Albert-Ludwigs-Universität Freiburg

Bachelorarbeit

Verzerrung der Inferenz bei Verwendung gemischter Modelle in latenten Repräsentationen

Autor:
Yannick Bantel
Betreuer:
Clemens Schächter

Betreuer:
Clemens Schächter

Abgabedatum: 4. Juni 2024

Zusammenfassung

Eine kurze Zusammenfassung der Arbeit auf Deutsch.

Abstract

A brief abstract of this thesis in English.

Inhaltsverzeichnis

1 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.

1.0.1 Motivation

2 Theoretische Grundlagen

Im Vorfeld der Erörterung der Methodik dieser Arbeit ist eine theoretische Aufarbeitung der behandelten Themen unabdingbar. In diesem Kapitel erfolgt eine ausführliche Beschreibung und Behandlung der theoretischen Aspekte dieser Arbeit. Es werden sowohl lineare gemischte Modelle als auch die Theorie hinter Variational Autoencodern eingeführt und beschrieben. Im Folgenden wird insbesondere auf die für die Analyse der Modelle notwendige Theorie eingegangen, wie beispielsweise die Likelihood-Berechnung und der Likelihood-Ratio-Test.

2.1 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 [FT01] und [FKL11]. 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:

$$y_{it} = x_{it}^t \beta + z_{it}^t b_i + \epsilon_{it} \tag{2.1}$$

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,

$$b_i \sim \mathcal{N}(0, Q) \tag{2.2}$$

mit der (q x q) Kovarianzmatrix $Cov(b_i) = Q > 0$, welche symmetrisch und positiv semi-definit ist. Eine ausführliche Beschreibung findet sich in [PB00] (Kapitel 2.2.1).
Aufgrund dieser Überlegungen lässt sich nun das Model 2.1 in einer allgemeineren Form beschreiben:

Definition 2.1.1 (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 2.1.2 (Marginales gemischtes Modell)**.**
Seien die Annahmen von 2.1.1 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.),

$$\epsilon_i^* \sim \mathcal{N}(0, V_i), \quad mit \quad V_i = \sigma_\epsilon^2 I + Z_i Q Z_i^t \tag{2.3}$$

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

Definition 2.1.3 (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:

$$y = X\beta + \epsilon^* \tag{2.4}$$

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

2.2 Likelihood Inferenz

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.

2.2.1 Likelihood Berechnung gemischter Modelle

Im Folgenden wird die Schätzung der unbekannten Parameter erörtert. Der Vorliegende Ansatz basiert auf den Ausführungen von [FKL11].

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 sowohl die Kovarianzmatrix σ_ϵ als auch der Parameter Q bekannt sind. In diesem Zusammenhang ist auch V gemäß 2.4 bekannt. Für die Schätzung von β, ausgehend vom marginalen Modell, bietet sich der ML-Schätzer

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y \tag{2.5}$$

an. Dieser Schätzer für β ist für dieses Modell äquivalent zu dem Kleinste-Quadrate-Schätzer. Siehe hierzu auch [FKL11] (Kap. 3).
Betrachten wir die log-Likelihood unter der Normalverteilungsannahme für β aus dem marginalen Modell

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

und leiten sie nach β ab.

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

Somit ergibt sich der Kleinste-Quadrate-Schätzer gemäß 2.5.
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. 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 2.2.1 (Schätzer für feste und zufällige Effekte).
Sei $y = X\beta + Zb + \epsilon$ ein lineares gemischtes Modell und sei $V = Var(y)$. 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 2.2.2 (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. Die Definition für den ML Wert folgt

Definition 2.2.3 (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

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

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

2.2.2 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.

Definition 2.2.4 (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.

2.3 Variational Autoencoder

Die Anwendung der gemischten Modelle auf einer latenten Repräsentation erfolgt mittels Variational Auto-Encoder (VAE). Variational Auto-Encoder sind generative Modelle, welche versuchen die zugrunde liegende Struktur der Inputdaten x im latenten Raum zu modellieren. Im Vergleich zu normalen Autoencoder, welche den latenten Raum durch feste Punkte modellieren, wird der latente Raum in der Erweiterung VAE durch eine Wahrscheinlichkeitsverteilung (Normalverteilung) modelliert.
Die Architektur eines VAE basiert auf zwei neuronalen Netzwerken: einem Encoder, der die Inputdaten x im latenten Raum als Verteilung kodiert, und einem Decoder, der aus den latenten Daten versucht die Originaldaten zu rekonstruieren. Diese Module lernen die wesentlichen Merkmale der Eingabedaten zu extrahieren und eine komprimierte Version dieser Daten zu erzeugen.
VAEs sind für die Modellierung latenter Repräsentationen von großem Interesse, da sie hochdimensionale Datensätze mit Hilfe des Encoders im latenten Raum niedrigdimensional darstellen können. Dies reduziert die Komplexität der Modellierung und ermöglicht es, gemischte Modelle effizienter und genauer zu betreiben. Im Gegensatz zu herkömmlichen Autoencodern ist der VAE in der Lage, nicht nur den Eingabedatensatz zu rekonstruieren, sondern auch neue Inhalte zu generieren. Dies wird durch die verbesserte Repräsentation ermöglicht(vgl. [Lub23]).

Latenter Raum
Im VAE werden die latenten Variablen z aus der prior-Verteilung $p_\theta(x)$ gezogen, welche eine multivariate Normalverteilung $p_\theta(x) = \mathcal{N}(z; 0, I)$ ist. Die latenten Daten werden aus den Inputdaten durch den Encoder gezogen, welcher die posterior 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 [DPK22] verwiesen.

2.3.1 Training VAE

Ein VAE stellt ein spezifisches Beispiel einer Variational Inference (VI) dar. Die Zielsetzung einer VI besteht in der Berechnung der Posteriori-Verteilung $p(z|x)$ der latenten Variablen in Abhängigkeit von den Input-Daten. Die Berechnung dieser Verteilung ist besonders bei komplexen Modellen mit Schwierigkeiten verbunden. Infolgedessen approximiert der VAE die Posterior-Verteilung durch eine einfachere Verteilung $q_\phi(x)$. Das Ziel von VIs besteht folglich in der Minimierung des Rekonstruktionsfehlers sowie der Kullback-Leibler-Divergenz zwischen der tatsächlichen Posterior-Verteilung $p(z|x)$ und der approximierten verteilung $q_\phi(x)$.

Definition 2.3.1 (Kullback-Leibler-Divergenz (KL-Divergenz))**.**
Sei $q_\phi(z|x)$ die approximierte Posterior-Gauß-Verteilung und $p(z)$ die Prior-Gauß-Verteilung. Dann ist die KL-Divergenz definiert als

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

Da die Berechnung von $p(z)$ mit Schwierigkeiten verbunden ist, wird stattdessen die Evidence Lower Bound (ELBO) durch Stochastic Gradient Descent (SGD) oder andere ähnliche Verfahren optimiert.
Das Objekt der Optimierung im VAE ist somit der ELBO.
Im Folgenden wird die log-Likelihood der Daten betrachtet, um den ELBO herzuleiten (vgl. [DPK]).

$$\begin{aligned}
\log p_\theta(x) &= \log p_\theta(x) * \overbrace{\int q_\phi(z|x)dz}^{=1} && (2.6)\\
&= \int \log p_\theta(x) q_\phi(z|x)dz && (2.7)\\
&= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x)] && (2.8)\\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{p_\theta(z|x)}\right]\right] && (2.9)\\
&= \mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)q_\phi(z|x)}{q_\phi(z|x)p_\theta(z|x)}\right]\right] && (2.10)\\
&= \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]\right]}_{\substack{=\mathcal{L}_{\theta,\phi}(x)\\ \text{(ELBO)}}} + \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[\log\left[\frac{q_\phi(z|x)}{p_\theta(z|x)}\right]\right]}_{=D_{KL}(q_\phi(z|x)||p_\theta(z|x))} && (2.11)
\end{aligned}$$

Der zweite Term in der Gleichung ist die nicht negative Kullback-Leibler-Divergenz (KL-Divergenz) zwischen $q_\phi(z|x)$ und $p_\theta(z|x)$). Der erste Term in Gleichung 2.11 stellt den Evidence-Lower-Bound, kurz ELBO, dar:

Definition 2.3.2 (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

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

Die Umstellung der Gleichung 2.11 zeigt, dass der ELBO eine untere Schranke für die log-likelihood der Daten darstellt, da die KL-Divergenz nicht negativ ist:

$$\begin{aligned}
\mathcal{L}_{\theta,\phi}(x) &= \log p_\theta(x) - D_{KL}(q_\phi(z|x)||p_\theta(z|x)) && (2.12)\\
&\leq \log p_\theta(x) && (2.13)
\end{aligned}$$

Die Gleichung kann alternativ mit der Jensenschen Ungleichung (vgl. [Jen]) wie folgt umgestellt werden:

$$\begin{aligned}
\log p_\theta(x) &= \log \int p_\theta(x,z)dz && (2.14)\\
&= \log \int p_\theta(x,z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz && (2.15)\\
&= \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] && (2.16)\\
&\overset{Jensen}{\geq} \mathbb{E}_{q_\phi(z|x)}\left[\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathcal{L}_{\theta,\phi}(x) && (2.17)
\end{aligned}$$

Es ist sofort ersichtlich, dass die log-likelihood $p_\theta(x)$ durch Maximieren der ELBO bzgl. θ und ϕ selbst maximiert wird. Folglich verbessert sich die Qualität unseres generatives Modells. Gleichzeitig wird dadurch die KL-Divergenz der Approximation $q_\phi(z|x)$ an den wahren Posterior $p_\theta(z|x)$ minimiert. Daher wird die Approximation $q_\phi(z|x)$ optimiert. Die ELBO kann durch stochastische Gradientenverfahren optimiert werden.

Die Berechnung des 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 θ kann problematisch sein. Der Monte-Carlo-Schätzer für Gradienten ist eine gängige Methode und kann wie folgt definiert werden (vgl. [SM20]):

$$\nabla_\phi \mathbb{E}_{q_\phi(z)}[f(z)] = \mathbb{E}_{q_\phi(z)}[f(z)\nabla_{q_\phi(z)}\log q_\phi(z)] \approx \frac{1}{L}\sum_{l=1}^{L} f(z)\nabla_{q_\phi(z^{(l)})}\log q_\phi(z^{(l)}) \tag{2.18}$$

wobei $z^{(l)} \sim q_\phi(z|x^{(i)})$ ist. In der Regel stellt dies keine Herausforderung dar. Allerdings ist $q_\phi(z|x)$ nun auch abhängig von ϕ, sodass der Monte Carlo Gradienten-Schätzer nicht direkt berechnet werden kann. 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 ermöglicht die Berechnung der Gradienten der ELBO durch eine Transformation der Zufallsvariable z in eine deterministische Funktion von einer Hilfsvariablen ϵ. Sei also die latente Variable z, die aus $q_\phi(z|x)$ gezogen wurde, gegeben. Die latente Variable z wird nun als deterministische Funktion einer Hilfsvariablen ϵ, unabhängig von x und ϕ, ausgedrückt. Die Transformation sieht dann wie folgt aus

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

$g_\phi(\epsilon, x)$ ist dabei 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

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

Dadurch können die Gradienten bezüglich μ und σ effizient berechnet werden.
Wir können den Trick direkt auf Gleichung 2.18 anwenden (vgl. [SM20]). Es gilt

$$\begin{aligned}
\nabla_\phi \mathbb{E}_{q_\phi(z)}[f(z)] &= \nabla_\phi \int q_\phi(z)f(z)dz && (2.19)\\
&= \nabla_\phi \int q_\epsilon(z)f(g_\phi(\epsilon;\phi))d\epsilon && (2.20)\\
&= \nabla_\phi \mathbb{E}_{q(\epsilon)}[g_\phi(\epsilon, z)] && (2.21)
\end{aligned}$$

2.4 Verzerrung (Bias) und ihre Messung

Erklärung von Verzerrung, wie sie entsteht und wie sie gemessen wird.

3 Minimierung der Verzerrung

4 Methodik

4.1 Vorgehen

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).

4.2 Datenbeschaffung

Quellen und Typen der verwendeten Daten

4.3 Modellierungstechniken

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

4.4 Analysemethoden

Verfahren zur Untersuchung der Verzerrung in den Inferenzergebnissen.

5 Experimente und Ergebnisse

5.1 Experimentelles Design

Aufbau der experimentellen Tests und Simulationen.

5.2 Durchführung

Beschreibung der durchgeführten Experimente und verwendeten Parameter.

5.3 Analyse der Ergebnisse

Diskussion der Ergebnisse im Hinblick auf die Verzerrung der Inferenz.

6 Diskussion

6.1 Interpretation der Ergebnisse

Tiefere Analyse der Ergebnisse und ihrer Implikationen.

6.2 Vergleich mit bestehenden Arbeiten

Wie sich die Ergebnisse zu bereits veröffentlichten Forschungen verhalten.

6.3 Limitationen und Herausforderungen

Kritische Betrachtung der Grenzen der Studie und mögliche Probleme.

7 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.

8 Anhang

9 Literaturverzeichnis

A Appendix

A.1 Supporting Data

A.2 Some Code Listings

Abbildungsverzeichnis

Tabellenverzeichnis

Literatur

[DPK] M. W. Diederik P. Kingma. „Än Introduction to Variational Autoencoders", Foundations and Trends in Machine Learning“. In: ().

[DPK22] M. W. Diederik P. Kingma. „Auto-Encoding Variational Bayes“. In: (2022-12-10).

[FKL11] L. Fahrmeir, T. Kneib und S. Lang. *Regression: Methoden, Modelle und Anwendungen.* 1. Aufl. Springer, 2011. ISBN: 978-3-642-01836-7.

[FT01] L. Fahrmeir und G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models.* 2nd. Springer, 2001. ISBN: 978-1-4419-2900-6.

[Jen] *Jensensche Ungleichung.*

[Lub23] D.-I. F. S. Luber. *Was ist ein Variational Autoencoder?* 2023-12-01. URL: `https://www.bigdata-insider.de/was-ist-ein-variational-autoencoder-a-e4507ba2894e870548d87da5a975d4` (besucht am 04. 06. 2024).

[PB00] J. C. Pinheiro und D. Bates. *Mixed-Effects Models in S and S-PLUS.* New York: Springer, 2000. ISBN: 978-1-4419-0318-1.

[SM20] M. F. A. M. Shakir Mohamed Mihaela Rosca. „Monte Carlo Gradient Estimation in Machine Learning“. In: (2020-07).