Economics Bulletin

Volume 36, Issue 1

Regularization parameter selection via cross-validation in the presence of dependent regressors: a simulation study

Yoshimasa Uematsu
The Institute of Statistical Mathematics

Shinya Tanaka
Otaru University of Commerce

Abstract

This letter reveals using simulation studies that regularization parameter selection via cross-validation (CV) in penalized regressions (e.g., Lasso) is valid even if the regressors are weakly dependent. In CV procedure, the time series structure of the data set is broken, meaning that there may occur a fatal problem unless the sample is i.i.d.; the estimation accuracy in the training step could be worse due to corruption of data continuity, which may furthermore lead to a bad choice of the regularization parameter. Even in such a situation, we find that CV works well as long as the sample size grows. These findings encourage us to apply the selection procedure via CV to macroeconomic empirical analyses with dependent regressors.

Uematsu acknowledges financial supports from a Grant-in-Aid for JSPS Fellows, 26-1905. Tanaka acknowledges financial support from Joint Usage and Research Center, Institute of Economic Research, Hitotsubashi University.
Citation: Yoshimasa Uematsu and Shinya Tanaka, (2016) "Regularization parameter selection via cross-validation in the presence of dependent regressors: a simulation study", *Economics Bulletin*, Volume 36, Issue 1, pages 313-319
Contact: Yoshimasa Uematsu - uematsu@ism.ac.jp, Shinya Tanaka - stanaka@res.otaru-uc.ac.jp.
Submitted: January 12, 2016. **Published:** March 17, 2016.

1 Introduction

In recent macroeconometrics and forecasting works, the estimation of a large number of parameters by a penalized regression, such as Lasso by Tibshirani (1996), has attracted much attention (see Fan et al. (2011) and Uematsu and Tanaka (2015), for example). As is well-known, the choice of a *regularization parameter*, λ, is essential in penalized regression, since the value of λ determines the sparsity of the estimated vector of coefficients; a larger (smaller) value of λ leads to more sparse (not sparse) estimates. One of the methodologies commonly used for choosing λ is the *k-fold cross-validation* (CV) (see Breheny and Huang (2011), for example). This procedure is as follows. The sample is partitioned into k subsamples. A subsample is retained as the *validation data* for testing the model, and the remaining $k-1$ subsamples are used as the *training data*. The CV procedure is repeated k times, with each of the subsamples used exactly once as the validation data. In this process, the time series structure of the data set is broken, meaning that there may occur a fatal problem unless the sample is i.i.d.; the estimation accuracy in the training step could be worse due to corruption of data continuity, which may furthermore lead to a bad choice of λ. See Figure 1 that illustrates the case $k=5$. To the best of our knowledge, there are few studies on CV for penalized regressions with dependent data. This is the motivation of this letter; we show using simulation studies that regularization parameter selection via CV in penalized regressions is valid even if the regressors are weakly dependent time series.

Let us consider the estimation of an h-step ahead forecasting regression model, $y_{t+h} = \alpha + \beta' x_t + u_t$, by penalized regression. The estimator of the regression coefficients, $(\hat{\alpha}, \hat{\beta}')'$, is then defined as a solution to the minimization problem of objective function $Q(\alpha, \beta)$, where

$$Q_T(\alpha, \beta) = \sum_{t=1}^{T} \left(y_{t+h} - \alpha - \beta' x_t\right)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{1}$$

Here, $p_\lambda(v)$ for $v \in [0, \infty)$ is a penalty function indexed by the *pre*-determined regularization parameter $\lambda (= \lambda_T) > 0$. The penalty function p_λ can be the L_1-penalty (Lasso) by Tibshirani (1996), the smoothly clipped absolute deviation (SCAD) penalty by Fan and Li (2001), or the minimax concave penalty (MCP) by Zhang (2010). From a theoretical point of view, $\lambda \to 0$ must be satisfied. Further, this convergence rate must appropriately be endowed to achieve the two statistically desirable properties, *oracle inequality* for prediction and *oracle property* for the estimator; see Fan et al. (2011) and Uematsu and Tanaka (2015). However, it is not necessarily clear what actual value should be assigned in practice. We thus check the validity of the choice of λ with CV using Monte Carlo simulation, in terms of these two theoretical viewpoints.

The remainder of this letter is organized as follows. Section 2 gives the model and its parameter configuration. Section 3 presents the result of Monte Carlo simulation. Section 4 concludes.

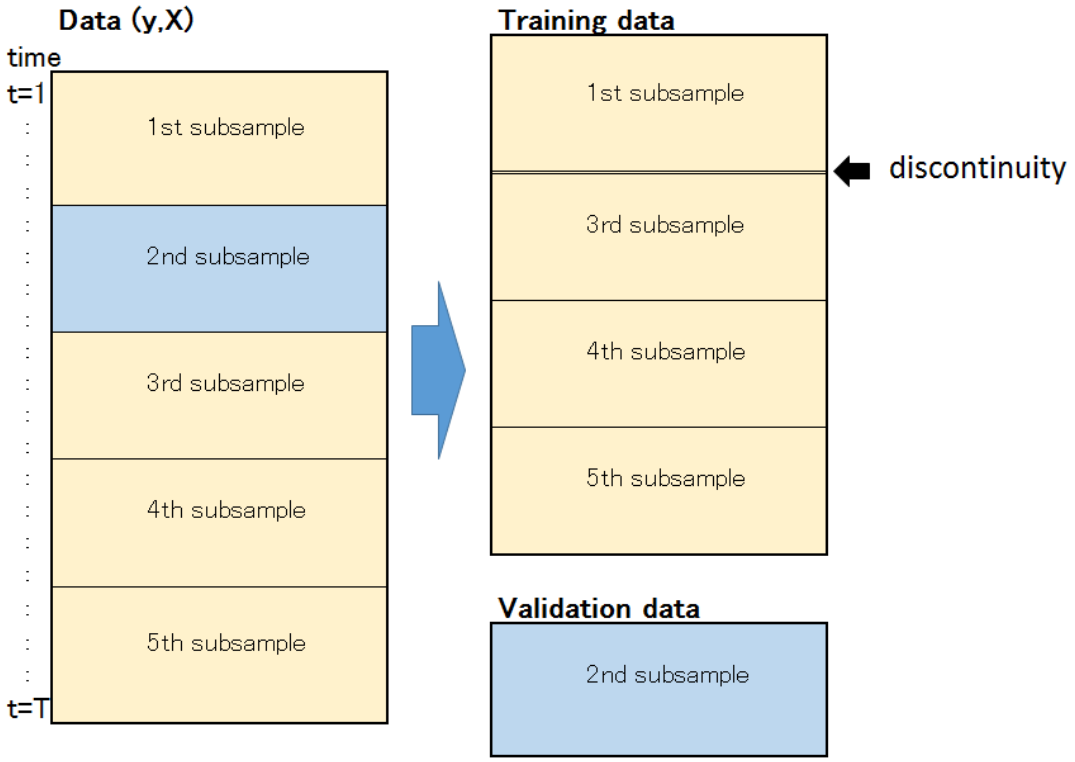

Figure 1: When jth subsample is taken as a validation data, the model is estimated by using the sample made of the other remaining subsamples. The estimation could be suffered due to corruption of time series structure unless $j = 1$ or 5. Even when $j = 1$, moreover, the estimation is made by using the "future observations" from the validation subsample.

2 Preliminary

2.1 Model and estimator

We investigate the validity of using CV in terms of h-step ahead forecasting and estimation of the coefficients of the model with dependent regressors:

$$y_{t+h} = \alpha + \beta' x_t + u_t, \tag{2}$$

$$x_t = \mu + \Phi x_{t-1} + v_t, \tag{3}$$

where $(\alpha, \beta')' \in \mathbb{R} \times \mathbb{R}^p$ is a parameter vector of interest, $x_t \in \mathbb{R}^p$ is a pre-specified stationary VAR(1) regressor with variance Σ_x, and $(u_t, v_t) \in \mathbb{R} \times \mathbb{R}^p$ is an i.i.d. Gaussian error process with mean zero and variance $\Sigma_{u,v}$. We consider a large dimensional sparse parameter vector, β, which is assumed to be split into two subvectors $\beta_A \in \mathbb{R}^q$ and $\beta_B \in \mathbb{R}^{p-q}$; β_A consists of q nonzero entries, but $\beta_B = 0$. Note that $p \geq T$ is allowed while $q < T$ must be satisfied. Regarding the estimation of the coefficient vector of (2), we adopt a penalized least squares to get the sparse estimate of β. Namely, we find the estimate $(\hat{\alpha}, \hat{\beta}')'$ by minimizing objective function $Q_T(\alpha, \beta)$ in (1) with p_λ being either Lasso, SCAD, or MCP.

Here we briefly introduce these penalty functions for the sake of completeness. Let θ denote a positive variable. The L_1-penalty (Lasso) is given by $p_\lambda(\theta) = \lambda\theta$, and we then obtain $p'_\lambda(\theta) = \lambda$ and $p''_\lambda(\theta) = 0$. The SCAD penalty is defined by

$$p_\lambda(\theta) = \begin{cases} \lambda\theta & \text{if} \quad \theta \le \lambda \\ \dfrac{\gamma\lambda\theta - 0.5(v^2+\lambda^2)}{\gamma-1} & \text{if} \quad \lambda < \theta \le \gamma\lambda \\ \dfrac{\lambda^2(\gamma^2-1)}{2(\gamma-1)} & \text{if} \quad \theta > \gamma\lambda \end{cases}$$

Its derivative is

$$p'_\lambda(\theta) = \lambda\left\{1(\theta \le \lambda) + \frac{(\gamma\lambda-\theta)_+}{(\gamma-1)\lambda}1(\theta > \lambda)\right\}$$

for some $\gamma > 2$. Then we have $p''_\lambda(\theta) = -(\gamma-1)^{-1}1\{\theta \in (\lambda, \gamma\lambda)\}$. The MCP is defined by

$$p_\lambda(\theta) = \begin{cases} \lambda\theta - \dfrac{\theta^2}{2\gamma} & \text{if} \quad \theta \le \gamma\lambda \\ \dfrac{1}{2}\gamma\lambda^2 & \text{if} \quad \theta > \gamma\lambda \end{cases}$$

Its derivative is $p'_\lambda(\theta) = \gamma^{-1}(\gamma\lambda - \theta)_+$ for some $\gamma \ge 1$. Thus, we have $p''_\lambda(\theta) = -\gamma^{-1}1\{\theta < \gamma\lambda\}$.

2.2 Parameter configuration

In our simulation study, we set $\mu = 0$ for simplicity and assume Φ to be diagonal to clarify the magnitude of dependence in x_t. Moreover, we denote by σ_u^2 and $\sigma_v^2 I_p$ the variance of u_t and the covariance matrix of v_t, respectively. The parameters and error terms are specified as follows:

$$\alpha = ca, \qquad \beta = c\tilde{\iota}, \qquad \Phi = \begin{pmatrix} \phi_A I_q & 0 \\ 0 & \phi_B I_{p-q} \end{pmatrix},$$
$$\Sigma_x = \begin{pmatrix} \sigma_A^2 I_q & 0 \\ 0 & \sigma_B^2 I_{p-q} \end{pmatrix}, \qquad \Sigma_{u,v} = \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 I_p \end{pmatrix},$$

where $\tilde{\iota} = (\iota', 0, \ldots, 0)' \in \mathbb{R}^p$ with $\iota = (1, \ldots, 1)' \in \mathbb{R}^q$, $c^2 = S\sigma_u^2/(\iota'\Sigma_x\iota)$ with S being the signal-to-noise ratio (SNR) of model (2) and $\sigma_j^2 = \sigma_u^2/(1-\phi_j^2)$ for $j = A, B$. Note that the multiplier c is introduced to make the SNR be constant for each model configuration. Specifically, we set the values $a = 0$, $\sigma_u^2 = 1$, $\sigma_v^2 = 0.3$, and $S = 4$. The dimension and the number of included nonzero variables are fixed at $p = 200$ and $q = 10$, respectively. In this situation, we consider the cases $h = 1$, $T = 100, 200$, and $\phi = \phi_A = \phi_B = 0, 0.3, 0.6, 0.9$. The optimization is executed by the coordinate descent algorithm in `ncvreg` of R, proposed by Breheny and Huang (2011). The values of λ are determined by 10-fold CV while we preset the tuning parameter $\gamma = 3.7$ for SCAD and $\gamma = 3$ for MCP. We also consider the cases $h = 2, 3, 4$, $T = 500$, and 5-fold CV in our preliminary simulation. However, these are omitted to save space since the results are essentially the same. These results are available upon request.

3 Results of Simulation Studies

3.1 Results observed from tables

Table 1 gives five criteria to measure the validity of choosing λ by CV. "Mean" and "S.D." are the mean and standard deviation of the selected λ based on 1,000 replications. "MSE" is the mean squared error of h-step ahead forecast $\hat{y}_{t+h}$ computed by

$$\frac{1}{1000}\sum_{r=1}^{1000}\left(y_{T+h}^{(r)}-\hat{y}_{T+h}^{(r)}\right)^2,$$

where $\hat{y}_{T+h}^{(r)}=\hat{\alpha}+\hat{\beta}'x_T^{(r)}$ with $(\hat{\alpha},\hat{\beta}')'$ estimated by the penalized regression of $y_{t+h}^{(r)}$ on $x_t^{(r)}$. "SC-*A*" and "SC-*B*" refer to the sign consistency of $\hat{\beta}_A$ and $\hat{\beta}_B$, describing the success rates $P\left(\mathrm{sgn}(\hat{\beta}_A)=\mathrm{sgn}(\beta_A)\right)$ and $P\left(\mathrm{sgn}(\hat{\beta}_B)=\mathrm{sgn}(\beta_B)\right)$, respectively. Note that $\mathrm{sgn}(\beta_A)\neq 0$ and $\mathrm{sgn}(\beta_B)=0$ by the definitions. Of course, higher values of these rates are desirable. Specifically, SCAD and MCP are expected to have SC-*A* and SC-*B* that approach one as T grows; it is well-known that they have the *oracle property* under regularity conditions with an appropriate choice of λ while Lasso does not have.

First, we consider the case where $T=100$. Overall, we find that Means and S.D.s are essentially the same for each ϕ and penalization method, except for the S.D.s of SCAD and MCP. We also find that MSE worsens in proportion to the value of ϕ. Likewise, SC-*A* is the smallest when $\phi=0.9$. These findings are not surprising because finite sample performance of the estimator with strongly dependent regressors tends to be distorted in general when T is small. However, SC-*B* does not seem to depend on ϕ. Next, we turn to the case where $T=200$ and compare it to the previous case. We first see that S.D.s uniformly decrease. A more interesting fact is that Means decrease for each penalization method; this corresponds to the asymptotic theory that requires $\lambda\to 0$ as $T\to\infty$ to obtain the oracle inequality and oracle property. Thus, the selection of λ by CV seems valid to achieve these desirable properties in a practical sense. Moreover, MSEs drastically improve for every penalization method. All MSEs tend to one regardless of the magnitude of ϕ. On the other hand, the behavior of SCs is different; Lasso fails to distinguish zeros from the nonzero coefficients even if T becomes larger, while the other penalties, SCAD and MCP, succeed in model selection even in highly dependent cases. Again, these findings are consistent with the theory that holds under the appropriate convergence rate of λ, implying that CV selects an appropriate λ in a practical sense.

3.2 Results observed from figures

This subsection focuses on the forecasting accuracy by observing Figures 2–4, which display scatterplots of the MSE (y-axis) against the selected λ (x-axis) for three regularization methods. Each figure consists of eight small pictures; the upper and lower rows show $T=100$ and 200, and from left to right, $\phi=0$, 0.3, 0.6, and 0.9, respectively. If the plots are densely distributed around the line MSE $=1$, it indicates good forecasting accuracy. In addition, a horizontally wide distribution hints at the robustness for λ. The value of λ does

Table 1: Finite sample performance of the penalized regression.

T	100				200			
ϕ	0	0.3	0.6	0.9	0	0.3	0.6	0.9
				Lasso				
Mean	0.105	0.102	0.094	0.092	0.090	0.089	0.082	0.062
S.D.	0.035	0.033	0.031	0.037	0.019	0.019	0.017	0.013
MSE	1.771	1.807	1.860	1.893	1.288	1.262	1.259	1.362
SC-*A*	99.8%	99.7%	98.3%	70.2%	100.0%	100.0%	100.0%	99.2%
SC-*B*	83.0%	83.2%	84.2%	88.7%	84.1%	84.3%	84.2%	85.8%
				SCAD				
Mean	0.133	0.133	0.136	0.158	0.134	0.132	0.128	0.095
S.D.	0.016	0.017	0.032	0.065	0.015	0.015	0.015	0.016
MSE	1.336	1.321	1.598	2.430	1.082	1.082	1.094	1.212
SC-*A*	99.9%	99.7%	96.3%	40.7%	100.0%	100.0%	100.0%	97.5%
SC-*B*	90.9%	91.5%	93.1%	95.0%	95.5%	95.5%	95.9%	95.8%
				MCP				
Mean	0.177	0.176	0.174	0.180	0.165	0.165	0.160	0.117
S.D.	0.024	0.023	0.044	0.069	0.020	0.020	0.020	0.022
MSE	1.326	1.330	1.586	2.595	1.089	1.085	1.099	1.234
SC-*A*	99.6%	99.4%	92.9%	33.3%	100.0%	100.0%	100.0%	95.3%
SC-*B*	96.4%	96.6%	97.0%	96.9%	98.4%	98.4%	98.6%	98.0%

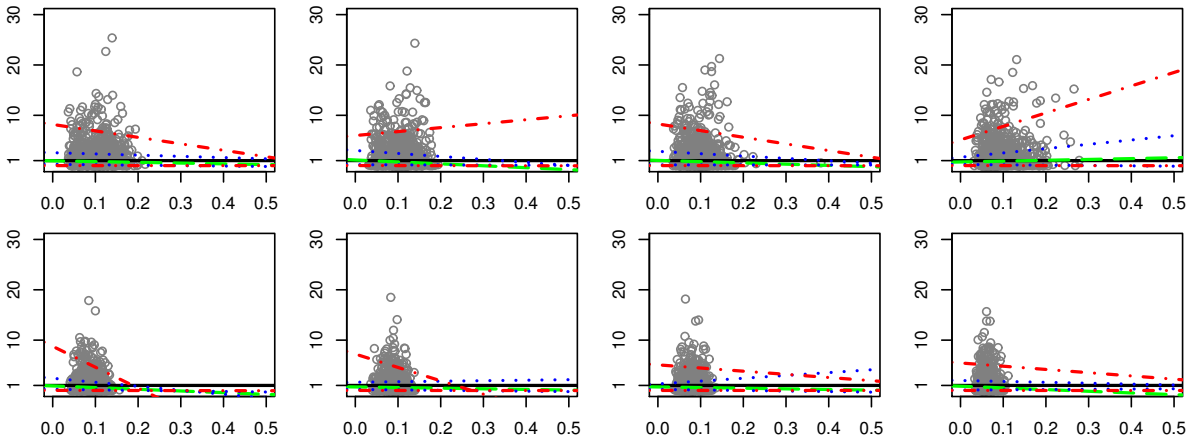

Figure 2: LASSO: Upper $T = 100$ and lower $T = 200$. From left to right, $\phi = 0$, 0.3, 0.6, and 0.9, respectively.

not seem to affect the value of MSE. These features may be well understood by observing the lines; a green dashed line indicates 50%, blue dotted lines show 25% and 75%, and red dash-dot lines denote 5% and 95% regression quantiles, respectively. A black solid line

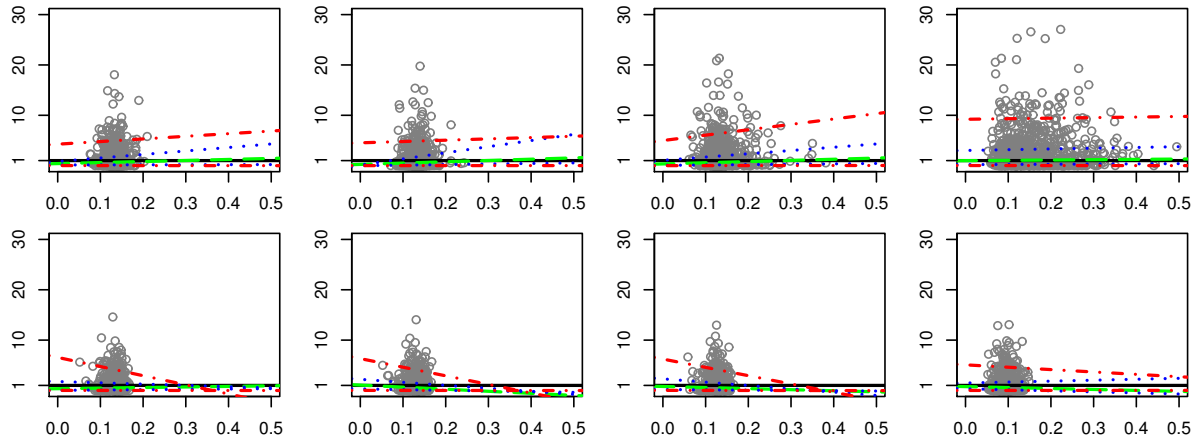

Figure 3: SCAD: Upper $T = 100$ and lower $T = 200$. From left to right, $\phi = 0$, 0.3, 0.6, and 0.9, respectively.

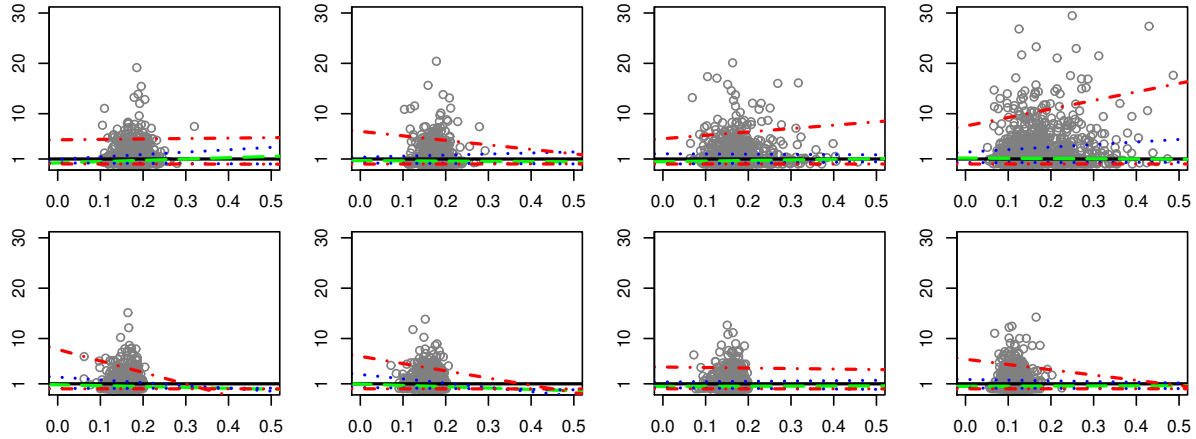

Figure 4: MCP: Upper $T = 100$ and lower $T = 200$. From left to right, $\phi = 0$, 0.3, 0.6, and 0.9, respectively.

designates 1. Note that these lines make sense only on areas where the data exist. Looking at these lines, we can confirm that they lie in lower positions in the case of $T = 200$ as compared to in the case of $T = 100$. This implies that CV works well in choosing λ as long as the sample size is larger even if the model has dependence.

4 Conclusion

In this letter, we have explored how well regularization parameter selection with CV works in penalized regressions with dependent regressors by Monte Carlo study. We have found from our simulation results that the selection using 10-fold CV performs well in terms of MSE and sign consistency as in the independent case. These findings encourage us to apply the selection procedure via CV to macroeconomic empirical analyses where regressors must be dependent.

Acknowledgements

Uematsu acknowledges financial supports from a Grant-in-Aid for JSPS Fellows, 26-1905. Tanaka acknowledges financial support from Joint Usage and Research Center, Institute of Economic Research, Hitotsubashi University.

References

[1] Breheny, P. and J. Huang (2011) "Coordinate descent algorithm for nonconvex penalized regression, with applications to biological feature selection" *Annals of Applied Statistics* **5**, 232–253.

[2] Fan, J. and R. Li (2001) "Variable selection via nonconcave penalized likelihood and its oracle properties" *Journal of the American Statistical Association* **96**, 1348–1360.

[3] Fan, J., J. Lv and L. Qi (2011) "Sparse high-dimensional models in economics" *Annual Review of Economics* **3**, 291–317.

[4] Tibshirani, R. (1996) "Regression shrinkage and selection via the lasso"*Journal of the Royal Statistical Society Series B* **58**, 267–288.

[5] Uematsu, Y. and S. Tanaka (2015) "Macroeconomic forecasting and variable selection with a very large number of predictors: A penalized regression approach" *arXiv*:1508.04217v1.

[6] Zhang, C. H. (2010) "Nearly unbiased variable selection under minimax concave penalty" *Annals of Statistics* **38**, 894–942.