An alternative for variable selection in high-dimensional linear regression models

Una alternativa para selección de variables en modelos de regresión lineal de alta dimensión.

Héctor M. Núñez [1]
Jesús Otero [2]

[1]Centro de Investigación y Docencia Económicas (CIDE), Mexico

[2]Facultad de Economía, Universidad del Rosario, Colombia

October, 2021

Introduction

- Chudik, Kapetanios, and Pesaran (2018), henceforth CKP, propose a one covariate at a time multiple testing (OCMT) procedure as an alternative approach to penalised regression for variable selection in high-dimensional linear regression models.
- The objective of OCMT is to find a set of predictors that is sufficient to approximate the true data generating process (DGP) that characterises the variable of interest.
- CKP motivate OCMT with several main advantages over penalised regression methods, including ease of interpretation, relation to classical statistical analysis, computational speed, and good performance in small samples.

Introduction

- To illustrate the usefulness of their novel approach, CKP use OCMT in a macroeconomic exercise in which a small subset of predictors is selected from a larger set of possibly relevant variables to forecast U.S. output growth and inflation.
- In this paper, we embark on a replication exercise of key OCMT results reported by CKP in their Monte Carlo study and empirical illustration sections.
- Our results are based on two new user-written commands using the Stata software, instead of the original MATLAB routines provided by CKP in their documentation files.

An overview of OCMT

- To illustrate the approach, suppose that a researcher is interested in explaining a target variable, y_t, in terms of n independent variables which correspond to what the authors call the "active" set $\mathcal{S}_{nt} = \{x_{i,t}, i = 1, 2, \ldots, n\}$, where the number of variables n can potentially be larger than the number of observations T.
- The variables in $\mathcal{S}_{nt}$ are classified in three categories:
  - k signals which collectively generate y_t;
  - $k^\star$ pseudo-signals which are correlated with signal variables but are not included in the model that generates y_t;
  - $n - k - k^\star$ noise variables which are not correlated with signals.
- The researcher does not know the identity of the signal variables though.

An overview of OCMT

- The DGP is:

$$y_t = \boldsymbol{a}' \boldsymbol{z_t} + \sum_{i=1}^{k} \beta_i x_{i,t} + u_t, \tag{1}$$

- where y_t is the variable of interest, $\boldsymbol{z_t}$ is a vector of known preselected variables, $x_{1,t}, \ldots, x_{k,t}$ is the set of k unknown signal variables, $0 < |\beta_i| \leq C < \infty$, with C a positive finite constant, for $i = 1, 2, \ldots, k$, u_t is the error term, and $t = 1, 2, \ldots, T$ is the number of observations available for estimation. Both $\boldsymbol{z_t}$ and $x_{i,t}$, where $i = 1, 2, \ldots, k$, are assumed to be uncorrelated with u_t, at time t.
- The vector of preselected variables $\boldsymbol{z_t}$ may include deterministic variables, e.g. intercept, trends and indicator variables; stochastic variables, e.g. past values of y_t and common factors; and other variables that are viewed as relevant based on theoretical considerations.

An overview of OCMT

- The OCMT approach to variable selection starts off by assessing the individual statistical significance of the variables in the active set through the estimation of ordinary least squares (OLS) regressions of y_t on $\boldsymbol{z_t}$ and $x_{i,t}$ one at a time, where $i = 1, 2, \ldots, n$.
- The variables whose t-statistics are greater (in absolute terms) than a given threshold are selected at the end of this first step. CKP define the threshold using the so-called critical value function:

$$c_p(n, \delta) = \Phi^{-1}\left(1 - \frac{p}{2f(n, \delta)}\right), \tag{2}$$

- where $\Phi(\cdot)$ denotes the standard normal distribution function, $f(n, \delta) = cn^{\delta}$ for some positive constants c and δ, where δ is referred to as the critical value exponent, $0 < p < 1$ is the nominal size of the individual tests statistics, and n is the number of variables in the active set.

An overview of OCMT

- All the variables that satisfy the condition already stated are next selected to conform an initial model specification for the k true or signal variables.
- In a second step, OCMT uses this initial specification and once more tests on an individual basis the statistical significance of the variables in the active set that were not selected in the previous step.
- OCMT continues on a multi-step fashion until no variables in the active set are found to be statistically significant.
- When this occurs the procedure finalises with OCMT yielding a linear regression model in which all the variables selected in the previous steps are included as joint determinants of y_t.

An overview of OCMT

- An important point here is that in the second and subsequent steps the critical value function is given by:

$$c_p(n, \delta^\star) = \Phi^{-1}\left(1 - \frac{p}{2f(n, \delta^\star)}\right), \tag{3}$$

- where it is required that $\delta^\star > \delta$. CKP indicate that the number of steps required for OCMT to converge is finite and bounded by the number of variables in the active set, that is n.
- The positive constants δ and $\delta^\star$ may be viewed as fine-tuning parameters that permit to adjust the critical values required for inference.

Replication of the empirical illustration section

- To this end, we develop the user-written Stata commands `baing` and `ocmt` which can be installed, along with their associated help files, by typing `ssc install baing` and `ssc install ocmt` in the Stata command window, respectively.
- The first command follows Bai and Ng (2002), who offer a data-dependent procedure to consistently estimate the number of common factors in large dimensional panels ($N, T \to \infty$) based on information criteria (IC).
- The reason for applying Bai-Ng follows the premise that when forecasting using a very large number of predictors, the information contained in them can be adequately summarised by a reduced number of estimated factors (see Stock and Watson, 2002, p.147).
- See Stata routines

```
h baing
h ocmt
```

Replication of the empirical illustration section

- The source of the data used by CKP is Stock and Watson (2002)
- After suitably merging the four data files, we end up with $74 + 1 + 35 + 1 = 111$ quarterly variables observed over the sample period 1960q1-2008q4, for a total of 196 observations.
- The variables output growth and inflation are referred to as target variables.
- In turn, the remaining 109 variables and the first four lags of the corresponding target variable conform the so-called active set, for a total of 113 variables.

Replication of the empirical illustration section

- The forecasting exercise is of the rolling type, using 74 sub-samples each of which consisting of 120 quarters; that is, 1960q3-1990q2, 1960q4-1990q3, and so on until reaching 1979q1-2008q4, so the effective sample size is 116 quarters in each sub-sample.
- In each rolling sample the common factors are determined using the target variable and those in the active set (excluding the lagged values of the target variable), based on the IC1 criterion of Bai and Ng (2002) with the maximum number of factors set equal to 5.
- For both output growth and inflation the command `baing` exactly reproduces the same number of common factors in all sub-samples; in the overwhelming majority of cases the number of factors is three (in few cases, a fourth factor is found).

Replication of the empirical illustration section

- Following CKP, we then apply OCMT using $c = 1$, $p = 0.05$, $\delta = 1$ and $\delta^{\star} = 2$, using regressions of each target variable conditional on the corresponding number of identified lagged common factors (which are always included in the set of preselected variables $\boldsymbol{z_t}$, along with the intercept term), and the variables in the active set $\mathcal{S}_{nt}$ lagged one period (let us recall that the active set also includes the first four lags of the target variable).

Replication of the empirical illustration section

Top five variables selected	CKP ($\delta = 1$)	This paper ($\delta = 1$)					
	$\delta^\star = 2$	$\delta^\star = 2$	$\delta^\star = 1.5$	$\delta^\star = 1.01$	$\delta^\star = 2$	$\delta^\star = 1.5$	$\delta^\star = 1.01$
Bai and Ng (2002) Information Criterion	IC(1)	IC(1)	IC(1)	IC(1)	IC(2)	IC(2)	IC(2)
Output growth:							
Residential price index	47.3	47.3	47.3	47.3	50.0	50.0	50.0
First lag of the dependent variable	45.9	45.9	45.9	45.9	32.4	32.4	32.4
Industrial production index - fuels	43.2	43.2	43.2	43.2	44.6	44.6	44.6
Labour productivity (output per hour)	37.8	37.8	37.8	37.8	27.0	27.0	27.0
Employees, nonfarm - mining	27.0	27.0	27.0	27.0	32.4	32.4	32.4
Average number of selected variables (excluding preselected factors)	2.2	2.2	2.2	2.2	2.1	2.1	2.2
Inflation:							
First lag of the dependent variable	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Third lag of the dependent variable	78.4	78.4	78.4	78.4	85.1	85.1	85.1
MZM money supply (FRB St. Louis)	71.6	71.6	71.6	71.6	43.2	43.2	43.2
Money Stock: M2	45.9	45.9	45.9	45.9		41.9	41.9
Recreation price index	33.8	33.8			41.9		
Second lag of the dependent variable			52.7	90.5	39.2	71.6	87.8
Average number of selected variables (excluding preselected factors)	4.0	4.0	4.5	4.9	4.1	4.4	4.5

References I

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Chudik, A., G. Kapetanios, and M. H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica 86*(4), 1479–1512.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147–162.