Department of Mathematics
TUM School of Computation, Information and Technology
Technische Universität München

Verzerrung der Inferenz bei der Verwendung gemischter Modelle in latentent Repräsentationen

Yannick Bantel

Wissenschaftliche Arbeit zur Erlangung des Grades

Bachelor of Science

an der TUM School of Computation, Information and Technology der Technischen Universität München

Prüfer(in):
Prof. Dr. Harald Binder

Betreuer(in):
Clemens Schächter

Eingereicht:
München, den 23. Juli 202

Ich versichere hiermit, dass ich die von mir eingereichte Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, den 23. Juli 202 Yannick Bantel

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, was sie besonders in den Bereichen der Biostatistik, der Sozialwissenschaften und der ökonomischen Modellierung beliebt macht. In dem Rahmen dieser Arbeit finden die gemischten Modelle Anwendung in der medizinischen Datenanalyse. 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 verschoben. In diesem Kontext gewinnen latente Repräsentationen an Bedeutung, da sie es ermöglichen, die inhärenten Strukturen innerhalb der Daten zu identifizieren und zu nutzen, um tiefergehende Einsichten zu gewinnen.

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

Die Fragestellung der Verzerrung ist besonders relevant, da eine fehlerhafte Inferenz zu falschen Entscheidungen führen kann, die in praktischen Anwendungen schwerwiegende Folgen haben könnten. Durch eine sorgfältige Analyse und Evaluation der gemischten Modellansätze in Verbindung mit latenten Repräsentationen strebt diese Arbeit an, einen Beitrag zur Verbesserung der Modellgenauigkeit und der Verlässlichkeit von Inferenzschlüssen zu leisten.

Diese Arbeit gliedert sich in mehrere Teile, die zunächst die theoretischen Grundlagen der gemischten Modelle und der latenten Repräsentationen behandeln, gefolgt von einer Diskussion der Methoden zur Messung und Korrektur von Verzerrungen. Anhand von experimentellen Studien werden diese Konzepte dann praktisch angewendet und evaluiert, um abschließend Empfehlungen für die Anwendung dieser Techniken in der Forschung und Praxis zu geben.

1.0.1 Motivation

2 Theoretische Grundlagen

Definition, Typen und Anwendungsbereiche.
Bevor wir uns die Methodik dieser Arbeit anschauen benötigt es noch eine theoretische Aufarbeitung der behandelten Themen. In diesem Kapitel werden die theoretischen Aspekte dieser Arbeit umfangreich beschrieben und aufgearbeitet. Sowohl lineare gemischte Modelle als auch die Theorie hinter Variational Autoencodern werden eingeführt und beschrieben. Insbesondere wird in diesem Kapitel die Theorie, welche zur Analyse der Modelle notwendig ist, wie zum Beispiel die Likelihood Berechnung und der Likelihood Ratio Test, behandelt

2.1 Gemischte Modelle

Um die theoretischen Aspekte gemischter Modelle aufzuarbeiten folgen wir den Notierungen in dem Buch [PB00] Mixed-Effects Models in S and S-Plus von José C. Pinheiro und Douglas M. Bates [PB00].

Ein gemischtes Modell (Mixed Modul) ist ein statistisches Datenanalyseverfahren, welches sowohl feste als auch zufällige Effekte (fixed und random Effects) modelliert. Ihre Anwendung finden die gemischten Modelle haupsächlich in der Analyse von Longitudinaldaten und Clusterdaten.
Die parametrisierten festen und zufälligen Effekte berechnen zusammen mit einem Fehlervektor die Antwortvariable. Das lineare gemischte Modell für eine Gruppe wird folgendermaßen definiert.

Definition 2.1.1 (Lineares gemischtes Modell für Longitudinal- oder Clusterdaten)**.**
Seien X (n_i x p) und Z (n_i x q) bekannte Designmatrizen für die festen und zufälligen Effekte. Seien β ein p-dimensionaler Vektor von p festen Effekten und b ein q-dimensionaler Vektor von q zufälligen Effekten und sei ϵ ein n_i-dimensionaler normalverteilter Fehlervektor.

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

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

$$b_i \sim \mathcal{N}(0, \sum), \epsilon_i \sim \mathcal{N}(0, \sigma^2 I)$$

definiert.

Dabei sind die Daten der zufälligen und festen Effekte 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. Für die immer auftretenden Messfehler oder unerwartete Einflüsse wird ein zufälliges Rauschen ϵ hinzugefügt.

Die zufälligen Effekte b_i und der Fehlervektor sind unabhängig voneinander in der selben Gruppe und unabhängig von anderen Gruppen. Wie man leicht erkennen kann, ist der Vektor der zufälligen Effekte ausschließlich von seiner Varianz-Kovarianz-Matrix $\sum$ charackterisiert, welche symmetrisch und postiv semidefinit ist.
Die einzelnen Cluster/Gruppen können zu einem einzigen allgemeinen linearen gemischten Modell zusammengefasst werden:

Definition 2.1.2 (Allgemeines lineares gemischtes Modell)**.**
Ein lineares gemischtes Modell ist definiert durch

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

mit

$$\begin{pmatrix} b \\ \epsilon \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sum & 0 \\ 0 & \sigma^2 I \end{pmatrix}\right)$$

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.

Die Vektoren b und ϵ sind also normalverteilt mit Varianz $Var(b) = \sum$ und $Var(\epsilon) = R = \sigma^2 I$. Also lässt sich die Varianz von y schreiben als $Var(y) = V = Z \sum Z^t + R$.

2.2 Likelihood Inferenz

Um den Einfluss eines festen Effekts zu testen vergleichen wir ein reduziertes Modell, ohne diesen festen Parameter, mit dem vollständigen Modell. Dazu nutzen wir die Likelihood-Ratio-Test (LRT) Statistik, welche üblicherweise dazu genutzt wird. Wie das Testverfahren genau funktioniert und wie die Implementierung des LRT erfolgt, schauen wir uns später an. Zuvor benötigen wir noch ein bisschen Theorie über die Likelihood-Berechnung.

2.2.1 Likelihood Berechnung gemischter Modelle

In diesem Abschnitt wird die Schätzung der unbekannten Parameter behandelt. Hier folgen wir der Herangehensweise von [**fahrmeir2010**] .

Wir nutzen die Maximum-Likelihood (ML) Methode zur Berechnung der Schätzer. Eine Alternative dazu wäre die restringierte ML Methode, welche allerdings nicht geeignet für den Likelihood-Ratio-Test ist. Deshalb greifen wir bei unserer Berechnung auf die ML-Methode zurück.

Die Schätzung der Parameter ist in einem gemischten Modell allerdings etwas komplizierter. Es ist nicht nur β unbekannt, sondern auch die b, $\sum$ und R. Es sind also sowohl die festen und zufälligen Effekte als auch die unbekannten Parameter, welche wir mit θ bezeichnen, in $\sum$ und R zu schätzen. Dies zwingt uns zu einer verschachtelten Schätzung.

Nehmen wir erst einmal an, dass X, V und Z bekannt sind.Für die Schätzung von β, ausgehend vom marginalen Modell,bietet sich der ML-Schätzer

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

an. Setzen wir hier $\hat{V}$ ein erhalten wir den besten linearen erwartungstreuen Schätzer (BLUE, best linear unbiased estimator) für die festen Effekte.
Für den Schätzer von b nutzen wir den bedingten Erwartungswert $E(b|y)$ von b, gegeben die Daten y.
Betrachten wir die gemeinsame Verteilung von b und y

$$\begin{pmatrix} y \\ b \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} V & Z\sum \\ \sum Z^t & \sum \end{pmatrix}\right)$$

erhalten wir $E(b|y) = \sum Z^t V^{-1}(y - X\beta)$.
Ersetzt man nun β durch den Schätzer $\hat{\beta}$ erhält man den Schätzer

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

für die zufälligen Effekte. 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

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

ein Schätzer für die zufälligen Effekte.

Wie schon erwähnt soll θ der Parametervektor sein, der alle unbekannten Parameter in $V = V(\theta), \sum = \sum(\theta)$ und $R = R(\theta)$ enthält. Falls wir nun einen Schätzer $\hat{\theta}$ haben, können wir die Kovarianzschätzer und somit auch die Schätzer der festen und zufälligen Effekte durch Einsetzen von $\hat{\theta}$ berechnen. Die ML-Methode für θ basiert auf dem marginalen Modell

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

. 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 liefert

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

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

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

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

Definition 2.2.2 (Kovarianz-Schätzer)**.**
Sei $y = X\beta + Zb + \epsilon$ ein lineares gemischtes Modell mit

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

und sei θ der unbekannte Parametervektor von $\sum$,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.

Definition 2.2.3 (Maximum-Likelihood)**.**
Sei

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

2.2.2 Likelihood-Ratio-Test

Mit der vorausgegangen Theorie über die ML-Methode ist die Berechnung der Likelihood Ratio Test (LRT) Statistik relativ einfach. Zur Erinnerung wollen wir ein reduziertes Modell mit dem vollständigen Modell vergleichen, um herauszufinden wie groß der Einfluss einer Störgröße ist. Dafür nutzen wir den LRT.

Definition 2.2.4 (Likelihood-Ratio-Test (LRT))**.**
Sei $\mathbf{L}_{full}$ *der Likelihood-Wert des vollständigen Modells und* $\mathbf{L}_{red}$ *der Likelihood-Wert des reduzierten Modells. Sei i die Anzahl der Freiheitsgrade. Dann ist die LRT Statistik gegeben durch*

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

Falls $\mathbf{L}_{full}$ und $\mathbf{L}_{red}$ wie in der Definition initialisiert sind, gilt $\mathbf{L}_{full} > \mathbf{L}_{red}$. Insbesondere gilt $\log(\mathbf{L}_{full}) > \log(\mathbf{L}_{red})$. Falls $\mathbf{L}_{full}$ und $\mathbf{L}_{red}$ nun schon die Log-Likelihood-Werte der Modelle sind, gilt für die Berechnung der LRT Statistik einfach $2(\mathbf{L}_{full} - \mathbf{L}_{red})$.

2.3 Variational Autoencoder

Ein Variational Autoencoder (VAE) besteht aus einem Encoder und einem Decoder. VAE sind sehr attraktiv für latente Repräsentationen, da sie hoch dimensionale Datensätze durch den Encoder im latenten Raum niedrig dimensional darstellen können. Der Decoder versucht dann aus der Darstellung im latenten Raum den originalen Datensatz zu rekonstruieren.
Bedeutung und Anwendung in verschiedenen Bereichen wie Deep Learning, Faktoranalyse etc.

2.4 Verzerrung (Bias) und ihre Messung

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

3 Methodik

3.1 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).

3.2 Datenbeschaffung

Quellen und Typen der verwendeten Daten

3.3 Modellierungstechniken

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

3.4 Analysemethoden

Verfahren zur Untersuchung der Verzerrung in den Inferenzergebnissen.

4 Experimente und Ergebnisse

4.1 Experimentelles Design

Aufbau der experimentellen Tests und Simulationen.

4.2 Durchführung

Beschreibung der durchgeführten Experimente und verwendeten Parameter.

4.3 Analyse der Ergebnisse

Diskussion der Ergebnisse im Hinblick auf die Verzerrung der Inferenz.

5 Diskussion

5.1 Interpretation der Ergebnisse

Tiefere Analyse der Ergebnisse und ihrer Implikationen.

5.2 Vergleich mit bestehenden Arbeiten

Wie sich die Ergebnisse zu bereits veröffentlichten Forschungen verhalten.

5.3 Limitationen und Herausforderungen

Kritische Betrachtung der Grenzen der Studie und mögliche Probleme.

6 Schlussfolgerung und Ausblick

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.

7 Anhang

8 Literaturverzeichnis

A Appendix

A.1 Supporting Data

A.2 Some Code Listings

Abbildungsverzeichnis

Tabellenverzeichnis

Literatur

[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.