Variable Selection and Potential Omitted Variable Bias

Project Proposal

Computational Statistics SS 2020

Instructor: JProf. Dr. Lena Janys

Max Schäfer

University of Bonn

July 26, 2020

Abstract

Abstract–

1 Idea for Term Paper

1.1 Motivation

In many observational studies a critical assumption to obtain unbiased estimates is *unconfoundedness* which requires that all variables explaining both the outcome and the variable of interest are included in the model. Given all confounding variables are observable, including them into the model solves this issue. However, in higher dimensions there may be a rationale to omit variables from the model. Some ML methods used for model selection – such as the *least absolute shrinkage and selection operator* (LASSO) – may drop variables based on their weak association with the outcome (given the variable of interest) albeit they are highly predictive of the variable of interest.

In the term paper I want to investigate this conflict of introducing biased coefficient estimates by omission of confounders in a setting where refraining from variable selection is not feasible [1]. By means of a simulation study I want to focus on the model selection properties of LASSO in general, and in settings with confounding in particular. I am curious if a data-driven rule-of-thumb in addition to subjective judgment (e.g. findings from earlier literature) can be found which helps to omit confounders less often.

Until now I could not decide upon a paper from which I want to derive my data generating process and research question from. Below are a few ideas I want to look into, but whose practicability and implementation require an informed decision as proposed methods and assessments highly depend upon the final research question and data at hand:

- To incorporate information which variables may be prone to be confounders I plan to assess two different model selection procedures where the set of covariates is selected by modeling the outcome only, and by modeling both the variable of interest and the outcome.
- Performance assessment of the adaptive LASSO where weights for variables can be defined to incorporate prior knowledge of potential confounders or important variables.

[1] In very high dimensions with more variables than observations in our dataset, we do not have a unique solution. Also, there is a rationale for variable omission if we have a strong believe the true model is sparse and we want to enhance interpretability.

- Prior lasso
- Group lasso
- IPF-Lasso

Addon