patienten zufällig ein Alter zugewiesen, wobei die Parameter gemäß Tabelle 1 berechnet werden. Insbesondere wird zu einem Zeitpunkt, welcher zufällig zwischen drei und zehn Jahren für jeden Patienten festgelegt wird, die Gewichtung der Parameter angepasst. Dies soll einen Behandlungsstart mit Medikamenten simulieren. So werden dann letztendlich mit einer Health-Score Formel

$$y = 150 + gewichte * feste\_Effekte + random\_slope * jahr + random\_intercept + \epsilon$$

die Testergebnisse nach einem gemischten Modell berechnet. $\epsilon \sim \mathcal{N}(0, 0.1)$ ist ein zufälliger Fehlervektor, welcher Messfehler berücksichtigt. Eine Beispielhafte Simulation der Daten ist in 2 für 20 ausgewählte Patienten dargestellt.

feste Effekte	Mittelwert	Standardabweichung	Gewicht	Gewicht nach Behandlungsstart
Systolischer Blutdruck	120	10	-0.1	-2
Diastolischer Blutdruck	80	10	-0.1	-2
Cholesterin	200	30	-0.2	-5
Triglyceride	150	20	-0.2	-3
Kreatinin	1	0.2	-0.1	-0.08
Body-Mass-Index (BMI)	25	4	-0.4	-0.4
Alter			-0.1	

Tabelle 1 Einfluss und Erstellung der Parameter des Health-Scores

LRT-Statistik

Um einen Likelihood-Ratio-Test durchzuführen, der eine Vergleichsstatistik für die spätere Analyse liefert, wird jedem Patienten zufällig ein Geschlecht zugewiesen. Das Geschlecht sollte keinen Einfluss auf die Testergebnisse haben und wird deswegen in der Berechnung des Health-Scores mit Null gewichtet.
Nun wurde ein Szenario entwickelt, in dem ein vollständiges Modell mit einem reduzierten Modell (ohne den Einfluss des Geschlechts) verglichen werden kann. Für die LRT-Statistik erfolgt ein Training beider Modelle jeweils 500 Mal auf einem neu simulierten Datensatz. Im Anschluss erfolgt ein Vergleich der bei jeder Simulation berechneten log-Likelihood-Werte mittels Likelihood-Ratio-Test, wobei das Ergebnis der Vergleichsanalyse zusammengetragen wird. Das Resultat wird in Form eines Histogramms, was zum Abgleich unter eine χ^2-Verteilung gelegt wird, in Abbildung 1 dargestellt. Wie man sieht folgt das grüne Histogramm ohne Verzerrung der roten Kurve, welche die χ^2-Verteilung beschreibt. Das bedeutet, dass bis auf einzelne Ausnahmen, welche durch Instabilitäten der Berechnung immer verursacht werden können, alle Ergebnisse des LRT-Statistik, wie zu erwarten, unter der χ^2-Verteilung bleiben.

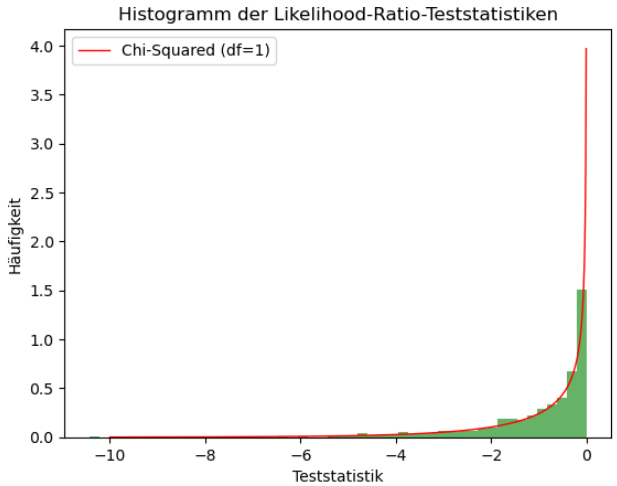

Abbildung 1 Histogramm der LRT-Statistik für vollständiges und reduziertes gemischtes Modell auf simulierten Daten

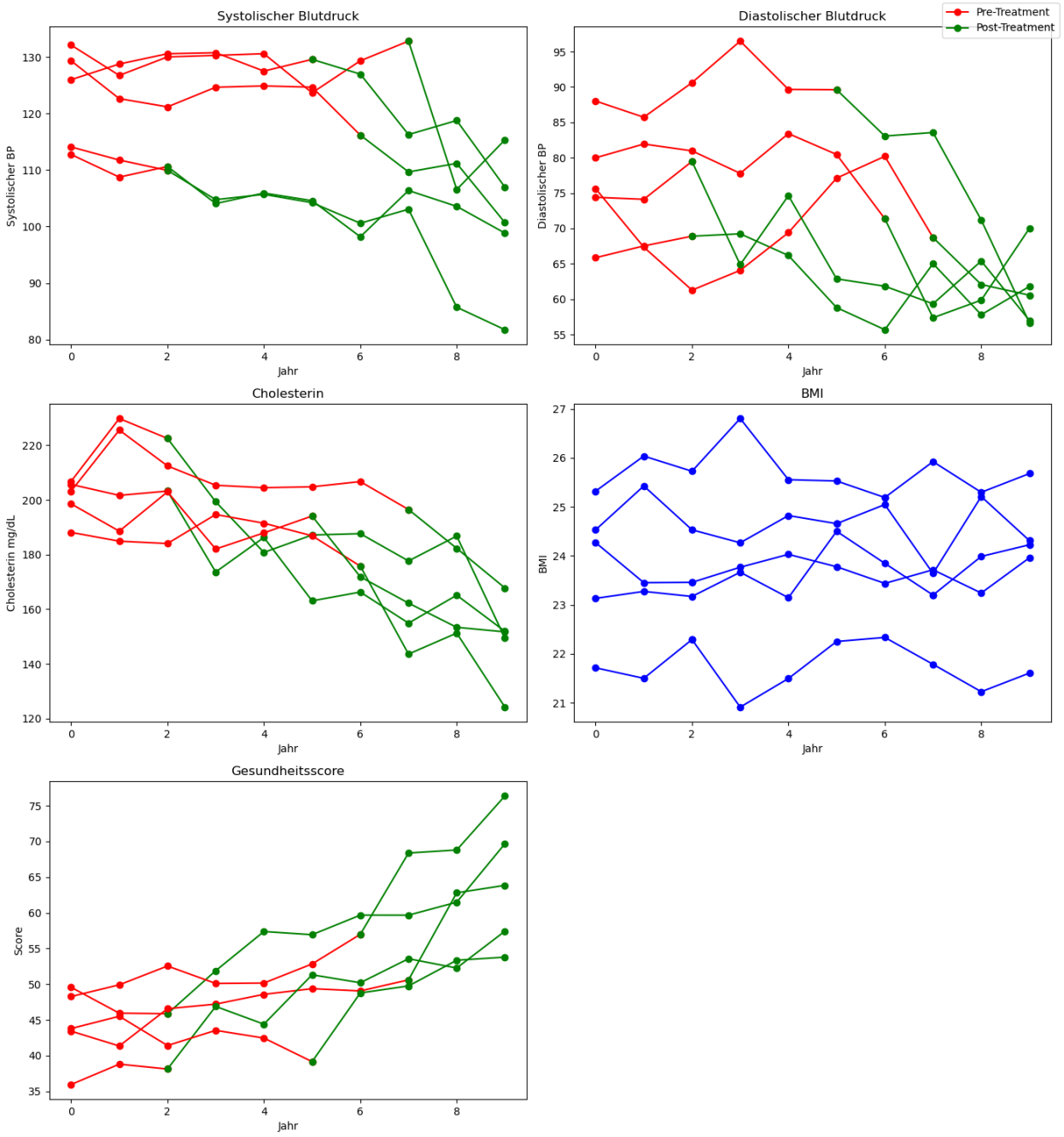

Abbildung 2 Simulierte Datensätze für 20 zufällig ausgewählte Patienten

2 gemischte Modelle in latenten Repräsentationen

In der bisherigen Betrachtung wurden die gemischten Modelle lediglich auf Basis der realen Daten evaluiert. Im Folgenden wird die Betrachtung der gemischten Modelle auf latenten Daten vorgenommen. Zur Analyse der gemischten Modelle auf latenten Daten wird ein Variational-Autoencoder verwendet. Die Grundlage unserer Analyse bildet das folgende Grundszenario: Im Folgenden wird zur Analyse nun einen komplexen longitudinalen medizinischen Datensatz betrachtet, der durch einen Encoder des VAEs im latenten Raum modelliert wird. Das Ziel ist es nun, das gemischtes Modell auf dieser latenten Datenwolke zu trainieren, um herauszufinden, ob es zu einer erwartbaren Verzerrung kommt. Dazu wird im Folgenden zuerst eine detaillierte Betrachtung des vorliegenden Datensatzes vorgenommen.

2.1 Der Datensatz

Im Rahmen dieser Bachelorarbeit basieren die Ergebnisse und Experimente, um die Verzerrung der Inferenz bei der Anwendung gemischter Modelle in latenten Repräsentationen zu untersuchen, auf einem generierten, hoch-dimensionalen, medizinischem Datensatz, welcher sich aus drei zentralen Datensätzen zusammensetzt. Diese Datensätze enthalten Informationen über die Basisdaten der Patienten, die Testergebnisse und zeitbezogene Informationen zu jedem Patienten. Der Datensatz wurde aus datenschutzrechtlichen Gründen einem originellen Datensatz nachgebaut.

Basisdaten
Der 'baseline_df' Datensatz enthält die grundlegenden Informationen der Patienten, welche mit einer eindeutigen Patienten-ID identifiziert werden. Zu jeder Patienten-ID sind folgenden Informationen gegeben:

1. 'family_affected': Gibt an, ob die Familie vorerkrankt ist.
2. 'sco_surg': Chirurgischer Score.
3. '$\leq$ 3': binäres Merkmal.
4. 'onset_age': Alter bei Eintritt der Krankheit.
5. 'presym_diag': Prä-symptomatische Diagnose (1: Ja, 0: Nein).
6. 'presymptomatic': Prä-symptomatischer Zustand (1: Ja, 0: Nein).
7. 'stand_lost': Gibt an, ob Patient Stehfähigkeit verloren hat (1: Ja, 0: Nein).
8. 'stand_gained': Gibt an, ob Patient Stehfähigkeit gewonnen hat (1: Ja, 0: Nein).
9. 'stand_never': Gibt an, ob Patient jemals stehen konnte (1: Ja, 0: Nein).
10. 'patient_id': Eindeutige Patienten-ID.

Eine beispielhafter Eintrag im Datensatz ist in Tabelle 2 wiedergegeben.

patient_id	sco_surg	≤3	onset_age	presym_diag	presymptomatic	stand_lost	stand_gained	stand_never	family_affected
0	0.0	1.0	0.039397	1.0	0.0	0.0	0.0	1.0	1.0
1	0.0	0.0	2.787249	0.0	0.0	0.0	1.0	0.0	-1.0
2	1.0	1.0	1.471984	0.0	0.0	0.0	0.0	1.0	0.0
3	0.0	1.0	1.092828	0.0	0.0	0.0	0.0	1.0	-1.0
4	0.0	0.0	13.150771	0.0	0.0	0.0	1.0	0.0	-1.0

Tabelle 2 Basisdaten der Patienten für Patient 0 bis 4 (baseline_df)

Testergebnisse
Der Datensatz 'test_scores' enthält die Ergebnisse von insgesamt 33 Tests, in denen die Patienten einen Score zwischen 1 und 6 erreichen können. Die Spalte eines Patienten besitzt einen Mobilitäts-Wert und zu jedem Test einen Eintrag (vgl. 3).

patient_id	mobility	test1	test2	test3	test4	test5	test6	test7	test8	...
0	3	2	2	2	2	2	2	2	2	...
0	6	2	2	2	2	2	2	2	2	...
0	6	2	2	2	2	2	2	2	2	...
0	6	2	2	2	2	2	2	2	2	...
0	6	2	2	2	2	2	2	2	2	...

Tabelle 3 Testergebnisse von Patient 0 (test_scores)

Zeitbezogene Daten
Der letzte Datensatz 'time_df' enthält zeitbezogene Informationen, wie das Alter. Des Weiteren gibt er an, seit wann ein Patient behandelt wird ('since_medication') und wieviel Zeit nach dem letzten Medikamentenwechsel vergangen ist ('since_switch'). Symbolisch für den Datensatz werden die Daten für die ersten zwei Patienten in Tabelle 4 erfasst.

patient_id	since_medication	since_switch	age
0	1.467488	0.000000	4.346177
0	1.793292	0.000000	4.671981
0	2.447639	0.000000	5.326328
0	2.773443	0.000000	5.652132
0	3.214237	0.383299	6.092926

Tabelle 4 Zeitbezogene Daten von Patient 0 (time_df)

Diese Datensätze bilden die Grundlage der Analyse dieser Arbeit. Zusammen ergeben sie einen komplexen Datensatz, welcher für 260 Patienten die Ergebnisse von 33 Mobilität-Tests enthält, die mehrmals wiederholt wurden. Somit lässt sich aus diesem Datensatz der Verlauf und die Schwere der Krankheit für jeden Patienten ablesen.

2.2 Gemischtes Modell auf latenter Datenwolke

Das VAE-Modell
Für den Variational Autoencoder benötigt es ein Encoder-Modell und ein Decoder-Modell. Für das Encoder- und Decoder-Modell wird ein vielseitig anpassbares Modell gewählt, welches allerdings recht einfach gehalten wird. Die latente Dimension wird dabei standardmäßig auf zwei gesetzt und der Encoder wird erstmal mit zwei Schichten und einer versteckten Dimension von 150 initialisiert. In der Verlustfunktion des VAEs wird der Reconstruction-Loss und die Kullback-Leibler-Divergenz optimiert.

Für das erste Szenario, was im Rahmen dieser Arbeit untersucht wird, trainieren wir einen VAE und im Anschluss auf den latenten Daten das vollständige und das reduzierte gemischte Modell.

Das gemischte Modell
Für die Berechnung der Schätzer benötigt es natürlich die Designmatrizen für die festen und die zufälligen Effekte.
Die festen Effekte des vorgestellten Datensatzes sind alle in Tabelle 2 aufgezählten Parameter und der künstlich hinzugefügte Parameter 'Geschlecht'. Aus den Werten der festen Effekte setzt sich die Designmatrix der festen Effekte für das vollständige gemischte Modell zusammen. Dementsprechend setzt sich die Designmatrix für die festen Effekte des reduzierten Modells nur aus den Effekten aus Tabelle 2.
Die Designmatrizen für die zufälligen Effekte beider Modelle setzten sich aus den Werten der zufälligen Effekte 'since_medication', 'since_switch' und 'intercept' zusammen.
Der Antwortvektor ist in diesem Szenario nun nicht durch den aus dem gemischten Modell berechneten Antwortvektor gegeben, sondern durch die aus dem Encoder gewonnene latente Datenwolke.

Wenn nun beide gemischte Modelle nach der vorangegangenen Theorie auf der latenten Datenwolke trainiert werden, können die Ergebnisse mit dem Likelihood-Ratio-Test analysiert werden. Für 100 Simulationen dieses Szenarios entsteht somit eine LRT-Statistik, welche wieder in Abbildung 3 unter eine χ^2-Verteilung gelegt wird. Wie man sieht ist schon für 100 Simulationen eine Verzerrung erkennbar. Die Daten erstrecken sich über einen größeren Bereich und liegen für geringere Werte nicht mehr unter der χ^2-Verteilung.

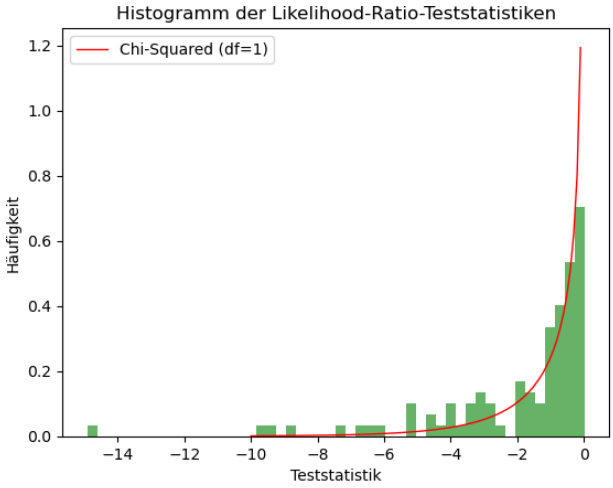

Abbildung 3 Histogramm der LRT-Statistik für vollständiges und reduziertes gemischtes Modell auf latenter Datenwolke

3 Modellierungstechniken

Beschreibung der spezifischen gemischten Modelle und der Techniken zur Gewinnung latenter Repräsentationen.

4 Analysemethoden

4.1 Post-Selection-Inferenz

Post-Selection Inference (PSI) bezieht sich auf statistische Methoden, die die Tatsache berücksichtigen, dass eine Hypothese oder ein Modell aufgrund der Daten ausgewählt wurde. Traditionelle statistische Inferenzmethoden setzen oft voraus, dass das Modell oder die Hypothese a priori festgelegt wurde, was in der Praxis selten der Fall ist. Bei PSI werden die statistischen Eigenschaften angepasst, um die zusätzliche Unsicherheit durch den Selektionsprozess zu reflektieren. Dies ist besonders wichtig in Bereichen wie maschinellem Lernen und Data Mining, wo oft zahlreiche Modelle getestet und das beste ausgewählt wird. Durch PSI können zuverlässigeren Konfidenzintervalle und p-Werte berechnet werden, die die Selektion berücksichtigen, was zu robusteren und aussagekräftigeren Ergebnissen führt. Verfahren zur Untersuchung der Verzerrung in den Inferenzergebnissen.

5 Experimente und Ergebnisse

5.1 Methodik

In den ersten Wochen habe ich mir selbst ein Simulationsdesign für einen longitudinalen medizinischen Datensatz ausgedacht und basierend darauf ein gemischtes Modell gefittet. Mit diesen simulierten Daten habe ich ein reduziertes Model mit dem vollständigen Model verglichen. Die LRT Statistik habe ich dann in einem Histogramm dargestellt.

Wir fügen dem gemischten Modell einen festen Effekt hinzu, welcher keinen Einfluss auf die Trajektorie haben soll. In unserem Fall ist dieser feste Effekt das Geschlecht, welches keinen Einfluss auf den Verlauf einer Krankheit haben sollte.

Mein zweites Projekt ist nun einen hoch dimensionalen medizinischen Datensatz durch den Encoder eines Variational Autoencoders im latenten Raum zu repräsentieren und dort mit einem gemischten Model

darzustellen. Ähnlich wie zuvor will ich wieder eine LRT Statistik erhalten, in dem ich ein reduziertes Modell mit dem vollständigen Model vergleiche. Dazu trainiere ich in einer Schleife den Encoder und das gemischte Model für jeden Iterationsschritt neu und vergleiche die negativen Maximum Likelihood-Werte (ML-Werte) durch den Likelihood Ratio Test. Am Ende der Schleife erhalte ich wieder eine LRT Statistik, welche durch ein Histogramm dargestellt wird. Im Optimalfall ähnelt das Histogramm einer Chi-Quadrat-Verteilung mit einem Freiheitsgrad (Da das reduzierte Model nur einen festen Effekt, das Geschlecht, weniger hat).

5.2 Experimentelles Design

Aufbau der experimentellen Tests und Simulationen.

5.3 Durchführung

Beschreibung der durchgeführten Experimente und verwendeten Parameter.

5.4 Analyse der Ergebnisse

Diskussion der Ergebnisse im Hinblick auf die Verzerrung der Inferenz.

Wir haben nun die nötigen theoretischen Kenntnisse, um die Verzerrung der Inferenz zu messen. Um die Verzerrung zu messen haben wir uns für eine Likelihood-Ratio-Test Statistik entschieden. Dabei fügen wir dem Datensatz einen neuen Parameter hinzu, welcher keinen Einfluss auf die Daten haben sollte. Bei unserem Datensatz haben wir jedem Patienten zufällig ein Geschlecht hinzugefügt. Das Geschlecht hat keinen Einfluss auf die Testergebnisse und demnach auch keinen Einfluss auf die Response-Variable. Nun können wir das vollständige, mit dem Geschlecht ergänzte, Modell mit dem reduzierten, ohne dem Geschlecht, Modell vergleichen, in dem wir den Likelihood-Ratio-Test anwenden. Um einen Richtwert zu haben, haben wir zuerst ein selbst überlegtes Simulationsdesign erstellt, welches einen niedrigdimensionalen medizinischen Datensatz simuliert. Dabei handelt es sich um einen Datensatz der die Herzgesundheit berechnet. Für jeden der n=500 Patienten wird Health-Score basierend auf verschiedensten Einflussfaktoren berechnet. Jeder der Patienten ist erkrankt und erhält nach frühestens drei Jahren eine Behandlung, die den Einfluss der einzelnen Parametern leicht verbessert. Der Start der Behandlung wird zufällig nach drei Jahren ausgemacht. Die Daten werden über 10 Jahre erhoben. Der Heath-Score setzt sich aus dem diastolischen- und systolischen Blutdruck, dem Cholesterinspiegel, dem Triglyceride-Wert, dem Creatininspiegel und dem BMI zusammen. Die Werte werden für jeden Patienten aus einer Normalverteilung gezogen.

$$
\begin{aligned}
bp_sys &\sim \mathcal{N}(120, 10) \\
bp_dia &\sim \mathcal{N}(80, 10) \\
cholesterol &\sim \mathcal{N}(200, 30) \\
triglyceride &\sim \mathcal{N}(150, 20) \\
creatinin &\sim \mathcal{N}(1, 0.2) \\
bmi &\sim \mathcal{N}(25, 4)
\end{aligned}
$$

Insbesondere wird jedem Patienten ein zufälliges Alter zwischen 30 und 60 Jahren zugeteilt, welches ebenfalls einen minimal negativen Effekt auf den Health-Score hat. Als zusätzlichen Effekt, welcher keinen Einfluss auf die Response-Variable (in diesem Fall der Health-Score) hat, wird jedem Patienten ein Geschlecht zugeteilt. Für immer zufällig auftretende Effekte wird ein Random Slope und ein Random Intercept in den Health-Score hinzugefügt.

Basierend auf diesem Simulationsdesign, welches einem Gemischten Modell folgt, können wir nun einen Likelihood-Ratio-Test durchführen. Wir trainieren dazu ein vollständiges gemischtes Modell und ein reduziertes gemischtes Modell ohne den Effekt 'Geschlecht'. Mit den log-Likelihood Werten für die trainierten Modelle führen wir den Likelihood-Ratio-Test durch. Nach 500 Wiederholungen ergibt sich ein Histogramm der Likelihood-Ratio-Test Statistik, welches, wie zu erwarten, einer χ^2 Verteilung folgt. Wir wir in Abb sieht folgt das Histogramm der Test Statistik der Roten Kurve, welche die χ^2 Verteilung beschreibt, ohne Verzerrung. Bis auf einzelne Ausnahmen, welche durch Instabilitäten der Berechnung immer verursacht werden können, sind die Ergebnisse immer unter der χ^2 Verteilung. Dies war allerdings auch so zu erwarten, da wir ein ganz normales gemischtes Modell betrachtet hatten.

5.5 Gemischtes Modell auf latenter Datenwolke mit separatem Training

Nun wollen wir das gemischte Modell in einer latenten Repräsentation betrachten. Dazu wählen wir, wie schon angeführt, einen Variational-Autoencoder. Außerdem haben wir nun einen hochdimensionalen medizinischen Datensatz gegeben. Dieser ist einem echten Datensatz bestmöglich nachgebaut, allerdings kann hier aus Datenschutzgründen kein wirklich echter Datensatz benutzt werden. Wir wählen zunächst ein recht simples Encoder-Modell mit zwei Schichten und einer zweidimensionalen latenten Dimension. Wir trainieren zuerst den VAE separat von den gemischten Modellen. Dazu optimieren wir in der Loss-Funktion den Reconstruction Loss und die KL-Divergenz. Das vollständige und reduzierte gemischte Modell trainieren wir dann auf der latenten Datenwolke jeweils nach dem abgeschlossenen Training des VAEs. So erhalten wir wieder zwei log-likelihood Werte welche wir mit dem Likelihood-Ratio-Test auswerten können. Fassen wir alle LRT-Werte gleichermaßen wie zuvor in einem Histogramm zusammen und vergleichen mit der χ^2 Verteilung, so sehen wir, dass eine bedeutende Masse dieses Mal über der χ^2 Verteilung liegt. Wir sehen also, dass es zu einer Verzerrung der Inferenz kommt.

5.6 Gemischtes Modell auf latenter Datenwolke mit gleichzeitigem Training

Wenn wir nun versuchen das gemischte Modell zusammen mit dem VAE in einer einzigen Loss-Funktion zu trainieren, sehen wir recht schnell, dass wir so nicht zu einem gewünschten Ergebnis kommen. Bei einem gemeinsamen Training unter der Voraussetzung, dass in der Loss-Funktion alle Parameter gleich gewichtet sind, geht die χ^2 Verteilung komplett verloren. Fügen wir der Lossfunktion auch nur den Mean-Squarred-Error zwischen dem Encoder Output und dem Output des vollständig trainierten gemischten Modells hinzu und gewichten diesen nur minimal, so verlieren wir schon die χ^2-Verteilung. Das bedeutet sobald der Encoder Einfluss auf das gemsischte Modell hat, geht die gewünschte Verteilung verloren. Dementsprechend geht die Verteilung auch verloren, wenn wir

Diskussion und Fazit

1 Interpretation der Ergebnisse

Tiefere Analyse der Ergebnisse und ihrer Implikationen.

2 Vergleich mit bestehenden Arbeiten

Wie sich die Ergebnisse zu bereits veröffentlichten Forschungen verhalten.

3 Limitationen und Herausforderungen

Kritische Betrachtung der Grenzen der Studie und mögliche Probleme.

Fazit

Zusammenfassung der wichtigsten Erkenntnisse Praktische Implikationen: Wie die Ergebnisse in der Praxis angewendet werden können. Empfehlungen für zukünftige Forschungen: Vorschläge für weiterführende oder ergänzende Studien.

Anhang

Appendix

1 Supporting Data

2 Some Code Listings

Abbildungsverzeichnis

1 Aufbau eines herkömmlichen Autoencoders vi
2 Darstellung der Architektur eines Variational Autoencoders (VAE) vi
3 Architektur eines VAEs mit Reparameterization Trick x
4 Chi-Quadrat-Verteilung für verschiedene Freiheitsgrade k. xvi

1 Histogramm der LRT-Statistik für vollständiges und reduziertes gemischtes Modell auf simulierten Daten xviii
2 Simulierte Datensätze für 20 zufällig ausgewählte Patienten xix
3 Histogramm der LRT-Statistik für vollständiges und reduziertes gemischtes Modell auf latenter Datenwolke xxii

Tabellenverzeichnis

1 Einfluss und Erstellung der Parameter des Health-Scores xvii
2 Basisdaten der Patienten für Patient 0 bis 4 (baseline_df) xx
3 Testergebnisse von Patient 0 (test_scores) xx
4 Zeitbezogene Daten von Patient 0 (time_df) xxi