Fighting Multicollinearity in Double Selection: A Bayesian Approach

Mateo Graciano Londoño*
Andrés Ramírez Hassan†

Universidad EAFIT
Medellín, Colombia

March 7, 2016

1 Problem statement

Consider the following structure (Belloni et al., 2014):

$$y_i = \alpha d_i + x_i^{'}\beta_g + \epsilon_i \tag{1}$$

$$d_i = x_i^{'}\beta_m + \zeta_i \tag{2}$$

where y_i is the response, β_g, β_m the structural and treatments effects of variables x_i respectively, d_i is the treatment, α is the treatment effect and ϵ_i, ζ_i are stochastic errors such that

$$E\left[\epsilon_i \mid x_i, d_i\right] = E\left[\zeta_i \mid x_i\right] = 0$$

Let n be the number of observations and $p = dim(x_i)$ with $p + 1 \gg n$. Considering the latter inference under OLS would be impossible given the absence of degrees of freedom $(n - p)$. One can say that given in (1) the most important value is α which is the impact of d_i over y_i so, it would be a good idea to select only a few variables (s) in x_i so that $n > s + 1$.

To select which variables to include is a question as important as the estimation process, and for that duty two different techniques are the LASSO estimator and Markov chain Monte Carlo model composition (MC^3) which are different approaches to the same problem, model selection.

The LASSO estimator consider an optimization problem as the following for the case of a simple lineal model:

$$\beta^* = \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[d_i - x_i^{'}\beta_m\right]^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{3}$$

where λ is a penalization coefficient. Let T be

$$T = \left\{ j \in 1, 2, ..., p \quad : \quad \mid \beta_j^* \mid > 0 \right\}$$

*Student of Mathematical Engineer, Universidad EAFIT , Medellín, Colombia
†Tutor Professor, Department of Economics, Universidad EAFIT , Medellín, Colombia

the post-LASSO estimator is defined as:

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[d_i - x_i^{'} \beta_m \right]^2 \quad : \quad \beta_j = 0 \quad \forall j \notin T$$

MC^3 is a Bayesian methodology which uses a stochastic search comparing different models by its posterior model probability.
Following Simmons et al. (2010), let $M = \{M_1, M_2, ..., M_m\}$ be the set of models under consideration, and d the observed data as in (2). The posterior model probability for model M_j is defined as:

$$P(M_j \mid d, M) = \frac{P(d \mid M_j)\pi(M_j)}{\sum_{i=1}^{m} P(d \mid M_i)\pi(M_i)} \qquad \forall j = 1, 2, ..., m$$

where

$$P(d \mid M_j, M) = \int ... \int P(d \mid \alpha_j, \ \ M_j)\pi(\alpha_j \mid M_j) d\alpha \qquad \forall j = 1, 2, ..., m$$

is the integrated likelihood of the model M_j, α_j is the vector of parameters of the model M_j, $\pi(\alpha_j \mid M_j)$ is the prior of parameters under M_j, $P(d \mid \alpha_j, \ M_j)$ is the likelihood and $\pi(M_j)$ is the prior of the prior probability that M_j is the true model.

2 Objectives

2.1 General objective

To propose a methodology based on MC^3 in order to compare its performance on the inference of a treatment based on frequentist results given by the post-double-LASSO (PD-LASSO) estimators.

2.2 Specific objectives

- Implement the PD-LASSO and MC^3 on simulations exercises.
- Gather real information and use both methodologies.
- Compare both methodologies and analyse how they perform based on simulation and real cases.

3 Preceding research

Econometricians always want to show the relationship between some variables (regresors) and a specific variable such as the gross domestic product (gdp), but maybe which variables are the ones which explain in a better way is one of the most important question. There are different ways to answer the latter, so that is the reason of why model selection have had attention among researchers, to know which is the best group of variables.

For instance, Tibshirani (1996) develop a methodology which shrinkage a linear model which leads to answer the question of which is the best model, on the other hand, there are the Bayesians methods which lead to answer the same questions but from a different perspective as we can see in a review made in Wasserman (2000) who gives a review in what is the basics of the Bayesian methodology for model selection via the posterior model probability.

Belloni et al. (2014) worked on the inference of a treatment effect using a model which had a lot of possible regresors and performing a model selection using a PD-LASSO estimation which is using the idea of the LASSO estimation as in Belloni and Chernozhukov (2011) but in two stages and also show that it had a better performance than the original methodology.

4 Justification

Usually, econometricians face a concern, which consists in not identifying the variables that can be useful for the model. Thus, it is commonly seen among researchers and may not help in the selection of a set of controls among a group of variables (Belloni et al., 2014).

The PD-LASSO technique, following the intuition on Tibshirani (1996) but with Belloni and Chernozhukov (2011) implementation, is an alternative used with the purpose of increasing accuracy in the variable selection process. Therefore, the principal aim of this project is to design an analog strategy to PD-LASSO using a Bayesian method called MC^3 for the issue of model selection in order to explain the effect of a treatment over a variable.

5 Scope

The project focuses on the development of a methodology based on MC^3 following the idea of the PD-LASSO estimator in Belloni et al. (2014). The benefits of using MC^3 had been widely prove as in Johnson and Rossell (2012); Simmons et al. (2010); Wasserman (2000); Eicher et al. (2012) but using a idea of a double selection as in the PD-LASSO estimator is clearly an interesting idea which could lead to excellent results in the model selection.

Furthermore, the idea is to compare both methodologies and see if our proposed methodology leads to better inference on treatment effects.

6 Methodology

The first stage is understanding the LASSO estimator and replicate a Monte Carlo simulation using the PD-LASSO procedure in Belloni et al. (2014) and then use MC^3 and see how it perform using the same simulated data.

The second and last stage is to replicate the exercise using real data as in Donohue III and Levitt (2001) for both approaches and compare how both perform.

7 Schedule

Dates	Activity
February 1st - 29th	Study Bayesian econometrics
February 7th - February 21st	Literature Review and PD-LASSO implementation
February 1st - 12nd	pre-project
February 19th	Proposal presentation
March 1st - April 8th	Methodology development
April 8th	Oral progress report
April 8th - April 22nd	Gather real data
April 22nd - May 6th	Check performance
May 1st - 20th	Write the final report
May 20th	Final project report
May 20th - June 7th	Preparation of final project presentation
June 7th	Final project presentation

8 Budget

This research will not required any budget, because EAFIT University provides data bases for the literature review, software licenses to implement the computer model and the tutor professor.

9 Intellectual property

According to the internal regulation on intellectual property within EAFIT University, the results of this investigation practice are product of Mateo Graciano Londoño as student and Andrés Ramírez Hassan as tutor professor.

In case further products, beside academic articles, should be generated from this work, the intellectual property distribution related to them will be directed under the current regulation of this matter determined by EAFIT University (Universidad EAFIT, 2009).

References

Belloni, A. and Chernozhukov, V. (2011). *High dimensional sparse econometric models: An introduction*. Springer.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Donohue III, J. J. and Levitt, S. D. (2001). The impact of legalized abortion on crime. *Quarterly Journal of Economics*, 116(2):379–420.

Eicher, T. S., Helfman, L., and Lenkoski, A. (2012). Robust fdi determinants: Bayesian model averaging in the presence of selection bias. *Journal of Macroeconomics*, 34(3):637–651.

Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Simmons, S. J., Fang, F., Fang, Q., and Ricanek, K. (2010). Markov chain Monte Carlo model composition search strategy for quantitative trait loci in a Bayesian hierarchical model. *World Academy of Science, Engineering and Technology*, 63:58–61.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Universidad EAFIT (2009). Reglamento de propiedad intelectual.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of mathematical psychology*, 44(1):92–107.