Feature Selection under Multicollinearity
&
Causal Inference on Time Series

Indranil Bhattacharya
MSc(Engg.), CSA, IISc

February 27, 2017

Abstract

In this work, we study and extend algorithms for *Sparse Regression* and *Causal Inference* problems. Both the problems are fundamental in the area of Data Science.

The goal of regression problem is to find out the "best" relationship between an output variable and input variables, given samples of the input and output values. We consider *sparse regression* under a high-dimensional *linear* model with *strongly correlated* variables, situations which cannot be handled well using many existing model selection algorithms. We study the performance of the popular feature selection algorithms such as LASSO, Elastic Net, BoLasso, Clustered Lasso as well as Projected Gradient Descent algorithms under this setting in terms of their running time, stability and consistency in recovering the true support. We also propose a new feature selection algorithm, BoPGD, which cluster the features first based on their sample correlation and do subsequent sparse estimation using a bootstrapped variant of the projected gradient descent method with projection on the non-convex L_0 ball. We attempt to characterize the efficiency and consistency of our algorithm by performing a host of experiments on both synthetic and real world datasets.

Discovering *causal* relationships, beyond mere *correlation*, is widely recognized as a fundamental problem. The Causal Inference problems use observations to infer the underlying causal structure of the data generating process. The input to these problems is either a multivariate time series or i.i.d sequences and the output is a *Feature Causal Graph* where the nodes correspond to the variables and edges capture the direction of causality. For high dimensional datasets, determining the causal relationships becomes a challenging task because of the *curse of dimensionality*. Graphical modeling of temporal data based on the concept of "Granger Causality" has gained much attention in this context. The blend of Granger methods along with model selection techniques, such as LASSO, enables efficient discovery of a "sparse" subset of causal variables in high dimensional settings. However, these temporal causal methods use an input parameter, L, the *maximum time lag*. This parameter is the maximum gap in time between the occurrence of the output phenomenon and the causal input stimulus. However, in many situations of interest, the maximum time lag is not known, and indeed, finding the range of causal effects is an important problem. In this work, we propose and evaluate a data-driven and computationally efficient method for Granger causality inference in the Vector Auto Regressive (VAR) model without foreknowledge of the maximum time lag. We present two algorithms *Lasso Granger++* and *Group Lasso Granger++* which not only constructs the hypothesis feature causal graph, but also simultaneously estimates a value of maxlag ($\hat{L}$) for each variable by balancing the trade-off between "goodness of fit" and "model complexity".