High-dimensional estimation using graphical models and clustering

Author(s):
Rütimann, Philipp Arthur

Publication Date:
2012

Permanent Link:

https://doi.org/10.3929/ethz-a-007558902 →

Diss. ETH No. 20489

High-dimensional estimation using graphical models and clustering

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by
PHILIPP ARTHUR RÜTIMANN
MSc ETH Math.
born June 25, 1983
citizen of Zug ZG

accepted on the recommendation of
Prof. Dr. Peter Bühlmann, examiner
Prof. Dr. Sara van de Geer, co-examiner

2012

Abstract

In the last few years high-dimensional statistics gained a lot of attention due to the increasing occurrence of high-dimensional data sets in many scientific fields. In high-dimensional statistical data, the number of unknown model parameters exceeds the number of observations by orders of magnitude. This poses several problems.
For the simple task of estimating the covariance matrix, the usual Gaussian maximum likelihood estimator is ill-posed in the high-dimensional setup. That is, to make high-dimensional statistical inference possible, we need to make additional assumptions.
A very common requirement is the assumption of a sparse setting, where only a few unknown model parameters are different from zero.

In the case of high-dimensional sparse covariance matrix estimation, sparsity means that only some of the matrix entries are non-zero. This can be achieved with various methods. Most of the techniques either regularise the estimated matrix or its Cholesky factor. However, in this thesis, we study a different approach. We consider a two step procedure where we first infer the sparse structure of the matrix and then estimate the entries of the matrix according to the structure.

Considering the high-dimensional regression problem, a very popular estimator is the Lasso. It enforces sparsity by setting some model parameters exactly to zero. But the Lasso has problems with correlated designs. Therefore, we show in this thesis how the Lasso handles correlated designs and present additional techniques in order to deal with this problem.

In Chapters 2 and 3 of this thesis we address the issue of high-dimensional sparse covariance estimation with two different approaches using graphical models. For both approaches we present theoretical results and show some

empirical outcomes by simulations and real data examples.

In Chapter 4 we consider the high-dimensional linear model with correlated design. We present a two step procedure combining clustering and penalized maximum likelihood estimation, in order to deal with the problems of the Lasso in such correlated designs.

In Chapter 5 we look at the problem of ranking causal intervention effects based on p-values. In this chapter the results have a more applied nature since most of the findings are based on simulations and real data examples.

Zusammenfassung

In den letzten Jahren wurde die hochdimensionale Statistik immer wichtiger, dies aufgrund der Zunahme von hochdimensionalen Datensätzen in vielen verschiedenen wissenschaftlichen Bereichen. Hochdimensional bedeutet, dass wir Datensätze betrachten in denen die Anzahl Beobachtungen viel kleiner ist als die Anzahl der Modellparameter. Dies wirft verschiedene Probleme auf.
Das heisst, wir brauchen zusätzliche Annahmen an unsere Modelle, um hochdimensionale statistische Inferenz möglich zu machen.
Eine weit verbreitete Annahme ist die der dünnen Besetztheit, bei der angenommen wird, dass nur wenige Modellparameter von Null verschieden sind.

Im Falle der hochdimensionalen Schätzung der Kovarianzmatrix bedeutet die Annahme der dünnen Besetztheit, dass nur wenige Matrixeinträge nicht Null sind. Dies wird erreicht durch unterschiedliche Ansätze. Die meisten dieser Ansätze beruhen auf dem Regularisieren der geschätzten Matrix selbst oder ihres Cholesky Faktors. In dieser Arbeit wird hingegen ein anderer Ansatz verfolgt. Wir betrachten eine zweistufige Prozedur, bei welcher wir im ersten Schritt die Struktur der Matrix erschliessen und danach basierend auf der Struktur die Einträge schätzen.

Für hochdimensionale Regressionsprobleme ist der Lasso sehr beliebt. Er generiert dünn besetzte Schätzungen, indem einige Regressionskoeffizienten exakt zu Null geschätzt werden. Doch der Lasso hat gewisse Probleme mit stark korrelierten Variablen. Deshalb zeigen wir in dieser Arbeit auf, wie der Lasso mit stark korrelierten Variablen umgeht und präsentieren zwei Lösungsansätze für dieses Problem.

In den Kapiteln 2 und 3 dieser Dissertation wird die Schätzung von hochdimensionalen dünn besetzten Kovarianzmatrizen mittels zwei verschiedenen graphischen Modellen studiert. Für beide werden theoretische Resultate und

Simulationsstudien präsentiert.

In Kapitel 4 wird das hochdimensionale lineare Modell mit stark korrelierten erklärenden Variablen betrachtet. Wir präsentieren ein zweistufiges Modellierungs- und Schätzverfahren als Alternative zum Lasso, welcher bei stark korrelierten Variablen Probleme aufzeigt.

In Kapitel 5 geht es um das Rangieren von kausalen Intervetionseffekten mittels P-Werten. In diesem Kapitel sind die Resultate eher angewandter Natur, da die meisten Erkenntnisse auf Simulationen und Datenbeispielen beruhen.