Fighting Multicollinearity in Double Selection: A Bayesian Approach

Mateo Graciano Londoño*
Andrés Ramírez Hassan†

EAFIT University
Medellín, Colombia

June 9, 2016

Abstract

We propose a model selection procedure when facing high multicollinearity levels applied to the inference over a treatment effect. We show different Frequentist and Bayesian approaches applied to a model selection procedure based on a post double estimation procedure. Our simulation results have evidence in favor of Bayesian procedures when the number of observations is not much higher than the number of possible controls. Finally, we perform a post double MC^3 procedure on real data regarding the impact of legalized abort on crimes rates.

Keywords: abortion, crime rates, model composition, treatment effect, post double selection.

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

1 Introduction

How can be explained the relationship between two specific variables? That is a question which many researchers have in a daily basis. For instance, one might be interested in some government policy and its effect on an important economic measure such as the gross domestic product, that would be important because no government would want to spend money in a policy which is leading to an undesirable result or maybe to nothing at all.

For instance, on the early 90's the people of United States of America wanted an explanation on the drop of the crime rate on their country, there were different approaches that tried to explain such phenomenon. As we see in Donohue III and Levitt (2001), there were arguments that support that the reason was the increasing use of incarceration, growth of the police force, declines in the crack cocaine trade, the economic growth and increasing precaution on victims but none of those reasons were enough to explain such drop in the crime rate in the whole country.

Donohue III and Levitt (2001) also showed that the legalized abort which leads to higher abortion rates had an important impact on that drop. In that case the authors had an empirical way to select the variables that were used in their work, but given the amount of failed explanation give us the path to which is one of the most important problems in an empirical analysis that is which variables should be included in the model. One may want to have an oracle which tells which are those variables to include, but in reality there is no such marvelous device.

In practice, researchers have a tool which is good but not efficient which is to rely on intuition followed by trial and error, that is why 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, and that is the reason of why model selection have had attention among researchers. 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 Bayesian 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.

As Scott et al. (2010) says, today there is a lot of available information, and that is a reason of why the model selection problem is becoming more relevant, when there is more information and there is no clear guidance on which is relevant information for a given duty, intuition is not the right compass. But since there is the information the results might be better that is why there is a lot of people focused on topics such as big data or data mining. But once there is a structural form defined as a linear model, model selection procedures leads to a better understanding about high-dimensional systems.

Perhaps when talking about model selection there is a big issue that should be taken into account which is model uncertainty. Hansen (2005) said that usual Frequestist methodologies does not include model uncertainty and only consider the fit as the unique measure of comparison between different model, those issues are solved when a Bayesian model selection procedure is performed since they includes model uncertainty.

On this work we followed Belloni et al. (2014) were the main idea is the inference over a treatment which can be taken as exogenous, so our problem framework will be:

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

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

where y_i is the response, β_g and 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 and ζ_i are independent stochastic errors such that

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

Let define n as the number of observations and $p = dim(x_i)$ as the number of potential controls. The whole work will be focused on a good inference over α. In this case x_i is a big set of controls for each i the idea is to select which controls should be included in the analysis.

This paper present a brief introduction to model selection procedures for both, with Frequentist and Bayesian, approaches and the framework for double selection procedures for inference over a treatment effect in section 2, on section 3 are presented simulation results which, apparently, gives evidence to Bayesian procedures to have a better performance in presence of high Multicollinearity levels. On section 4 there are results using real data and a comparison with previous results in the literature, finally on section 5 we conclude and present a summary about the advantages of Bayesian procedures in presence of Multicollinearity.

2 Methodology

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.

We consider model selection procedures based on a common linear model as the following:

$$y = X\beta + \epsilon \tag{3}$$

where X is a set of possible controls, y an exogenous variable and ϵ is a common white noise with mean zero and variance σ^2.

2.1 Frequentist approach

2.1.1 T-test

This is the most common test for check if a variable is significant after a linear regression is done, the statistic in the case in which we are checking if a variable is significant is defined as:

$$T_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{s.e(\hat{\beta}_i)} \sim T_{n-k}$$

where $s.e(\hat{\beta}_i)$ is the standard error of β_i estimation, k is the number of regressors and T_{n-k} is a T-student distribution with $n-k$ degrees of freedom.
The most common model selection procedure is to make a regression with a bunch of possible controls and then discard the one associated to the highest p-value greater than 0.05. That process is performed until there are no insignificant variables in the model.

2.1.2 LASSO

The LASSO (Tibshirani, 1996) estimator is obtained considering an optimization problem as the following for the case of a simple linear model:

$$\beta^* = \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left[y_i - x_i^{'} \beta_m \right]^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{4}$$

where λ is a penalization coefficient. Let T be

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

the post-LASSO estimator is defined as:

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

2.2 Bayesian approach

MC^3 is a Bayesian methodology which uses a stochastic search comparing different models by its posterior model probability. The whole idea came from Raftery et al. (1997) where they performed, on the first part, a not practical procedure because it was needed to had information about the whole possible models which are 2^p. Since today procedures usually allows the researchers to have a big set of data, that would require a lot of computation time. They also knew it so the proposed a Markov chain Monte Carlo approach that directly approximates the exact solution, which leads to the almost the same answer without calculating 2^p different models.

Following Simmons et al. (2010), let $M = \{M_1, M_2, ..., M_m\}$ be the set of models under consideration, and y the observed data as in (3). The posterior model probability (PMP) for model M_j is defined as:

$$P(M_j \mid y, M) = \frac{P(y \mid M_j)\pi(M_j)}{\sum_{i=1}^{m} P(y \mid M_i)\pi(M_i)} \qquad \forall j = 1, 2, ..., m \tag{5}$$

where

$$P(y \mid M_j) = \int ... \int P(y \mid \alpha_j, \;\; M_j)\pi(\alpha_j \mid M_j) d\alpha_j \qquad \forall j = 1, 2, ..., m \tag{6}$$

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, $P(y \mid \alpha_j, \;\; M_j)$ is the likelihood and $\pi(M_j)$ is the prior probability that M_j is the true model.

The a prori acknowledge of the probability of model j of being the true model is the term $\pi(M_j)$ in (5) so it is intuitive to think that is equal to $1/m$ for each of m considered model. But we can see in Scott et al. (2010) that, although that choice is the more intuitive it is not the best, in fact, they use a prior based on a Binomial-Beta distribution, so we have:

$$\pi(M_j) = \pi(M_j \mid prob) = prob^{k_j}(1 - prob)^{p - k_j} \qquad \forall j = 1, 2, ..., m \tag{7}$$

where $prob \sim beta(a, b)$ and k_j is the number of selected variables in model j and that is the prior used for Bayesian procedures in this paper.

We can see in (6) that there are some assumptions over the prior of parameters $\pi(\alpha_j \mid M_j)$. Those assumption can be the usual and most intuitive local prior which is as the presented case 2 in Barbieri and Berger (2004). There are other alternatives such as the non-local priors, as those presented in Johnson and Rossell (2012).

For every model there should be priors for every parameter on it, for the linear regression model those priors include assumptions over σ^2 and β. There are different possibilities for selecting those priors but in general some may use $\sigma^2 \sim Inverse\ gamma(a, b)$ where a and b are hyper-parameters but since there is a difficult regarding the choose of a and b there is also another commonly used prior which is $\sigma^2 \propto \frac{1}{\sigma}$.

The most common (local) prior for β is $\beta \mid M,\ \sigma \sim N_k(0, \sigma^2(gX'X)^{-1})$ which is a k-dimensional normal distribution with mean zero and covariance matrix $\sigma^2(gX'X)^{-1}$. For the case of non-local priors for β we refer to the appendix of this paper.

So far the given methodology leads to the best m models in terms of posterior model probability, but it does not tell which are the variables which leads to the best model. Intuitively one can say that the variables to include would be those which appears in the best model (in terms of PMP), but as Barbieri and Berger (2004) shows, the best model is the *median probability model* in term of prediction.

The *median probability model* is the one which includes every variable which has posterior inclusion probability (PIP) higher than 0.5. The PIP for variable i is defined as

$$PIP_i = \sum_{j=1}^{m} P(M_j \mid y, M) * I_{i,j}$$

where

$$I_{i,j} = \begin{cases} 1 & if \quad x_i \in M_j \\ 0 & if \quad x_i \notin M_j \end{cases}$$

2.3 Double selection procedure

Following the Belloni et al. (2014) idea behind the post double LASSO, we consider a general post double estimation which can be performed regardless the model selection procedure. Consider (1) and (2) a post double selection estimation for α would be a three staged procedure:

1. Let T_1 be a set of selected controls after model selection in (1) excluding d.
2. Let T_2 be a set of selected controls after model selection in (2).
3. Let $T = T_1 \cup T_2$ the set of selected controls in at least one of the previous stages, then make X=T and perform an usual OLS estimation in (1) which leads to a estimation of α.

3 Simulation Results

Considering (1) and (2), we define $dim(x_i) = 40$, $\alpha = 0$, β_g such that there are only eigth non zero coefficients and β_m with only four non zero coefficients.

We also define:
$x_{i1} = N_{10}(0, \Sigma)$
$x_{i2} = N_5(0, I)$
$x_{i3} = x_{i,j} = f_j(x_{i1}, x_{i2}) \qquad \forall j \in \{1, 2, ..., 25\}$
where f_j is a non linear function so that in X_3 there are high order terms of X_1 and X_2 and interactions between them, let define: $x_i = (x_{i1}, x_{i2}, x_{i3})$.

we define three different types of Σ to generate x_i1[1]:
1) Σ so that $\sigma_{ij} \in (0.5, 0.9)$ (defined as type 1)
2) Σ so that $\sigma_{ij} \in (0, 0.5)$ (defined as type 2)
3) $\Sigma = I_{10}$ (defined as type 3)
we also set the signal to noise ratio ($\sigma_{X\beta}/\sigma_\epsilon$) equals to 1, 2 or 5 in both, the structural and the treatment equation. We consider the case where the sample size n is 50, 100 or 500. Finally we define our simulation as:
$y_i = 0.8x_{1,i} + 0.8x_{2,i} + 0.5x_{5,i} - 0.7x_{10,i} + 0.8x_{11,i} + 0.4x_{15,i} - 0.5x_{25,i} + 0.7x_{35,i} + \epsilon_i$
$d_i = 0.6x_{1,i} + 0.8x_{8,i} + 0.9x_{11,i} - 0.5x_{18,i} + \zeta_i$
were both, ϵ and ζ are white noises.

Table 1: *Multinollinearity level*

Measure	Type 1	Type 2	Type 3
	$n = 50$		
VIF	167.34	14.56	9.31
Condition number	318.90	61.03	47.76
	$n = 100$		
VIF	81.50	4.11	2.86
Condition number	152.40	18.56	17.75
	$n = 500$		
VIF	8.23	2.34	1.65
Condition number	21.42	7.61	5.81

The results after the simulations are presented on the following tables[2]:

[1] We consider the VIF of the whole matrix as the mean of the VIF for each column.
[2] Not Rejection (NR) Rate is the rate in which the null hypothesis $\alpha = 0$ is not rejected with significance level of 5%.

Table 2: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_{\epsilon}} = 1$*, type 1*

	MSE	MAE	Range	NR Rate
$n = 50$				
PD T	0.355	0.487	1.667	0.824
PD LASSO	0.376	0.536	1.510	0.746
PD L prior	0.204	0.351	1.755	0.949
PD NL Prior	0.068	0.201	0.991	0.94
PD Oracle	0.204	0.361	1.812	0.947
$n = 100$				
PD T	0.094	0.247	0.808	0.806
PD LASSO	0.093	0.240	0.952	0.867
PD L prior	0.038	0.153	0.762	0.951
PD NL Prior	0.038	0.154	0.764	0.951
PD Oracle	0.037	0.154	0.775	0.951
$n = 500$				
PD T	0.008	0.071	0.355	0.946
PD LASSO	0.008	0.070	0.355	0.949
PD L prior	0.006	0.064	0.327	0.96
PD NL Prior	0.008	0.070	0.354	0.948
PD Oracle	0.008	0.070	0.352	0.948

Table 3: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_{\epsilon}} = 1$*, type 2*

	MSE	MAE	Range	NR Rate
$n = 50$				
PD T	0.081	0.236	0.812	0.796
PD LASSO	0.111	0.301	0.685	0.619
PD L prior	0.041	0.160	0.772	0.941
PD NL Prior	0.070	0.210	1.060	0.946
PD Oracle	0.045	0.168	0.801	0.940
$n = 100$				
PD T	0.028	0.134	0.597	0.917
PD LASSO	0.052	0.182	0.792	0.912
PD L prior	0.022	0.120	0.592	0.951
PD NL Prior	0.023	0.122	0.592	0.952
PD Oracle	0.023	0.120	0.594	0.952
$n = 500$				
PD T	0.004	0.051	0.270	0.966
PD LASSO	0.004	0.051	0.270	0.957
PD L prior	0.006	0.059	0.306	0.955
PD NL Prior	0.004	0.050	0.268	0.963
PD Oracle	0.004	0.050	0.227	0.965

Table 4: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_\epsilon} = 1$*, type 3*

	MSE	MAE	Range	NR Rate
		$n = 50$		
PD T	0.087	0.230	1.085	0.920
PD LASSO	0.047	0.169	0.965	0.971
PD L prior	0.084	0.228	1.103	0.927
PD NL Prior	0.061	0.193	0.928	0.956
PD Oracle	0.081	0.226	1.119	0.948
		$n = 100$		
PD T	0.008	0.072	0.340	0.951
PD LASSO	0.016	0.101	0.431	0.917
PD L prior	0.007	0.068	0.328	0.950
PD NL Prior	0.008	0.070	0.330	0.943
PD Oracle	0.007	0.068	0.328	0.943
		$n = 500$		
PD T	0.003	0.050	0.219	0.949
PD LASSO	0.003	0.045	0.219	0.941
PD L prior	0.003	0.046	0.227	0.947
PD NL Prior	0.003	0.045	0.218	0.946
PD Oracle	0.003	0.045	0.218	0.948

Table 5: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_\epsilon} = 2$*, type 1*

	MSE	MAE	Range	NR Rate
		$n = 50$		
PD T	0.609	0.652	1.762	0.694
PD LASSO	0.515	0.653	1.490	0.613
PD L prior	0.206	0.360	1.750	0.941
PD NL Prior	0.070	0.204	1.062	0.954
PD Oracle	0.223	0.373	1.822	0.943
		$n = 100$		
PD T	0.096	0.240	0.814	0.821
PD LASSO	0.097	0.247	1.121	0.934
PD L prior	0.041	0.160	0.775	0.941
PD NL Prior	0.042	0.161	0.774	0.944
PD Oracle	0.042	0.161	0.779	0.944
		$n = 500$		
PD T	0.008	0.074	0.356	0.953
PD LASSO	0.008	0.075	0.360	0.953
PD L prior	0.006	0.065	0.326	0.952
PD NL Prior	0.008	0.073	0.355	0.954
PD Oracle	0.008	0.072	0.353	0.952

Table 6: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_\epsilon} = 2$*, type 2*

	MSE	MAE	Range	NR Rate
		$n = 50$		
PD T	0.100	0.255	0.930	0.798
PD LASSO	0.147	0.360	0.693	0.416
PD L prior	0.043	0.163	0.786	0.949
PD NL Prior	0.081	0.226	1.119	0.944
PD Oracle	0.040	0.158	0.796	0.952
		$n = 100$		
PD T	0.040	0.161	0.600	0.870
PD LASSO	0.064	0.203	0.975	0.949
PD L prior	0.023	0.121	0.598	0.946
PD NL Prior	0.023	0.120	0.596	0.950
PD Oracle	0.023	0.120	0.591	0.950
		$n = 500$		
PD T	0.004	0.054	0.269	0.953
PD LASSO	0.005	0.055	0.276	0.954
PD L prior	0.006	0.065	0.307	0.934
PD NL Prior	0.004	0.053	0.267	0.958
PD Oracle	0.004	0.053	0.266	0.954

Table 7: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_\epsilon} = 2$*, type 3*

	MSE	MAE	Range	NR Rate
		$n = 50$		
PD T	0.120	0.268	1.195	0.920
PD LASSO	0.050	0.169	0.991	0.967
PD L prior	0.082	0.228	1.172	0.961
PD NL Prior	0.066	0.198	0.966	0.937
PD Oracle	0.079	0.219	1.134	0.950
		$n = 100$		
PD T	0.008	0.073	0.348	0.933
PD LASSO	0.026	0.132	0.519	0.885
PD L prior	0.008	0.070	0.331	0.942
PD NL Prior	0.007	0.070	0.327	0.943
PD Oracle	0.008	0.070	0.330	0.943
		$n = 500$		
PD T	0.003	0.043	0.218	0.961
PD LASSO	0.003	0.044	0.222	0.958
PD L prior	0.003	0.045	0.227	0.958
PD NL Prior	0.003	0.043	0.218	0.953
PD Oracle	0.003	0.043	0.217	0.955

Table 8: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_\epsilon} = 5$*, type 1*

	MSE	MAE	Range	NR Rate
$n = 50$				
PD T	1.310	0.968	2.048	0.523
PD LASSO	0.738	0.715	1.541	0.521
PD L prior	0.231	0.372	1.813	0.945
PD NL Prior	0.147	0.271	1.386	0.951
PD Oracle	0.210	0.363	1.816	0.948
$n = 100$				
PD T	0.070	0.201	0.764	0.869
PD LASSO	0.162	0.322	1.501	0.937
PD L prior	0.042	0.162	0.783	0.939
PD NL Prior	0.043	0.162	0.777	0.931
PD Oracle	0.041	0.159	0.777	0.931
$n = 500$				
PD T	0.008	0.0.071	0.356	0.954
PD LASSO	0.008	0.075	0.372	0.954
PD L prior	0.007	0.066	0.325	0.954
PD NL Prior	0.008	0.070	0.354	0.955
PD Oracle	0.008	0.070	0.353	0.957

Table 9: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_\epsilon} = 5$*, type 2*

	MSE	MAE	Range	NR Rate
$n = 50$				
PD T	0.130	0.288	1.066	0.840
PD LASSO	0.141	0.334	0.791	0.438
PD L prior	0.047	0.171	0.835	0.956
PD NL Prior	0.134	0.264	1.380	0.953
PD Oracle	0.040	0.158	0.801	0.938
$n = 100$				
PD T	0.093	0354	0.607	0.630
PD LASSO	0.119	0.275	1.317	0.947
PD L prior	0.022	0.118	0.602	0.959
PD NL Prior	0.021	0.114	0.594	0.960
PD Oracle	0.021	0.115	0.590	0.960
$n = 500$				
PD T	0.004	0.052	0.269	0.966
PD LASSO	0.005	0.057	0.296	0.953
PD L prior	0.006	0.062	0.305	0.952
PD NL Prior	0.004	0.051	0.267	0.962
PD Oracle	0.004	0.051	0.267	0.962

Table 10: *Performance with* $\frac{\sigma_{X\beta}}{\sigma_{\epsilon}} = 5$*, type 3*

	MSE	MAE	Range	NR Rate
		$n = 50$		
PD T	0.177	0.320	1.335	0.881
PD LASSO	0.169	0.304	1.201	0.815
PD L prior	0.094	0.242	1.206	0.952
PD NL Prior	0.102	0.229	1.157	0.944
PD Oracle	0.085	0.234	1.138	0.952
		$n = 100$		
PD T	0.007	0.070	0.344	0.944
PD LASSO	0.054	0.191	0.679	0.848
PD L prior	0.007	0.068	0.332	0.952
PD NL Prior	0.007	0.067	0.330	0.946
PD Oracle	0.004	0.068	0.230	0.946
		$n = 500$		
PD T	0.003	0.045	0.217	0.953
PD LASSO	0.004	0.047	0.230	0.954
PD L prior	0.003	0.045	0.226	0.950
PD NL Prior	0.003	0.050	0.217	0.949
PD Oracle	0.003	0.050	0.217	0.950

So far those results presented on Tables 2 to 10 show that the most important parameter for a good inference over α is the sample size, in fact, no procedure is very sensible to the signal to noise ratio. The results show that, as expected, they may vary as the level of mulltycolinearity increases. The results show that there are not significant differences between estimation results when $n = 500$.

The inference using a PD LASSO estimation shows that when $n = 50$ (just a bit higher than 40 which is the number of possible controls) the procedure does not perform well, moreover when $n = 100$ PD LASSO performs better but the simulation results show that there is no evidence of obtaining the expected significance level which is 5%.

Our results show that there is evidence in favor of the Bayesian procedures, but there is no significant differences between using local priors and non local priors. The procedures shows that regardless the size n or the signal to noise ratio the significance level is always obtained. Also it is pretty remarkable the fact that both Bayesian procedures perform as well as the post double oracle, which is not plausible, and also is the best methodology when can be performed (just for simulations).

4 Real data results

Donohue III and Levitt (2001) model has the following form:

$$y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit} \tag{8}$$

where i is the index for state, t index of time and $c \in \{violence,\ property,\ murder\}$ is the index of type of crime, ϵ_{cit} the error, δ_{ci} are state-specific effects for time invariant state specific characteristics, γ_{ct} are time specifics effects, w_{it} is a set of control variables and finally a_{cit} is a measure of abortion rate relevant for type of crime c.[3] The set of control variables that where used were the log of lagged prisoners per capita, the log of lagged police per capita, the unemployment rate, per-capita income, the poverty rate, AFDC (Aid to Families with Dependent Children) generosity at time $t-15$, a dummy for concealed weapons law, and beer consumption per capita. Belloni et al. (2014) consider the following model on first differences

$$y_{cit} - y_{ci(t-1)} = \alpha_c(a_{cit} - a_{ci(t-1)}) + z_{cit}^{'}\beta_c + \delta_{ci} + g_{ct} + \eta_{cit} \tag{9}$$

where g_{ct} are time effects and η_{cit} is the error for this case. Both models suggest the same implication of the abortion rate on the crime rate. On this new model they also said that abortion rate should be taken as exogenous conditioned to the data at a given time. That leads to the possibility of an auxiliary equation and then a possible double selection procedure in order to have a better inference on α_c.

They also consider z_{cit} to have a richer set of controls, z_{cit} includes higher order terms and interaction between the originals control variables, they also considered initial conditions of w_{it} (the original set of controls) and a_{cit} and average by states of w_{it}.

First, we make some simulation exercises using the same LASSO procedure as in Belloni et al. (2014), the usual t-statistic and a Bayesian procedure using local priors. The set of controls were the same that were for the murder crimes rates. The results are shown in the Table 11. The simulation was set as:

$$y = 0 * d + \beta Z + \epsilon_y$$
$$d = \delta Z + \epsilon_d$$

where ϵ_y, ϵ_d are independent stochastic errors such that

$$E\left[\epsilon_y \mid Z, d\right] = E\left[\epsilon_d \mid Z\right] = 0$$

Table 11: *Using the same regressors as in Belloni et al. (2014)*

	MSE	MAE	NR Rate
PD T	0.002	0.035	0.884
PD LASSO	0.001	0.030	0.947
PD L prior	0.001	0.034	0.940

Table 11 shows that for that given data set, it is both, PD LASSO and using local priors leads to a better inference on α. Those results shows that using the most common procedure, which is the

[3] This measure is widely explained in Donohue III and Levitt (2001).

t-test, does not give the best results. For this case n and p was 576 and 291 respectively, in this case n was not sufficiently large so there are still differences between procedures.

Considering the double selection procedure on (9) both selection procedures lead to different results as we can see in the Table 12:

Table 12: *Inference on the impact abortion over crime rates*

	Violent crime		Property crime		Murder	
	Effect	$s.e(\hat{\alpha})$	Effect	$s.e(\hat{\alpha})$	Effect	$s.e(\hat{\alpha})$
Donohue III and Levitt (2001)	-0.129	0.024	0.091	0.018	-0.121	0.047
First-difference	-0.152	0.034	-0.108	0.022	-0.204	0.068
Belloni et al. (2014) PD LASSO	-0.104	0.107	0.030	0.055	-0.125	0.151
PD local prior	0.096	0.387	-0.143	0.119	1.059	1.712

On the Table 12 are shown different estimations of the effect of the abort on crime rates, at first, it shows the original idea in which their estimation said that three abort rates were significant, the first difference model shows also the same result and also it shows that abort rates are significant.

Taking into account a more vastly control set makes the model conclusion more desirable since we are finding out which is the real effect of the abortion rate and not giving an explanation of a variable when the true reason is other. After model selection procedures both, the PD LASSO and MC^3, the results shows that the abortion rates are not significant, and therefore implies that there is no real impact of the abortion rate over the crime rates and the true reason were other controls, in other words, it is true that there is evidence in favor of Donohue III and Levitt (2001) statement but, apparently, that happened by indirect reason and the real (direct) reason were hide on the controls proposed by Belloni et al. (2014).

5 Concluding remarks

So far our simulation results confirm that double selection procedures leads to a better estimation of a treatment effect. The real question was what would happen in presence of multicollinearity, Frequentist procedures have problems facing a high multicollinearity level due to problems computing $\left(X'X\right)^{-1}$ (when X is the design matrix) which is needed in a OLS procedure. In the frequentist case there are some procedures to solve that problems using shrinkage estimators such as LASSO or ridge, also there are alternatives in Bayesians approaches which leads to avoid such problem since bayesian estimator are also shrinkage procedures. The latter idea leads to have a preference of a Bayesian procedure over a Frequentist when facing a multicollinearity problem when talking about estimation but that did not implicate that it would be better a Bayesian approach than a Frequentist one in the problem of model selection in presence of multicollinearity, but based on our presented simulations result there is evidence in favor of Bayesian procedures in such case.

A post double estimation procedure have shown better inference on a treatment effect, in fact, we have shown via simulation that inference over the impact of a treatment reaches the theoretical significance level when there is a good model selection regardless the multicollinearity level or the number of samples (n). There is evidence in favor of Bayesian procedures in a model selection context when n is not big enough, when n is sufficiently large there is no evidence of in favor of any model selection procedure, moreover, even the simplest procedure (using t-test) shows the same performance as the LASSO based procedures or both explained Bayesian approaches.

Results using real data for the impact of legalized abort on crime rates show an outcome which could be a little ambiguous, at a first sight there it is not a good result since our results show that there is no relationship between the abort rate and crimes rates which contradicts Donohue III and Levitt (2001) results, but what it says is that there are other factors which affects the abort rate and also the crime rate that were not included in the original model. Since in the original result they did not take into account those factors, their results are given an explicative effect in a variable when the true factor is another one which is the same result that showed Belloni et al. (2014). As in the latter work there is needed to say that this does not directly says that the original result is wrong, but it is a big argument against it since apparently there is a big issue in the model specification which imply problems on estimation and inference based on it.

Further work should consider a model selection procedure using non-local priors in the case when the response variable is not continuous, and also a consider the same methodology applied to a different data base. It is recommended to include others, one-staged, model selection procedures for instance the focused information criteria (FIC) or the Akaike information criterion (AIC).

References

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, pages 870–897.

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.

Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, 21(01):60–68.

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

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Scott, J. G., Berger, J. O., et al. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

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.

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

6 Appendix

Consider the following prior for β as in Johnson and Rossell (2012):

$$\pi(\beta \mid \tau,\, \sigma^2,\, r) = d_p(2\pi)^{-p/2}(\tau\sigma^2)^{-rp-p/2} \mid A_p \mid^{1/2} exp\left\{-\frac{1}{2\tau\sigma^2}\beta' A_p \beta\right\} \prod_{i=1}^{p} \beta_i^{2r} \tag{10}$$

Consider the case were σ^2 is known, the sampling density under model k would be:

$$Y_n \mid \sigma^2,\ \beta_k,\ \ M_k \sim N(X_k\beta_k, \sigma^2 I_n) \tag{11}$$

The marginal distribution would be defined as

$$\int_\beta P(Y \mid \sigma^2,\ \beta\ , M_k)\pi(\beta \mid \tau,\ \sigma^2,\ r)d\beta$$

$$\begin{aligned} &= (2\pi)^{-n/2}(\sigma^2)^{-n/2}d_k(2\pi)^{-k/2}(\sigma^2)^{-rk-k/2}\tau^{-rk-k/2} \mid A_p \mid^{1/2} \\ &\int_\beta exp\left\{-\frac{1}{2\sigma^2}\left(y'y - 2\beta' X_k' y + \beta'(X_k' X_k + \frac{1}{\tau}A_p)\beta\right)\right\} \prod_{i\in K}^{p} \beta_{k_i}^{2r} \end{aligned} \tag{12}$$

We define $C_k = X_k' X_k + \frac{1}{\tau} A_p$ and the integral in (12)

$$\int_\beta exp\left\{-\frac{1}{2\sigma^2}\left(y'y - 2\beta' C_k C_k^{-1} X_k' y + \beta' C_k \beta + y' X_k C_k^{-1} C_k C_k^{-1} X_k' y - y' X_k C_k^{-1} X_k' y\right)\right\} \prod_{i\in K}^{p} \beta_{k_i}^{2r} \tag{13}$$

Define $\mu = C_k^{-1} X_k' y$ and reformulate (13) we obtain:

$$exp\left\{-\frac{1}{2\sigma^2} y' \left(I_n + X_k C_k^{-1} X_k'\right) y\right\} \int_\beta exp\left\{-\frac{1}{2\sigma^2}\left((\beta_k - \mu)' C_k (\beta_k - \mu)\right)\right\} \prod_{i\in K}^{p} \beta_{k_i}^{2r} \tag{14}$$

Completing the integral in (14) to the form of a normal multivariate we obtain:

$$\int_\beta exp\left\{-\frac{1}{2\sigma^2}\left((\beta_k - \mu)' C_k (\beta_k - \mu)\right)\right\} \prod_{i\in K}^{p} \beta_{k_i}^{2r} = (2\pi)^{k/2} \mid C_k \mid^{-1/2} (\sigma^2)^{n/2} E_n\left[\prod_{i\in K}^{p} \beta_{k_i}^{2r}\right] \tag{15}$$

where $E_n[.]$ is the expected value operator for a multivariate normal with mean μ and covariance matrix $\sigma^2 C_k^{-1}$.
Let $R_k = y'\left(I_n + X_k C_k^{-1} X_k'\right) y$ and replace (15) in (14) and then in (12) we obtain the marginal distribution as:

$$\begin{aligned} P(y \mid M_k) =& (2\pi)^{-n/2}(\sigma^2)^{-n/2}d_k(2\pi)^{-k/2}(\sigma^2)^{-rk-k/2}\tau^{-rk-k/2} \mid A_k \mid^{1/2} \\ & exp\left\{-\frac{1}{2\sigma^2}R_k\right\} (2\pi)^{k/2} \mid C_k \mid^{-1/2} (\sigma^2)^{n/2} E_n\left[\prod_{i\in K}^{p} \beta_{k_i}^{2r}\right] \\ =& (2\pi)^{-n/2} d_k (\sigma^2)^{-rk-n/2}\tau^{-rk-k/2}\left[\frac{\mid A_k \mid}{\mid C_k \mid}\right]^{1/2} exp\left\{-\frac{1}{2\sigma^2}R_k\right\} E_n\left[\prod_{i\in K}^{p} \beta_{k_i}^{2r}\right] \end{aligned} \tag{16}$$

Consider the more general case where the variance is not known, a common inverse gamma prior with parameters γ and α, the marginal distribution is defined as:

$$\int_{\beta}\int_{0}^{\infty} P(Y \mid \sigma^2, \beta\ , M_k)\pi(\beta \mid \tau,\ \sigma^2,\ r)\pi(\sigma^2 \mid\ \ \gamma,\ \alpha)d\sigma^2 d\beta$$

$$\begin{aligned} = &\left(d_k(2\pi)^{-n/2}\tau^{-rk-k/2}(2\pi)^{-k/2} \mid A_k \mid^{1/2} \frac{\gamma^{\alpha}}{\Gamma(\alpha)} \int_{\beta} \prod_{i\in K}^{p} \beta_{k_i}^{2r}\right) \cdot \\ &\left(\int_{0}^{\infty} exp\left\{-\frac{1}{2\sigma^2}\left((\beta-\mu)' C_k(\beta-\mu) + R_k + 2\gamma\right)\right\}(\sigma^2)^{-\frac{n+2rk+k+\alpha}{2}-1} d\sigma^2 d\beta\right) \end{aligned} \tag{17}$$

Let $v = n + 2rk + 2\alpha$ and:

$$M = (\beta-\mu)' C_k(\beta-\mu) + R_k + 2\gamma = \left(1 + \frac{v}{v(R_k+2\gamma)}(\beta-\mu)' C_k(\beta-\mu)\right)(R_k+2\gamma)$$

$$\frac{M}{2\sigma^2} = a \quad and \quad d\sigma^2 = \frac{-M}{2}a^{-2}da$$

We can calculate the integral over σ^2 in (17) as follows:

$$\int_{0}^{\infty} e^{-a}\left(\frac{2}{M}\right)^{\frac{k+v}{2}+1} a^{\frac{k+v}{2}+1}\left(\frac{2}{M}\right)^{-1} a^{-2}da = \frac{2^{\frac{k+v}{2}}\Gamma\left(\frac{k+v}{2}\right)}{M^{\frac{k+v}{2}}} \tag{18}$$

Completing the result in (18) to the form of a multivariate t with location vector μ, scale matrix $C_k^* = \frac{R_k+2\gamma}{v}C_k^{-1}$ and v degrees of freedom replacing in (17) we get the marginal distribution as:

$$\begin{aligned} P(y \mid M_k) = &d_k(2\pi)^{-n/2}\tau^{-rk-k/2}(2)^{-k/2} \mid A_k \mid^{1/2} \frac{\gamma^{\alpha}}{\Gamma(\alpha)} \\ &2^{k/2}2^{v/2}\Gamma(v/2)v^{k/2}\left(\frac{R_k+2\gamma}{v}\right)^{k/2} \mid C_k \mid^{-1/2} (R_k+2\gamma)^{-\frac{k+v}{2}} E_t\left[\prod_{i\in K}^{p} \beta_{k_i}^{2r}\right] \\ = &d_k(2\pi)^{-\frac{n}{2}}\tau^{-rk-\frac{k}{2}}2^{v/2}\left[\frac{\mid A_k \mid}{\mid C_k \mid}\right]^{1/2} \frac{\gamma^{\alpha}}{\Gamma(\alpha)}(R_k+2\gamma)^{-v/2}\Gamma(v/2)E_t\left[\prod_{i\in K}^{p} \beta_{k_i}^{2r}\right] \end{aligned} \tag{19}$$