Fighting Multicollinearity in Double Selection: A Bayesian Approach

Research practice 2: Final presentation

Mateo Graciano-Londoño
Mathematical Engineering
Student

Andrés Ramírez-Hassan
Department of Economics
Tutor

Universidad EAFIT, Medellín
Colombia

June 7^{th}, 2016

Intuition on what we want to do

How can be explained an 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.

Common model selection problem

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

$$y = X\beta + \epsilon \tag{1}$$

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

Frequentist: 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.

Frequentist: LASSO

The Lasso estimator as introduced in Tibshirani [1996] is an optimization problem which solves the following:

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

where λ is a penalization coefficient. Let T be

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

Bayesian: MC^3

Markov chain Monte Carlo model composition (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 y the observed data as in (1). The posterior model probability (PMP) for model M_j is defined as:

$$P(M_j \mid y, M) = \frac{P(y \mid M_j)\pi(M_j)}{\sum_{i=1}^{m} P(y \mid M_i)\pi(M_i)} \quad \forall j = 1, 2, ..., m \qquad (3)$$

Bayesian: MC^3

Let

$$P(y \mid M_j) = \int \ldots \int P(y \mid \alpha_j, \ M_j)\pi(\alpha_j \mid M_j)d\alpha_j \quad \forall j = 1, 2, ..., m \tag{4}$$

be 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.

Bayesian: MC^3, defining priors

The a prori acknowledge of the probability of model j of being the true model is the term $\pi(M_j)$ in (3) 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::

$$\pi(M_j) = \pi(M_j \mid prob) = prob^{k_j}(1 - prob)^{p-k_j} \quad \forall j = 1, 2, ..., m \quad (5)$$

where $prob \sim beta(a, b)$ and k_j is the number of selected variables in model j.

Bayesian: MC^3, defining priors

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 $\sigma^2 \sim \mathit{InverseGamma}(a, b)$ 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 $\sigma^2 \propto \frac{1}{\sigma}$.

Bayesian: MC^3, defining priors

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-variate normal distribution with mean 0 and covariance matrix $\sigma^2(gX'X)^{-1}$.
The idea of a nonlocal (to 0) prior is to effectively eliminate models with unnecessary explanatory variables, for instance consider the following nonlocal prior proposed by Johnson and Rossell [2012]:

$$\pi(\beta \mid \tau,\ \sigma^2,\ r,\ A_p) = d_p(2\pi)^{-p/2}(\tau\sigma^2)^{-rp-p/2} \mid A_p \mid^{1/2} exp\left\{-\frac{1}{2\tau\sigma^2}\beta' A_p \beta\right\} \prod_{i=1}^{p} \beta_i^{2r} \quad (6)$$

where τ, r, A_p are hyper-parameters for the prior.

Bayesian: MC^3, choosing which variables to include

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.

Bayesian: MC^3, choosing which variables to include

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

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

Double selection: Problem statement

Consider the following structure [Belloni et al., 2014]:

$$y_i = \alpha d_i + x_i^{'} \beta_g + \epsilon_i \tag{7}$$

$$d_i = x_i^{'} \beta_m + \zeta_i \tag{8}$$

where y_i is the response, β_g, β_m are 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

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

Double selection: How to do it

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 (7) and (8) 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 7 excluding d.
2. Let T_2 be a set of selected controls after model selection in 8.
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 (7) which leads to a estimation of α.

General objective

Propose a double post MC3 estimators based on local and non local prior distributions, and compare its performance with the frequentist counterpart under different multicollinearity degrees. ✓

Specific objectives

- Implement the post double selection and MC^3 on simulations exercises. ✓
- Gather real information as in Donohue III and Levitt [2001], and use both methodologies. ✓
- Compare both methodologies and analyse how they perform based on simulation and real cases. ✓

Simulation settings

Considering (7) and (8) 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:
$x_i = (x_{i1}, x_{i2}, x_{i3})$

Simulation settings

We define three different types of Σ to generate $x_i 1$

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 consider the case where the sample size n is 50, 100 or 500.

Simulation settings

Finally we define our simulation as:

$$\begin{aligned} y_i =& 0.8x_{1,i} + 0.8x_{2,i} + 0.5x_{5,i} - 0.7x_{10,i} \\ &+ 0.8x_{11,i} + 0.4x_{15,i} - 0.5x_{25,i} + 0.7x_{35,i} + \epsilon_i \end{aligned} \tag{9}$$

$$d_i = 0.6x_{1,i} + 0.8x_{8,i} + 0.9x_{11,i} - 0.5x_{18,i} + \zeta_i \tag{10}$$

were both, ϵ and ζ are white noises

Multicollinearity levels

Table: 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

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

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

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

Summary

So far those results show that the most important parameters 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 $n = 500$.

Model formulation

Donohue III and Levitt [2001] model has the following form:

$$y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit} \tag{11}$$

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

Which were those controls w_{it}?

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.

Another aproach

Belloni et al. [2014] consider the following model on first differences

$$y_{cit} - y_{ci(t-1)} = \alpha_c(a_{cit} - a_{ci(t-1)}) + z_{cit}^{'}\beta_c + \delta_{ci} + g_{ct} + \eta_{cit} \quad (12)$$

where g_{ct} are time effects and η_{cit} is the error for this case. 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}.

PD Selection?

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.

Comparing results

Table: 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

What happened?

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].

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.

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

References

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.

Any questions?