DISSERTATION

Sufficient Dimension Reduction for Structured and Big Data

ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften unter der Leitung von

Univ. Prof. Ph.D Efstathia Bura
E105 – Institut für Stochastik und Wirtschaftsmathematik, TU Wien
Angewandte Statistik

eingereicht an der Technischen Universität Wien
Fakultät für Mathematik und Geoinformation

von

Daniel Kapla
Matrikelnummer: 01128052
Lokalbahnzeile 67
2500 Baden

Diese Dissertation haben begutachtet:

1. **Univ. Prof., Ph.D. Efstathia Bura**
   Institut für Stochastik und Wirtschaftsmathematik, Technische Universität Wien, Österreich
2. **Prof., Ph.D. Liliana Forzani**
   Facultad de Ingeniería Química, Universidad Nacional del Litoral, Argentina
3. **Prof., Ph.D. Xin Zhang**
   Department of Statistics, Florida State University, USA

Wien, am 14. März 2024

Kurzfassung

Regression modelliert die bedingte Verteilung einer oder mehrerer Zufallsvariablen (Antwortvariablen) gegeben den Prädiktoren. Wenn die Prädiktoren hochdimensional sind, wird die Regressionsaufgabe noch herausfordernder. *Sufficient Dimension Reduction* (SDR; auf Deutsch: Ausschöpfende Dimensions Reduktion) begegnet diesem Problem, indem die Dimensionalität der Prädiktoren reduziert wird, während sämtliche relevante Information über die Zielvariable erhalten bleibt. Unter den SDR Methoden tragen lineare Projektionen auf niedrigdimensionale Teilräume nicht nur dazu bei, das Problem handhabbarer zu machen, sondern auch – was besonders wichtig ist – die Interpretierbarkeit zu verbessern.

In dieser Arbeit werden drei lineare SDR Methoden oder Methodenklassen vorgestellt. Die erste Methode, *Neural Network SDR* (NNSDR), behandelt effizient Datensätze mit enormem Datenvolumen und hoher Prädiktor-Dimensionalität. NNSDR kombiniert die lineare Projektion der Prädiktoren mit dem gleichzeitigen Lernen neuronaler Netzwerke. Dabei wird sowohl von der Anpassung des neuronalen Netzwerks als auch von vorwärts SDR gleichzeitig profitiert, um einen Prozess aufzubauen, der eine Vorhersagegenauigkeit bietet, die mit Neuronalen Netzwerken vergleichbar ist. Gleichzeitig bleibt die Transparenz bezüglich der Relevanz der Prädiktoren und der intrinsischen Regressionsdimension erhalten.

Die heute gesammelten Daten können strukturelle Merkmale aufweisen, die von früheren Regressionsmethoden nicht berücksichtigt werden und zu einem Verlust in Inferenz führt. Unsere zweite Methodenklasse führt SDR Inferenz bei Regressionen mit matrixwertigen Prädiktoren durch. Das allgemeine Problem eines Inferenzmodels für die bedingte Verteilung der Antwortvariablen gegeben tensorwertige Prädiktoren wird in unserer dritten Methodenklasse behandelt. *Generalized Multi-Linear Model* SDR ermöglicht flexible lineare Reduktionen, die die Informationen in beliebigen tensorwertigen Daten mit Verteilung in der quadratischen Exponentialfamilie erfassen. Wir zeigen die Konsistenz und asymptotische Normalität der ausschöpfenden Reduktion. Für kontinuierliche tensorwertige Prädiktoren entwickeln wir ein rechnerisch effizientes Schätzverfahren für ihre ausschöpfenden Reduktionen, das auch auf Situationen anwendbar ist, in denen die Dimension der Reduktion die verfügbare Stichprobengröße übersteigt. Für Regressionen mit binären tensorwertigen Prädiktoren orientiert sich das Schätzverfahren an Algorithmen, die zum Trainieren neuronaler Netzwerke verwendet werden, um große Mengen von Beobachtungen zu verarbeiten, die in allgemeinen hochdimensionalen diskreten Umgebungen oft erforderlich sind.

Abstract

Regression models the conditional distribution of a random variable(s) (response(s)) given a set of predictors. When the predictors are high-dimensional, the regression task becomes even more challenging. *Sufficient Dimension Reduction* (SDR) addresses this issue by reducing the dimensionality of the predictors while retaining all relevant information about the response variable. Among SDR methods, linear projections to lower dimensional subspaces make the problem not only more manageable but, importantly, also more interpretable.

Three linear SDR methods, or class of methods, are presented in this thesis. The first method, *Neural Network SDR*, efficiently handles datasets of huge data size and high predictor dimensionality. NNSDR combines linear projection of the predictors with neural network learning simultaneously borrowing strength from NN fitting and forward SDR to build a process that enjoys predictive accuracy comparable to NN methods while being transparent about predictor importance and intrinsic regression dimension.

The data collected today can exhibit structural features that past regression techniques do not take into consideration and that may result in inferential loss. Our second class of methods carry out SDR inference on regressions with matrix-valued predictors. The general problem of building inference models for the conditional distribution of a response given multi-way array-valued predictors is addressed in our third class of methods. *Generalized Multi-Linear Model* SDR, allows for flexible linear reductions capturing the information in arbitrary tensor-valued data with distribution in the quadratic exponential family. We prove the consistency and asymptotic normality of the sufficient reduction. For continuous tensor-valued predictors, we develop a computationally efficient estimation procedure of their sufficient reductions, which is also applicable to situations where the dimension of the reduction exceeds the available sample size. For regressions with binary tensor-valued predictors, the estimation procedure draws inspiration from algorithms used for training neural networks to be able to process large amounts of observations often required in general high-dimensional discrete settings.