DISSERTATION

Sparse and robust modeling for high-dimensional data

ausgeführt zum Zwecke der Erlangung des akademischen Grades eines
Doktors der technischen Wissenschaften

unter der Leitung von
Univ.-Prof. Dipl.-Ing. Dr.techn. Peter Filzmoser,
Institut für Stochastik und Wirtschaftsmathematik (E105)

eingereicht an der Technischen Universität Wien an der Fakultät für Mathematik und Geoinformation von

Dipl.-Ing. Irene Hoffmann

Matrikelnummer 0825098

Diese Dissertation haben begutachtet:

Peter Filzmoser	Tim Verdonck	Beata Walczak
TU Wien	KU Leuven	University of Silesia

Wien, 10. Oktober 2017

Irene Hoffmann

Abstract

The development of statistical methods for high-dimensional data has become an important focus in recent research. Classical regression and classification approaches require full rank data matrices, with more observations than variables. In many areas of application (e.g. bioinformatics and chemometrics) this assumption is not met. Sparse methods describe a class of approaches where a penalty is imposed on the coefficient estimate to favour exact zero values and so intrinsically perform variable selection.

Another challenge in many applications are outliers in the data, which are observations that do not follow the structure of the majority of the data and so violate the distribution assumptions which are necessary for classical model estimation. Robust methods give stable estimates when outliers are present and model the relationship of the majority of the data.

The focus of this thesis is on the development of regression and classification methods, which are appropriate for high-dimensional data and data with outliers. Sparse partial robust M regression is a robust and sparse regression method. A robust subspace is identified, including only a subset of the original variables, where a robust regression model is estimated. This approach is then extended to binary classification problems. With the help of the optimal scoring approach, regression methods can be applied to classification problems. Robust sparse optimal scoring is a classification method based on least trimmed squares regression. Finally, sparse and robust linear regression and logistic regression methods are introduced based on least trimmed squares with an elastic net penalty, which induces sparsity and at the same time favours similar coefficient estimates for highly correlated variables.

Kurzfassung

Ein wichtiger Fokus in der Entwicklung neuer statistischer Methoden liegt seit einigen Jahren auf der Analyse hochdimmensionaler Daten. Klassische Regressions- und Klassifikationsmethoden benötigen Datenmatrizen mit vollem Rang, die mehr Beobachtungen als Variablen beinhalten. In vielen Anwendungsgebieten (z.B. der Bioinformatik oder Chemometrie) kann diese Anforderung aus praktischen Gründen nicht erfüllt werden. *Sparse modeling* umfasst eine Klasse von Methoden, die durch einen Strafterm Nullwerte bei der Koeffizientenschätzung bevorzugen und dadurch intrinsisch Variablen selektieren.

Eine weitere Herrausforderung in vielen Anwendungsgebieten sind Ausreißer in den Daten. Als Ausreißer werden Beobachtungen bezeichnet, die nicht der Struktur oder dem Trend der Mehrheit der Daten entsprechen und dadurch die Verteilungsannahmen klassischer Methoden verletzen und Modellschätzungen verzerren. Robuste Methoden beschränken den Einfluss extremer Werte auf die Modellschätzung und liefern stabile Modelle.

Diese Arbeit befasst sich mit robusten Methoden, die *sparse modeling* Ansätze integrieren und dadurch anwendbar auf hochdimensionale Daten sind. *Sparse partial robust M regression* ist eine robuste Methode, die partielle kleinste Quadrate Regression mit *sparse modeling* verbindet. Die latenten Variablen eines niedrigdimensionalen Raumes werden aus Linearkombinationen einer Teilmenge der originalen Variablen erzeugt. Mit den latenten Variablen wird ein robustes Regressionsmodell erzeugt. Die Methode wird für binäre Klassifikationsprobleme erweitert. *Robust sparse optimal scoring* ist eine weitere robuste Klassifikationsmethode, die auch auf Mehrgruppenprobleme angewandt werden kann und auf *least trimmed squares regression* basiert. Zuletzt werden zwei robuste Methoden vorgestellt, die durch einen *elastic net* Strafterm sowohl Variablenselektion integrieren als auch regularisierend auf die Koeffizienten wirken, wenn Variablen stark korrelieren.

Contents

Acknowledgements

First of all, I would like to thank my supervisor Prof. Peter Filzmoser who encouraged me and supported me in many ways. Thank you for all the opportunities and the contacts you provided me with, for your advise and many fruitful discussions.

I would also like to thank Prof. Varmuza for the insight he gave me into chemometrics and spectroscopy and for the many discussions on data preprocessing. Many thanks also to Sven Serneels and Prof. Christophe Croux. I learned many things from you and I highly appreciate the opportunity of working together.

My special thanks go to my friends and colleagues who shared not only an office with me, but also many challenging and joyful moments. Thank you to all my friends and especially to my family for encouraging words, for trying to understand my research and for practical support.

Finally, I express my gratitude to the Austrian Science Fund (FWF) for financial support through the project P 26871-N20.