On LASSO for Predictive Regression

Ji Hyung Lee * Zhentao Shi†

March 2018

Abstract

This paper studies possibilities of using shrinkage methods for predictive regression. The variable selection in predictive regression is important since there is a variety of potential predictor variables. The commonly used predictors typically have various degrees of persistence, and exhibit low signal strength in explaining the dependent variable. We investigate the pitfalls and possibilities of the LASSO methods in this framework of predictive regression with mixed degrees of persistence. We show that the adaptive LASSO methods have the consistent variable selection and the oracle properties under the presence of stationary, unit root and cointegrated predictors. The conventional LASSO methods under this environment are also studied, signifying some practical concerns. Exploratory simulation results are reported, and some empirical practices are performed for illustration.

Key words: cointegration, LASSO, nonstationary time series, shrinkage estimation, variable selection

JEL code: C22, C53, C61

*Department of Economics, University of Illinois. Email: jihyung@illinois.edu

†Department of Economics, the Chinese University of Hong Kong. Email: zhentao.shi@cuhk.edu.hk. Shi acknowledges the financial support from the Hong Kong Research Grants Council Early Career Scheme No.24614817.

1 Introduction

Predictive regression models are extensively used in empirical macroeconomics and finance. A leading example is the stock return regression model where predictability has been a long standing goal. The first central econometric issue in these models is severe test size distortion in the presence of highly persistent predictors coupled with the regression endogeneity. When persistence and endogeneity are present, the conventional inferential tools designed for stationary data are no longer valid. Another major challenge in predictive regression is the well-known low signal-to-noise ratio (SNR). The regression coefficient representing the predictive relation is small albeit it could be statistically significant. Thus it is hard to detect the significant relation, and is often dominated by the estimation error when the predictive relation is exploited for forecasting. The predictive regression literature has developed econometric methods for overcoming the inferential difficulties and for improving prediction.

The shrinkage methods have been popular in the era of high dimensional data. We have witnessed unprecedented abundance of data sources across many disciplines such as computer science, neuroscience, engineering and statistics. This data-rich environment also provides new challenges and opportunities in using machine learning technique for economic data analysis. Machine learning methods, in particular, the shrinkage methods are increasingly popular for the econometric inference and prediction in view of its variable selection and regularization property. A leading technique in the shrinkage methods is the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), which has received much attention in the past two decades.

We study the property of LASSO methods in predictive regression. The intrinsic low SNR in predictive regression creates challenges hence naturally calls for a shrinkage method. A researcher may throw in *ex ante* a pool of candidate regressors hoping to catch a few important predictors. The more variables the researcher attempts, the more important is a data-driven method for variable screening, since many of these variables *ex post* demonstrate little to none predictability. LASSO-type shrinkage methods are therefore attractive in the predictive regression as they enable researchers to select the important predictors and excluding the irrelevant or unimportant ones. However, time series regressors in predictive regressions have heterogeneous degrees of persistence. Some may exhibit short memory (e.g., T-bill), while others are highly persistent (e.g., most of financial/macro predictors). Moreover, the multiple persistent predictors can be cointegrated. For example, DP ratio is essentially a cointegrating residual between the dividend and price. The so-called cay data (Lettau and Ludvigson, 2001) is another cointegrating residual between consumption, asset holdings and labor income. The property of LASSO methods under the mixed regressor persistence has not been systematically studied yet.

The performance of LASSO procedure crucially relies on the choice of the tuning parameter.

In this paper, we examine whether a single tuning parameter can cope with the heterogeneous degrees of regressors. In particular, we explore the plain LASSO (Tibshirani, 1996), the standardized LASSO (see below for the definition) and the adaptive LASSO (Zou, 2006) with three categories of regressors; non-cointegrated unit root (I(1)) regressors, cointegrated regressors, and short memory (I(0)) regressors. The different degrees of persistence of the regressors challenges the conventional wisdom of the variable screening property of the plain LASSO and the standardized LASSO. We find that the last two methods with the commonly used tuning parameter cannot deliver proper variable screening. In contrast, a proper choice of the tuning parameter in the adaptive LASSO framework is shown to achieve the oracle property and consistent variable selection. Our exploration in this paper paves a stepping stone toward the automated variable selection in a high-dimensional predictive regression with heterogeneously persistent regressors.

Literature Review Since the seminal LASSO paper by Tibshirani (1996), a variety of nontrivial extension of LASSO has been proposed; the adaptive LASSO (Zou, 2006) and Elastic net (Zou and Hastie, 2008), to name a few. In econometrics, Caner (2009) and Caner and Zhang (2014) employ the LASSO-type procedures in GMM contexts. Belloni and Chernozhukov (2009), Belloni, Chen, Chernozhukov and Hansen (2012), Belloni, Chernozhukov and Hansen (2014), Belloni, Chernozhukov, Chetverikov and Wei (2015) develop new methodology and uniform statistical theory for estimation and inference in various microeconometric settings.

In comparison with the vast literature of LASSO in cross sectional regressions, shrinkage methods are less studied in time series context. Medeiros and Mendes (2016) study the adaptive LASSO method in high-dimensional stationary time series models. Kock and Callot (2015) discuss LASSO in a VAR system. In the time series forecasting context, Inoue and Killian (2006) apply various model selection and model averaging methods to forecast U.S. consumer price inflation. Hirano and Wright (2017) develop a local asymptotic framework with iid orthonormalized predictors to study the risk properties of various machine learning estimators.

There are also a few papers on LASSO with nonstationary data. Caner and Knight (2013) discuss the bridge estimator, a generalization of LASSO, for the augmented Dicky-Fuller test in autoregression. Under the same setting, Kock (2016) investigates consistent variable selection by adaptive LASSO. In a VECM framework, Liao and Phillips (2015) use the adaptive LASSO for cointegration rank selection.

In predictive regression context, Kostakis et al.(2014), Lee (2016) and Phillips and Lee (2013, 2016) provide some valid inference in the presence of multiple predictors with various degrees of persistence. Xu (2017) studies variable selection and inference in predictive regression with possible cointegration among the I(1) predictors. Koo et al. (2016) recently

investigates the use of the plain LASSO in predictive regressions, in which they invoke the restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009). The last two papers are closely related to this paper. We, however, advocate the usage of adaptive LASSO in predictive regression under mixed degrees of persistence.

The paper is organized as follows. Section 2 introduces the unit root regressors into a simple LASSO framework to clarify the idea. Section 3 substantially generalizes the model to include I(0), I(1) and cointegrated regressors. The theoretical results are explored through a set of empirically relevant simulation designs in Section 4. We also examine the stock return regressions via these LASSO methods in Section 5.

Notation We use standard notation. We define $\|\cdot\|_1$ and $\|\cdot\|_2$ as the usual vector l_1-norm and l_2-norm respectively. $\Longrightarrow$, $\longrightarrow^p$ and $\longrightarrow^{a.s.}$ represent convergence in distribution, convergence in probability and almost sure convergence, respectively. All limit theory assumes $n \to \infty$ so we oftentimes omit this condition. $\sim$ signifies "being distributed as" either exactly or asymptotically, depending on the contexts. $O(1)$ and $o(1)$ ($O_p(1)$ and $o_p(1)$) are (stochastically) asymptotically bounded or negligible quantities.

2 LASSO Theory with Unit Roots

In this Section, we study the theory of LASSO with p-dimensional unit root regressors. To fix ideas, we investigate the asymptotic behavior of the adaptive LASSO, plain and standardized LASSO under a simple nonstationary regression model. This simple framework helps us understand the technical issues in LASSO arising from nonstationary predictors with the conventional choices of tuning parameters. Section 3 generalizes the simplistic model to include I(0), I(1) and cointegrated predictors altogether.

Assume the dependent variable y_i is from the linear model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_{jn}^* + u_i,\ i = 1, \ldots, n. \tag{1}$$

Let $y = (y_1, \ldots, y_n)'$ be the response vector, and $X = [x_1, \ldots, x_p]$ be the predictor matrix, where each $x_j = (x_{1j}, ..., x_{nj})'$ for $j = 1, \ldots, p$ are unit root predictors with $x_{ij} = x_{i-1,j} + e_{ij} = \sum_{k=1}^{i} e_{kj}$, $e_{ij} \sim iid\ (0, \sigma_{e_j}^2)$. For simplicity, let $e_{0j} = 0$ for all j. In a $p \times 1$ vector notation

$$x_{i\cdot}' = \sum_{k=1}^{i} e_{k\cdot}'. \tag{2}$$

where $e_{k\cdot} = (e_{k1}, ..., e_{kp})$.

We assume the following iid assumption on innovations. This assumption will be substantially generalized to the linear process assumption in Section 3.

Assumption 2.1 *The vector of innovation* $e_{i\cdot} = (e_{i1}, \ldots, e_{ip})$ *(*$1 \times p$ *vector) and* u_i *(scalar) follow the joint iid distribution:*

$$\begin{pmatrix} e_{i\cdot}' \\ u_i \end{pmatrix}_{(p+1)\times 1} \sim iid \left(0, \Sigma = \begin{pmatrix} \Sigma_{ee} & \Sigma_{eu} \\ \Sigma_{eu}' & \sigma_u^2 \end{pmatrix}\right).$$

Under Assumption 2.1, the following functional central limit theorem holds:

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor nr \rfloor} \begin{pmatrix} e_{k\cdot}' \\ u_k \end{pmatrix}_{(p+1)\times 1} \Longrightarrow \begin{pmatrix} B_x(r) \\ B_u(r) \end{pmatrix} \equiv BM(\Sigma) \tag{3}$$

The regression equation (1) can be equivalently written as

$$y = \sum_{j=1}^{p} x_j \beta_{jn}^* + u = y = X\beta_n^* + u. \tag{4}$$

where $\beta_n^* = \left(\beta_{1n}^*, \ldots, \beta_{pn}^*\right)'$. The true coefficient in (1) $\beta_{jn}^* = \beta_{0j}^*/n^{\delta_j}$, where $\beta_{0j}^* \in \mathbb{R}$ is a fixed constant independent of the sample size, and $\delta_j \in [0, 1)$. Here β_{jn}^* varies with the sample size if $\beta_{0j}^* \neq 0$ and $\delta_j \in (0, 1)$[1].

Note that the pure I(1) regressor model in (4) is a direct extension of the common predictive regression application with a single unit root predictor (e.g., D/P-ratio). The mixed roots case in Section 3 will be more relevant in practical applications with multivariate predictors.

The literature of predictive regression focuses on the non-standard statistical inference caused by persistent regressors and weak signal; the discussion is usually confined to a reasonable number of candidate predictors, but not with a huge number of them. Following this literature, we also consider the asymptotic framework in which p is fixed and the sample size $n \to \infty$[2]. This simple asymptotic framework allows us to concentrate on the contrast between the stationary regressors and the nonstationary ones in the penalized estimation methods. We need not introduce the complex and unstable large sample Gram matrix theory, for which the restriction on the eigenvalues must be imposed (Bickel et al., 2009).

[1] This type of local-to-zero coefficient is designed to balance the I(0)-I(1) relation between the stock return and the unit root predictors, as well as modeling the weak SNR in predictive regressions. See Phillips and Lee (2013) and Timmermann and Zhu (2017) for the recent discussion. Note that the case of $\delta_j = 1$ (Pitman drift) is excluded not to have the effect of nuisance parameter in the limit, unilke Hirano and Wright (2016). Please see Remark (3.6) below for the related discussion and clarification.

[2] Koo et al (2016) allow the number of I(0) regressors to increase while still having the number of I(1) regressors fixed.

Under this framework, one can learn about the unknown coefficients β_n^* from the data by running the OLS

$$\widehat{\beta}^{ols} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2.$$

whose asymptotic behavior is now well understood (Phillips, 1987). Let $\Omega = \int_0^1 B_x(r) B_x(r)' dr$, where B_x is the $p \times 1$ vector Brownian motion, $BM(\Sigma_{ee})$, as given in (3). Let $W = \int_0^1 B_x(r) dB_u(r)$ is a stochastic integral whose distribution depends on Σ. Then we have

$$n\left(\hat{\beta}^{ols} - \beta_n^*\right) = \left(\frac{X'X}{n^2}\right)^{-1} \frac{X'u}{n} \Longrightarrow \Omega^{-1} W.$$

In addition to the low SNR in predictive regressions, some true coefficients β_{0j}^* in (4) could be identically zero. Let $A^* = \{j : \beta_{0j}^* \neq 0\}$ be the set of the relevant regressors and $A^{*c} = \{1, \ldots, p\} \setminus A^*$ be the set of the redundant regressors. Let $p^* = |A^*|$ be the number of relevant regressors. If we have prior knowledge about A^*, ideally we can estimate the unknown parameter by OLS

$$\widehat{\beta}^{oracle} = \arg\min_{\beta \in \mathbb{R}^{p^*}} \|y - \sum_{j \in A^*} x_j \beta_j\|_2^2.$$

We call this the *oracle* estimator. The oracle information about A^* is infeasible in practice. Since $\widehat{\beta}^{oracle}$ is estimated by OLS, it is straightforward to see that its asymptotic distribution is

$$n\left(\hat{\beta}^{oracle} - \beta_n^*\right) \Longrightarrow \Omega_{A^*}^{-1} W_{A^*},$$

where Ω_{A^*} is the $p^* \times p^*$ submatrix $\left(\Omega_{jj'}\right)_{j,j' \in A^*}$ and W_{A^*} is the $p^* \times 1$ subvector $(W_j)_{j \in A^*}$.

2.1 Adaptive LASSO with Unit Root Regressors

The adaptive LASSO is known to enjoy the oracle property in regressions with stationary and weakly dependent regressors (Medeiros and Mendes, 2016). To accommodate the predictive regression applications, we investigate whether the adaptive LASSO maintains the oracle property in the regression with p-dimensional unit root regressors.

The adaptive LASSO estimator for (1) is given by

$$\hat{\beta}^{alasso} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|, \tag{5}$$

where the weight is $\hat{w}_j = |\hat{\beta}_j^{\text{init}}|^{-\gamma}$ for some initial estimator $\widehat{\beta}^{\text{init}}$. In this paper we discuss the case with $\gamma \geq 1$ and $\hat{\beta}^{\text{init}} = \widehat{\beta}^{ols}$.

We introduce additional notation. An index set associated with non-zero coefficients is

called an *active set* in the literature of variable selection. Denote $A_n = \{j : \hat{\beta}_j^{alasso} \neq 0\}$ as the selected active set by the adaptive LASSO (5), while let $A^* = \{j : \beta_{0j}^* \neq 0\}$ be the true active set. For a generic index set A and vector an $p \times 1$ vector β, we denote $\beta_A = (\beta_j)_{j \in A}$. Let $\overline{\delta} = \max_{j \leq p} \delta_j$.

We provide a modified version of Zou (2006, Theorem 2) in the presence of unit root regressors.

Theorem 2.1 *Suppose the linear model (1) satisfies Assumption 2.1. If the tuning parameter λ_n is chosen such that $\lambda_n \to \infty$ and $\frac{\lambda_n}{n^{1-\gamma \cdot \overline{\delta}}} + \frac{1}{\lambda_n n^{\gamma-1}} \to 0$, then*

(a) Variable selection consistency: $P(A_n = A^*) \to 1$.

(b) Asymptotic distribution of $\hat{\beta}_{A^*}^{alasso}$: $n(\hat{\beta}^{alasso} - \beta_n^*)_{A^*} \Longrightarrow \Omega_{A^*}^{-1} W_{A^*}$.

Theorem 2.1 confirms the oracle property of the adaptive LASSO with unit root regressors. The first result indicates that the selected active set coincides with the true active set with probability approaching to one. The second result shows that the asymptotic distribution of adaptive LASSO estimator in the true active set is the same as the oracle estimator.

This delicate adaptive argument is only valid through the proper choice of $\hat{w}_j = |\hat{\beta}_j^{\text{ols}}|^{-\gamma}$ in this nonstationary regression. In essence, when the true coefficients are not zero, $\hat{w}_j$ provides a penalty of the order $\lambda_n n^{\gamma\overline{\delta}-1} = o(1)$ so is negligible, recovering OLS limit theory. On the other hands, if the true coefficients are zero, $\hat{w}_j$ provides a heavier penalty of the order $\lambda_n n^{\gamma-1} \to \infty$ thereby achieving consistent variable selection. This intuition was originally provided in Zou (2006, Remark 2) in the deterministic regressor design.

Remark 2.2 *In Theorem 2.1, we observe some interconnected rate conditions between λ_n, δ_j and γ. To achieve the oracle property in the active set, we need a rate condition of $\frac{\lambda_n}{n^{1-\gamma \cdot \overline{\delta}}} \to 0$. In the meantime, $\lambda_n n^{\gamma-1} \to \infty$ is required to penalize the zero coefficients. Consider the formulation of the usual tuning parameter $\lambda_n = c_\lambda b_n n^{\frac{1}{2}}$, then we need*

$$\frac{b_n}{n^{1/2-\gamma \cdot \overline{\delta}}} + \frac{n^{1/2-\gamma}}{b_n} \to 0.$$

When $\gamma = 1$, and $\overline{\delta} = 1/2$ (a balancing order for I(0)-I(1) regression), the corresponding condition is $b_n + \frac{1}{b_n n^{1/2}} \to 0$ so a slowly shrinking sequence such as $b_n = (\log \log n)^{-1}$ satisfies the rate condition. This is a commonly imposed rate condition in the adaptive LASSO literature.

Since we now have the positive results about adaptive LASSO with unit root regressors, we continue to study the plain LASSO (Tibshirani, 1996), and a simple variant, which we call the standardized LASSO.

2.2 Plain LASSO with Unit Roots

The plain LASSO can be viewed as a special case of the penalized estimation in (5) with the weights $\hat{w}_j$, $j = 1, \ldots, p$, fixed at unity. In this paper, we call it the *plain LASSO* estimator

$$\hat{\beta}^{lasso} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 . \tag{6}$$

LASSO is proposed by Tibshirani (1996) to produce a parsimonious model as it tends to select the relevant variables. The following results characterize the asymptotic behavior of the conventional LASSO according to various choices of λ_n when we use unit root regressors. For exposition, we define a function $D : \left(\mathbb{R}^{\dim(\theta)}\right)^3 \mapsto \mathbb{R}^+$ as

$$D(s, v, \theta) := \sum_{j=1}^{p} s_j \left(v_j \text{sgn}(\theta_j) I(\theta_j \neq 0) + |v_j| I(\theta_j = 0) \right)$$

for three generic vectors s, v, and θ of the same dimension.

Corollary 2.3 *Suppose the linear model (1) satisfies Assumption 2.1.*

(a) If $\lambda_n \to \infty$ and $\lambda_n / n \to 0$, then

$$n(\hat{\beta}^{lasso} - \beta_n^*) \Longrightarrow \Omega^{-1} W$$

(b) If $\lambda_n \to \infty$ and $\lambda_n / n \to c_\lambda \in (0, \infty)$, then

$$n(\hat{\beta}^{lasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega v - 2v'W + c_\lambda D(\mathbf{1}_p, v, \beta_0^*) \right\} .$$

(c) If $\lambda_n / n \to \infty$, and $\lambda_n / n^{2-\bar{\delta}} \to 0$,

$$\frac{n^2}{\lambda_n}(\hat{\beta}^{lasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega v + D(\mathbf{1}_p, v, \beta_0^*) \right\}$$

Remark 2.4 *Corollary 2.3 extends the results of Zou (2006, Section 2) to the unit root regressor case. Following the same discussion as Zou's, we conclude that the plain LASSO's selected variables are in general inconsistent when the unit root regressors are present.*

The above Corollary 2.3 shows that the conventional tuning parameter $\lambda_n \sim \sqrt{n}$ is too small for variable selection with nonstationary regressors. Moreover, without the adaptive argument as in the adaptive LASSO case, the consistent variable selection is not guaranteed.

In this paper, we call the phenomenon that LASSO shrinks some estimated coefficient to exactly zero (whether or not the true coefficients are zeros) as the *variable screening* effect, instead of the *variable selection* effect (which means that LASSO shrinks those truly zero coefficients). Such effect will be further discussed in the paragraphs following Corollary 3.7.

2.3 Standardized LASSO with Unit Roots

In view of the problem that the usual choice of λ_n is too small for LASSO to conduct variable screening in nonstationary regression, one may consider an alternative implementation which is common in practice. LASSO is scale-variant in the sense that if we change the unit of x_j by multiplying it with a non-zero constant c, such a change is not reflected in the penalty term of in (6) so the LASSO estimator does not change proportionally to $\widehat{\beta}_j^{lasso}/c$. To keep LASSO scale-invariant to the choice of unit of x_j, which can be arbitrary, researchers often scale-standardizes LASSO as

$$\hat{\beta}^{slasso} = \arg\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j |\beta_j| . \tag{7}$$

where $\widehat{\sigma}_j = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$ is the sample standard deviation of $(x_{ij})_{i=1}^n$. In this paper, we call (7) the *standardized LASSO*. Such standardization is the default option for LASSO in many statistical packages, for example the `R` package `glmnet`.

We can view the standardized LASSO is another alternative of (5) by setting $\hat{w}_j = \hat{\sigma}_j$. When such a scale standardization is carried out with stationary and weakly dependent regressors, these $\widehat{\sigma}_j$'s converge in probability to the finite population variance. As it does not change the rate of the tuning parameter, it has no asymptotic effect to the estimation. In contrast, when x_j has a unit root, from (3) we have

$$\frac{\widehat{\sigma}_j}{\sqrt{n}} = \sqrt{\frac{1}{n^2}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \Longrightarrow d_j = \sqrt{\int B_{x_j}^2(r)dr - \left(\int B_{x_j}(r)\,dr\right)^2} \tag{8}$$

so that $\widehat{\sigma}_j = O_p(\sqrt{n})$. It thus imposes a much heavier penalty on the associated coefficients with unit root regressors than the stationary ones. Adopting a standard argument for LASSO as in Knight and Fu (2000) and Zou (2006), we have the following asymptotic distribution for $n(\hat{\beta}^{slasso} - \beta_n^*)$. Let $d = (d_1, \ldots, d_p)'$.

Corollary 2.5 *Suppose the liner model (1) satisfies Assumption 2.1.*

(a) If $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to c_\lambda \in [0, \infty)$, then

$$n(\hat{\beta}^{slasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{v'\Omega v - 2v'W + c_\lambda D(d, v, \beta_0^*)\right\},$$

where $c_\lambda = 0$ case restores OLS limit theory.

(b) If $\lambda_n/\sqrt{n} \to \infty$ and $\lambda_n/n^{\frac{3}{2}-\bar{\delta}} \to 0$,

$$\frac{n^{3/2}}{\lambda_n}(\hat{\beta}^{slasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{v'\Omega v + D(d, v, \beta_0^*)\right\}$$

Remark 2.6 *In Corollary 2.5, $D(d, v, \beta_0^*)$ is the term that generates the variable screening effect. In Corollary 2.5(a), $D(d, v, \beta_0^*)$ appears under the usual choice of tuning parameter $\lambda_n \sim \sqrt{n}$. In contrast, its counterpart $D(\mathbf{1}_p, v, \beta_0^*)$ emerges in Corollary 2.3 when $\lambda_n \sim n$. The random vector d, the first argument of $D(\cdot, v, \beta_0^*)$, introduces an extra source of randomness in the variable screening in the standardized LASSO, whereas its counterpart in the plain LASSO is the unit vector $\mathbf{1}_p$. We confirm that the standardized LASSO cannot achieve consistent variable selection in general.*

To summarize, in the regression with unit root predictors, the adaptive LASSO retains the oracle property under the usual choice of the tuning parameter. For the plain LASSO to exhibit the variable screening effect, we need to lift the tuning parameter up to the order of n. For the standardized LASSO, although $\lambda_n \sim \sqrt{n}$ is sufficient for variable screening, the sample variance of the nonstationary regressors brings the random vector d into the limit theory, affecting the variable screening.

The unit root regressors are shown to alter the asymptotic properties of the conventional LASSO methods. In practice, we often encounter a multitude of candidate predictors, exhibiting various dynamic patterns. Some are stationary, while others can be highly persistent and may be cointegrated. In the following section, we will show that the conventional LASSO methods behave even more irregularly under the mixed persistence environment.

3 LASSO Theory with Mixed Roots and Cointegration

In this section, we generalize the model of Section 2 by considering I(0) and I(1) regressors with possible cointegration among those I(1) regressors. In applications of predictive regression with multiple predictors, the model and LASSO theory of this section can provide a general guidance.

3.1 OLS theory with mixed roots

We first study OLS theory since OLS estimator is used as the initial estimator for the adaptive LASSO. The dependent variable y_i is generated from the linear model

$$y_i = \sum_{j=1}^{p_z} z_{ij}\alpha_j^* + \sum_{j=1}^{p_c} x_{ij}^c \phi_{jn}^* + \sum_{j=1}^{p_x} x_{ij}\beta_{jn}^* + u_i = \alpha^{*\prime} Z_i + \phi_n^{*\prime} X_i^c + \beta_n^{*\prime} X_i + u_i, \tag{9}$$

for $i = 1, \ldots, n$, where $Z_i = (z_{i1}, ..., z_{ip_z})'$, $X_i^c = \left(x_{i1}^c, ..., x_{ip_c}^c\right)'$, and $X_i = (x_{i1}, ..., x_{ip_x})'$ represent the stationary, cointegrated and unit root regressors, respectively. Equivalently,

$$y = Z\alpha^* + X^c\phi_n^* + X\beta_n^* + u := W\theta + u, \tag{10}$$

where the response vector $y = (y_1, \ldots, y_n)'$, the observation matrix of predictors

$$\underset{n\times p}{W} = \left[\underset{n\times p_z}{Z}, \underset{n\times p_c}{X^c}, \underset{n\times p_x}{X}\right],$$

and the stacked parameter of $\underset{p\times 1}{\theta} = (\alpha^{*\prime}, \phi_n^{*\prime}, \beta_n^{*\prime})'$ with $p = p_z + p_c + p_x$.

As in Section 2, each $x_j = (x_{1j}, ..., x_{nj})'$ for $j = 1, \ldots, p_x$ is a unit root predictor (initialized at zeros for simplicity) with $x_{ij} = x_{i-1,j} + e_{ij} = \sum_{k=1}^{i} e_{kj}$ where the DGP for $e_{i\cdot}' = (e_{i1}, ..., e_{ip_x})'$ is given in Assumption 3.1 below. In a $p_x \times 1$ vector notation $x_{i\cdot}' = \sum_{k=1}^{i} e_{k\cdot}'$. On the other hand, the $p_c \times 1$ predictor X_i^c has the triangular representation (Phillips, 1991)

$$\begin{aligned} \underset{p_1\times p_c}{A} X_i^c = X_{1i}^c - \underset{p_1\times p_2}{A_1} X_{2i}^c = v_{1i}\ , \\ \triangle X_{2i}^c = v_{2i}\ , \end{aligned} \tag{11}$$

where $A = [I_{p_1}, -A_1]$, $X_i^c = (X_{1i}^{c\prime}, X_{2i}^{c\prime})'$, and $p_c = p_1 + p_2$. Hence p_1 is the cointegration rank, and p_2 is the number of unit roots in the system. This is a convenient but general representation of cointegrated system, and Xu (2017) recently used this structure in predictive regression framework. Now we let $\phi_n^* = \left(\phi_1^*, \ldots, \phi_{p_1}^*, \phi_{p_1+1,n}^*, ..., \phi_{p_1+p_2,n}^*\right)'$ so that, for $j \in \{p_1 + 1, ..., p_1 + p_2\}$, $\phi_{jn}^* = \phi_{0j}^*/n^{\delta_j}$ with $\delta_j \in [0, 1)$ to ensure the regression model validity, similarly to Section 2. Recall $\overline{\delta} = \max_{j\leq p} \delta_j$.

We stack the cointegrating residual vector and the innovation from (11) and define $p_c \times 1$ vector $v_i = (v_{1i}', v_{2i}')'$. We then assume the following linear process for innovation and cointegrating residual vectors. In contrast to the unrealistic iid assumption from Section 2, the linear process assumption is fairly general, including many practical dependent processes (stationary AR and MA processes, for example) as special cases.

Assumption 3.1 *[Linear Process] The vector of stacked innovation follows the linear process:*

$$\underset{(p+1)\times 1}{\xi_i} := \begin{pmatrix} z_{i\cdot}' \\ v_i \\ e_{i\cdot}' \\ u_i \end{pmatrix} = F(L)\varepsilon_i = \sum_{j=0}^{\infty} F_j \varepsilon_{i-j},$$

$$\underset{(p+1)\times 1}{\varepsilon_i} = \begin{pmatrix} \varepsilon_{zi} \\ \varepsilon_{vi} \\ \varepsilon_{ei} \\ \varepsilon_{ui} \end{pmatrix} \sim iid \left(0, \underset{(p+1)\times(p+1)}{\Sigma_\varepsilon} = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zv} & \Sigma_{ze} & 0 \\ \Sigma_{zv}' & \Sigma_{vv} & \Sigma_{ve} & 0 \\ \Sigma_{ze}' & \Sigma_{ve}' & \Sigma_{ee} & \Sigma_{eu} \\ 0' & 0' & \Sigma_{eu}' & \Sigma_{uu} \end{pmatrix} \right),$$

where $F_0 = I_{p+1}$, $\sum_{j=0}^{\infty} j \left\| F_j \right\| < \infty$, $F(z) = \sum_{j=0}^{\infty} F_j z^j$ *and* $F(1) = \sum_{j=0}^{\infty} F_j > 0$.

Remark 3.1 *Following the cointegration and predictive regression literature, we allow the correlation between the regression error* ε_{ui} *and the innovation of nonstationary predictors* ε_{ei}. *However, in order to ensure identification we rule out the correlation between* ε_{ui} *and either the innovation of stationary or the cointegrated predictors.*

Define the long-run covariance matrices associated with the innovation vector as $\Omega = \sum_{h=-\infty}^{\infty} \mathbb{E}\left(\xi_i \xi_{i-h}'\right) = F(1)\Sigma_\varepsilon F(1)'$, where $F(1) = \left(F_z'(1), F_v'(1), F_e'(1), F_u(1)\right)'$. Moreover, define the sum of one-sided autocovariance as $\Lambda = \sum_{h=1}^{\infty} \mathbb{E}\left(\xi_i \xi_{i-h}'\right)$, and $\Delta = \Lambda + \mathbb{E}\left(\xi_i \xi_i'\right)$. We use the functional law (Phillips and Solo, 1992) under Assumption 3.1 to derive

$$\frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor ns \rfloor} \xi_j = \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor ns \rfloor} \begin{pmatrix} z_{i\cdot}' \\ v_i \\ e_{i\cdot}' \\ u_i \end{pmatrix} = \begin{pmatrix} B_{zn}(r) \\ B_{vn}(r) \\ B_{en}(r) \\ B_{un}(r) \end{pmatrix} \Longrightarrow \begin{pmatrix} B_z(r) \\ B_v(r) \\ B_e(r) \\ B_u(r) \end{pmatrix} \equiv BM(\Omega).$$

Note that the observation matrix W can be decomposed as $\underset{n\times p_z}{Z}, \underset{n\times p_1}{X_1^c}, \underset{n\times p_2}{X_2^c}, \underset{n\times p_x}{X}$. From (11), $v_1 = X_1^c - X_2^c A_1'$ is an $n \times p_1$ matrix of I(0) cointegrating residuals. Define

$$R_n = \begin{pmatrix} \sqrt{n} \cdot I_{p_z+p_1} & 0 \\ 0 & n \cdot I_{p_2+p_x} \end{pmatrix},$$

which will serve as a normalizing matrix for any cointegrating rank p_1 with $0 < p_1 < p_c$. We "extend" the I(0) regressors as $Z^+ = [Z, v_1]$ and the I(1) regressors as $X^+ := [X_2^c, X]$. Let us

denote

$$\Omega = \begin{pmatrix} \Omega_{zz} & \Omega_{zv} & \Omega_{ze} & 0 \\ \Omega_{zv}' & \Omega_{vv} & \Omega_{ve} & 0 \\ \Omega_{ze}' & \Omega_{ve}' & \Omega_{ee} & \Omega_{eu} \\ 0' & 0' & \Omega_{eu}' & \Omega_{uu} \end{pmatrix}$$

according to the explicit form of Σ_ε. Then the left-top $p \times p$ submatrix of Ω, which we denote as $[\Omega]_{p\times p}$, can be also represented conformably,

$$[\Omega]_{p\times p} = \begin{pmatrix} \Omega_{zz} & \Omega_{zv} & \Omega_{ze} \\ \Omega_{zv}' & \Omega_{vv} & \Omega_{ve} \\ \Omega_{ze}' & \Omega_{ve}' & \Omega_{ee} \end{pmatrix} = \begin{pmatrix} \Omega_{zz}^{+} & \Omega_{zx}^{+} \\ \Omega_{zx}^{+\prime} & \Omega_{xx}^{+} \end{pmatrix}.$$

Using the BN decomposition and weak convergence to stochastic integral, it is easy to show

$$\begin{pmatrix} Z^{+\prime}u/\sqrt{n} \\ X^{+\prime}u/n \end{pmatrix} \Longrightarrow \begin{pmatrix} \xi_{Z^+} \sim N\left(0, \Sigma_{uu}\Omega_{zz}^{+}\right) \\ \xi_{X^+} \sim \int B^{+}(r)dB_\varepsilon(r)'F_u(1)' + \Delta_{+u} \end{pmatrix} := \xi^{+} \tag{12}$$

where the one-sided long-run covariance matrix $\Delta_{+u} = \sum_{h=0}^{\infty} \mathbb{E}\left(\tilde{u}_i u_{i-h}\right)$ with $\tilde{u}_i = (v_{2i}', e_{i\cdot})'$.

Give the definition of these quantities, we establish the following theorem about the asymptotic distribution of the OLS estimator $\widehat{\theta}_n^{ols} = (W'W)^{-1}W'y$.

Theorem 3.2 *If the linear model (9) satisfies Assumption 3.1, then*

$$R_n\left(\widehat{\theta}_n^{ols} - \theta_n^*\right) \Longrightarrow \left(\Omega^{+}\right)^{-1}\xi^{+}.$$

where $\Omega^{+} = \begin{pmatrix} \Omega_{zz}^{+} & 0 \\ 0 & \Omega_{xx}^{+} \end{pmatrix}$, *and* ξ^{+} *is given in (12) above.*

Remark 3.3 *Theorem 3.2 shows that an asymptotic bias term* Δ_{+u} *appears in the limit distribution of OLS with nonstationary predictors. This asymptotic bias arises from the serial dependence in the innovations. However, the asymptotic bias does not affect the rate of convergence, so* $\hat{\theta}_n^{ols} - \theta_n^* = O_p(R_n^{-1})$. *This rate of convergence is critical in the study of the asymptotic behavior of the adaptive LASSO using* $\hat{\theta}_n^{ols}$ *as the initial estimator.*

Next, we study the asymptotic behavior of the adaptive LASSO in this mixed roots scenario.

3.2 Adaptive LASSO with mixed roots

Similarly to Section 2.1, we define the adaptive LASSO estimator under the system of (10) as

$$\hat{\theta}^{alasso} = \arg\min_{\beta \in \mathbb{R}^p} \|y - W\theta\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\theta_j|, \tag{13}$$

where $\hat{w}_j = |\hat{\theta}_j^{\text{ols}}|^{-\gamma}$. The following theorem confirms that the adaptive LASSO maintains the oracle property and variable selection consistency in the presence of stationary, unit root and cointegrated regressors. Similarly to Section 2, denote $A_n^+ = \{j : \hat{\theta}_j^{alasso} \neq 0\}$ as the selected active set by the adaptive LASSO (13), while let $A^{+*} = \{j : \theta_{0j}^* \neq 0\}$ be the true active set.

Theorem 3.4 *Suppose that the linear model (9) satisfies Assumption 3.1. If the tuning parameter λ_n is chosen such that $\lambda_n \to \infty$ and*

$$\frac{\lambda_n}{n^{(1/2)\wedge\left(1-\gamma\cdot\overline{\delta}\right)}} + \frac{1}{\lambda_n n^{(\gamma-1)/2}} \to 0, \tag{14}$$

then, we have

(a) Variable selection consistency: $P\left(A_n^+ = A^{+*}\right) \to 1$.

(b) Asymptotic distribution of $\hat{\theta}_{A^{+*}}^{alasso}$: $R_n(\hat{\theta}^{alasso} - \theta_n^*)_{A^{+*}} \Longrightarrow \left(\Omega_{A^{+*}}^+\right)^{-1} \xi_{A^{+*}}^+$.

Remark 3.5 *The rate condition for the tuning parameter λ_n in Theorem 3.4 implies the conditions in Theorem 2.1 as a special case, as long as $\gamma \geq 1$, and $\overline{\delta} \geq 1/2$. The condition (14) is reasonable because, (i) we choose $\gamma \geq 1$ in practice to prevent the adaptive LASSO implementation from being non-convex optimization, and (ii) $\overline{\delta} \geq 1/2$ is the balancing order of I(0)-I(1) predictive regression applications. Being agnostic to the presence of stationary, unit root and cointegrated regressors, we can choose the tuning parameter λ_n following the guidance in Theorem 3.4.*

Remark 3.6 *Another related results in the literature are uniformly valid inference and forecasting after the LASSO model selection, see Belloni, Chernozhukov and Kato (2015, 2018) or Hirano and Wright (2016), for example. These papers allow the so-called model selection mistake by LASSO, and provide the valid inference or prediction by introducing local limit theory with small departures from the true models. Combining these recent developments with our current LASSO theory with mixed roots would be interesting future research but we do not pursue here.*

Given what we learn from Caner and Knight (2013) and Kock (2016), the theoretical results in Theorem 3.4 may be expected. These papers, however, work in the pure autoregressive setting with iid error processes. We complement this line of nonstationary LASSO literature by allowing a general regression framework with mixed degrees of persistence. We also generalize the error processes to the commonly used dependent processes, which is important in practice. For example, the long-horizon return regressions in Section 5 requires this type of dependence in their error structure because of the overlapping return construction. Moreover, our research provides a valuable guidance for practice. Faced with a variety of potential predictor variables with uncertain orders of integration, we prefer not to sort them into different persistence categories in predictive regressions.

3.3 Conventional LASSO with mixed roots

We now study the asymptotic theory of the plain LASSO estimator

$$\hat{\theta}^{lasso} = \arg\min_{\theta \in \mathbb{R}^p} \|y - W\theta\|_2^2 + \lambda_n \|\theta\|_1 , \tag{15}$$

under the system of (10). Following the notation in Section 3.1, let

$$\theta_n^* = \left(\alpha^{*\prime}, \phi_n^{*\prime}, \beta_n^{*\prime}\right)' := \left(\alpha^{+*\prime}, \beta_n^{+*\prime}\right)' ,$$

where α^{+*} is the $(p_z + p_1) \times 1$ parameter vector associated with the stationary and cointegrated predictors, and $\beta_n^{+*\prime}$ is the $(p_2 + p_x) \times 1$ local-to-zero parameter vector associated with the unit root predictors.

Corollary 3.7 *Suppose the linear model (9) satisfies Assumption 3.1.*

(a) If $\lambda_n \to \infty$ and $\lambda_n/\sqrt{n} \to 0$, then $R_n(\hat{\theta}^{lasso} - \theta_n^*) \Longrightarrow (\Omega^+)^{-1} \xi^+$.

(b) If $\lambda_n/\sqrt{n} \to c_\lambda \in (0, \infty)$, then

$$R_n(\hat{\theta}^{lasso} - \theta_n^*) \Longrightarrow \left(v_{I(0)}', v_{I(1)}'\right)'$$

where

$$\begin{aligned} v_{I(0)} &\equiv \arg\min_{v \in \mathbb{R}^{p_z + p_1}} \left\{ v'\Omega_{zz}^+ v - 2v'\xi_{Z^+} + c_\lambda \cdot D\left(\mathbf{1}_{p_z+p_1}, v, \alpha^{+*}\right) \right\}, \\ v_{I(1)} &\equiv \left(\Omega_{xx}^+\right)^{-1} \xi_{X^+}. \end{aligned}$$

(c) If $\lambda_n/\sqrt{n} \to \infty$ and $\lambda_n/n \to 0$, then

$$\frac{\sqrt{n}}{\lambda_n} R_n(\hat{\theta}^{lasso} - \theta_n^*) \Longrightarrow \left(w'_{I(0)}, w'_{I(1)}\right)'$$

where

$$\begin{aligned} w_{I(0)} &\equiv \arg \min_{v \in \mathbb{R}^{p_z+p_1}} \left\{ v' \Omega_{zz}^{+} v + D\left(\mathbf{1}_{p_z+p_1}, v, \alpha^{+*}\right) \right\}, \\ w_{I(1)} &\equiv 0. \end{aligned}$$

Remark 3.8 *In Corollary 3.7(a), the tuning parameter is too small and the limit theory of plain LASSO is equivalent to that of OLS; there is no variable screening effect. When the tuning parameter is raised to the case of (b), the plain LASSO screens variables in the stationary part, but the tuning parameter is still too small for variable screening in the nonstationary part. Such difficulty is caused by the different rates of convergence between the estimated coefficients associated with the stationary regressors and the nonstationary ones. Since the plain LASSO has one single rate for the tuning parameter, it is not adaptive to deal with these two types of predictors. There is no way to achieve, simultaneously in both types of predictors, the same rate of convergence as OLS and screening variables. If we further increase the tuning parameter as in the case of (c), then the slower rate of convergence of the I(0) part drags the rate of $\widehat{\beta}^{+lasso}$ from n down to $n^{3/2}/\lambda_n$. Moreover, it implies inconsistency of $\hat{\alpha}^{+lasso}$ if $\lambda_n/n \to c_\lambda \in (0, \infty)$.*

Let us now turn to standardized LASSO, defined as

$$\hat{\theta}^{slasso} = \arg \min_{\theta \in \mathbb{R}^p} \|y - W\theta\|_2^2 + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j \left|\theta_j\right|. \tag{16}$$

We investigate whether the standardized LASSO restores the optimal rate of convergence and variable screening effect in this linear regression with mixed roots. In standardized LASSO, the stationary regressors are accompanied with $\widehat{\sigma}_j = O_p(1)$ for $j = 1, \ldots, p_z + p_1$, while the nonstationary regressors are coupled with $\widehat{\sigma}_j = O_p(\sqrt{n})$ for $j = p_z + p_1 + 1, \ldots p$. According to the following results on the asymptotic properties of the standardized LASSO, we do not have the rate adaptiveness.

Corollary 3.9 *Suppose the linear model (9) satisfies Assumption 3.1.*

(a) When $\lambda_n = [0, \infty)$, then

$$R_n(\hat{\theta}^{slasso} - \theta^*) \Longrightarrow \arg \min_{v \in \mathbb{R}^p} \left\{ v' \Omega^{+} v - 2v' \xi^{+} + c_\lambda D\left(\left(d_j, v_j, \phi_{0j}^*\right)_{j=p_z+1}^{p_z+p_1} \right) \right\}$$

where $c_\lambda = 0$ case restores OLS limit theory.

(b) When $\lambda_n \to \infty$ and $\lambda_n / n^{(1-\bar{\delta})\wedge 0.5} \to 0$, then

$$\frac{R_n}{\lambda_n}(\hat{\theta}^{slasso} - \theta^*) \Longrightarrow \arg\min_{v\in\mathbb{R}^p} \left\{ v'\Omega^+ v + c_\lambda D\left((d_j, v_j, \phi_{0j}^*)_{j=p_z+1}^{p_z+p_1} \right) \right\}.$$

Remark 3.10 *The difficulty of standardized LASSO arises from the coefficients associated with the cointegrating residuals. In OLS, these estimates $\widehat{\phi}_j$, $j = 1, \ldots, p_1$, converges at $\sqrt{n}$ rate. However, in the standardized LASSO their corresponding penalty have the multipliers of $\widehat{\sigma}_j = O_p(\sqrt{n})$, instead of the desirable $O_p(1)$. In other words, the penalty level is too heavy for these parameters. The overwhelming penalty produces variable screening effect as soon as $\lambda_n = c_\lambda \in (0, \infty)$, as shown in Corollary 3.9(a). Moreover, (b) implies that for the consistency of $\widehat{\phi}_1$ the tuning parameter λ_n must be small enough so that $\lambda_n/\sqrt{n} \to 0$. In this case, no variable screening is possible for all other coefficients in θ. If we further raise λ_n to $\lambda_n/\sqrt{n} \to c_\lambda \in (0, \infty)$, those $\widehat{\phi}_1$ will be an inconsistent estimator for ϕ_1^0.*

To sum up this section, in the general model with various types of regressors, the adaptive LASSO maintains the oracle property under the standard choice of the tuning parameter. It echoes our finding in Section 2, which is one special case of the model in this Section. In contrast, the plain LASSO using the single tuning parameter does not adapt to the different order of magnitudes of the stationary and nonstationary regressors. The standardized LASSO suffers from overwhelming penalties for those coefficients associated with the cointegration residuals. Keeping an agnostic view about the persistence property of the regressors, we recommend the adaptive LASSO in the multivariate predictive regression with mixed regressor persistence.

4 Monte Carlo Simulation

In this Section, we examine the performance of forecasting and variable screening of LASSO methods via simulation. We consider the different sample sizes n to demonstrate the validity of limit theory as well as the finite sample performance. All the comparison is based on the one-period-ahead forecast $\hat{y}_{n+1}$.

4.1 Simulation Design

To evaluate the finite sample performance of various estimators, we consider two data generating processes (DGPs), one with unit root regressors and the other with mixed roots and cointegration. In the Appendix B.1, two more DGPs using lagged dependent variable as regressors are included.

DGP 1 (Unit roots). We consider a linear model with eight unit root predictors, $x_i =$

$\begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{i8} \end{pmatrix}'$ where x_{ij} are drawn from independent random walk processes $x_{ij} = x_{i-1,j} + e_{ij}$, $e_{ij} \sim$ i.i.d. $N(0,1)$. The dependent variable y_i is generated by $y_{i+1} = \gamma^* + x_i' \beta_n^* + u_i$ where the intercept $\gamma^* = 0.25$, and $\beta_n^* = (1,1,1,1,0,0,0,0)'/\sqrt{n}$. The idiosyncratic error u_i follows i.i.d. standard normal distribution, so does those u_i's in the other three DGPs.

DGP 2 (Mixed roots and cointegration). This DGP corresponds to the generalized model in Section 3. The dependent variable y_i is generated by $y_i = \gamma^* + \sum_{j=1}^{2} z_{ij}\alpha_j^* + \sum_{j=1}^{4} x_{ij}^c \phi_{jn}^* + \sum_{j=1}^{2} x_{ij}\beta_{jn}^* + u_i$, where $\theta^* = (\alpha^*, \phi_n^*, \beta_n^*) = \left(0.4,\ 0,\ 0.3,\ -0.3,\ 0,\ 0,\ \frac{1}{\sqrt{n}},\ 0\right)$ and $\gamma^* = 0.3$. The stationary regressors z_{i1} and z_{i2} follow two independent AR(1) processes with the same AR(1) coefficient 0.5. $X_i^c \in \mathbb{R}^4$ is an I(1) process with cointegrating rank 2 based on the vector error correction model (VECM) $\Delta X_i^c = \Gamma' \Lambda X_{i-1}^c + e_i$, where $\Lambda = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$ and $\Gamma = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$ are the cointegrating matrix and the loading matrix, respectively. In the error term $e_i = (e_{i1},\ e_{i2},\ e_{i3},\ e_{i4})'$, we set $e_{i2} = e_{i1} - \mu_i$ and $e_{i4} = e_{i3} - \nu_i$ where μ_i and ν_i are AR(1) processes with the AR(1) coefficient 0.2. x_{i1} and x_{i2} are independent random walk as those in DGP 1.

As we develop our theory with fixed-dimensional regressors, the OLS is a natural benchmark. Another benchmark is the oracle OLS, in which the oracle reveals the true model. In reality, the oracle OLS estimator is of course infeasible. The sample sizes in our exercise range from $n = 40$, 80, 120, 200 and 400. For each simulation setting, we generate data with 1000 burn-in periods and run 1000 replications for each sample size n.

For the shrinkage estimators, we do not penalize the intercept in the simulations as well as in the empirical application. Each shrinkage estimator relies on its tuning parameter λ_n, which is the standard convergence rate $\sqrt{n}$ multiplied by a constant c_λ. We use 10-fold cross validation to guide the choice of c_λ. Specifically, we set $n = 200$ and run an exploratory simulation for 100 times for each method that needs a tuning parameter[3]. In each replication, we use the 10-fold cross-validation to obtain $c_\lambda^{(1)}, \ldots, c_\lambda^{(100)}$. Then we fix $c_\lambda = \text{median}\left(c_\lambda^{(1)}, \ldots, c_\lambda^{(100)}\right)$ in the full-scale 1000-time simulation.

The tuning parameter of plain LASSO and standardized LASSO are $\lambda_n = c_\lambda^{plasso}\sqrt{n}$, and $\lambda_n = c_\lambda^{slasso}\sqrt{n}$. The OLS estimator is used as the initial estimator in the adaptive LASSO, and the tuning parameter λ_n is set similarly.

4.2 Performance Comparison

Table 1 reports the out-of-sample prediction accuracy in terms of the mean prediction squared error (MPSE), MPSE $= E\left[(y_{T+1} - \hat{y}_{T+1})^2\right]$. By the simulation design, the variance of the id-

[3]For comparison, we also include simulation results in which the tuning parameter is determined by cross-validation for each sample size separately in Appendix B.3.

iosyncratic error is 1, which is the unpredictable part. Table 2 summarizes the variable screening performance. Recall that the set of relevant regressors as $A^* = \left\{j \in \{1,\ldots,p\} : \theta_j^* \neq 0\right\}$ and the estimated active set is $\widehat{A} = \left\{j \in \{1,\ldots,p\} : \hat{\theta}_j \neq 0\right\}$. We define two *correct ratios* for variable screening:

$$CR_1 = \frac{1}{|A^*|} E\left[\left|\left\{j : j \in \widehat{A},\, j \in A^*\right\}\right|\right], \quad CR_2 = \frac{1}{|A^{*c}|} E\left[\left|\left\{j : j \in A^{*c},\, j \in \widehat{A}^c\right\}\right|\right].$$

Here CR_1 is the percentage of the correct selection in the active set, whereas CR_2 is the percentage of correct elimination of the zero coefficients. We also report the overall correct ratio

$$CR = \frac{1}{p} E\left[\left|\left\{j : \text{ sign}\left(\theta_j^*\right) = \text{sign}\left(\widehat{\theta}_j\right)\right\}\right|\right].$$

These expectations are computed by the average in the 1000 simulations replications.

Table 1: Mean Prediction Squared Error (MPSE)

	n	Oracle	OLS	alasso	plasso	slasso
DGP 1	40	1.2064	1.4841	1.3388	**1.2259**	1.2695
	80	1.1886	1.2677	1.2540	**1.2267**	1.2294
	120	1.1035	1.1710	1.1459	1.1340	**1.1289**
	200	1.0940	1.1689	1.1429	1.1349	**1.1303**
	400	0.9775	1.0047	0.9969	**0.9941**	0.9959
DGP 2	40	1.2626	1.4900	1.3883	**1.3766**	1.4301
	80	1.1029	1.2156	**1.1906**	1.2051	1.2097
	120	1.0984	1.1640	**1.1465**	1.1566	1.1583
	200	1.1017	1.1523	**1.1241**	1.1386	1.1390
	400	0.9569	0.9722	**0.9605**	0.9660	0.9676

Note: Bold numbers are for the best performance among all the feasible estimators.

According to Table 1, plain LASSO and standardized LASSO achieve better forecasting performance than adaptive Lasso in DGP 1. In DGP 2, adaptive LASSO outperforms the competitors in MPSE except for the smallest sample size $n = 40$. The MPSE results can be explained by the variable screening results in Table 2.

The parameter tuning in adaptive LASSO is not as good as plain LASSO and standardized LASSO due to the estimated weights from the first step estimation, which means adaptive LASSO achieves better variable screening at the cost of estimation loss in finite sample. In DGP 1 with pure unit-root regressors, plain LASSO and standardized LASSO achieve good performance in terms of CR, not far behind adaptive LASSO in large sample size cases and even better than adaptive LASSO in small sample size cases. Considering the trade-off between variable screening and coefficient estimation accuracy, it is understandable that plain LASSO and standardized LASSO have better forecasting performance in DGP 1. The advantage of

Table 2: Variable Screening

		CR			CR_1			CR_2		
	n	alasso	plasso	slasso	alasso	plasso	slasso	alasso	plasso	slasso
DGP 1	40	0.5885	**0.6366**	0.6000	0.7653	0.6408	**0.8178**	0.4118	**0.6325**	0.3823
	80	0.6606	**0.6776**	0.6339	0.8268	0.8248	**0.8918**	0.4945	**0.5305**	0.3760
	120	**0.7080**	0.6860	0.6581	0.8868	0.9095	**0.9395**	**0.5293**	0.4625	0.3768
	200	**0.7619**	0.6739	0.6735	0.9365	0.9673	**0.9713**	**0.5873**	0.3805	0.3758
	400	**0.8311**	0.6361	0.6794	0.9810	**0.9930**	**0.9930**	**0.6813**	0.2793	0.3658
DGP 2	40	**0.6665**	0.5780	0.5431	0.8213	0.9165	**0.9628**	**0.5118**	0.2395	0.1235
	80	**0.7636**	0.6065	0.5664	0.9245	0.9855	**0.9910**	**0.6028**	0.2275	0.1418
	120	**0.8060**	0.5986	0.5750	0.9610	0.9950	**0.9973**	**0.6510**	0.2023	0.1528
	200	**0.8385**	0.5915	0.5808	0.9880	0.9990	**0.9993**	**0.6890**	0.1840	0.1623
	400	**0.8728**	0.5905	0.5994	0.9978	**1.0000**	**1.0000**	**0.7478**	0.1810	0.1988

Note: Bold numbers are for the best performance.

adaptive LASSO in variable screening is prominent in DGP 2 as the DGP becomes more sophisticated with mixed roots and cointegration. So the adaptive LASSO outperforms the others in forecasting performance.

Adaptive LASSO outperforms the others in CR and CR_2 in both DGPs. As sample size increases, all CR, CR_1 and CR_2 of adaptive LASSO increases in both DGPs, which verifies the variable screening consistency of adaptive LASSO. The asymptotic theory suggests $\lambda_n \sim \sqrt{n}$ is too small for plain LASSO to eliminate 0 coefficients corresponding to I(1) regressors, which is consistent to the CR_2 results of plain LASSO that CR_2 decreases as the sample size increases. Plain LASSO and standardized LASSO achieve high CR_1 at the cost of low CR_2, i.e. they tend to keep more regressors even some of the selected ones are redundant. As sample size increases, the difference in CR_1 among methods becomes negligible.

According to Table 2, standardized LASSO has the lowest variable elimination correct ratio CR_2, whereas in asymptotics it imposes heavier penalty on coefficients of I(1) regressors than plain LASSO does due to the presence of $\hat{w}_j = \hat{\sigma}_j = O_p(\sqrt{n})$ in the penalty term. The reason is that we fix c_λ^{plasso} and c_λ^{slasso} by cross-validation separately. The cross-validation selects tuning parameters based in-sample MSE and hence favors c_λ achieving lower MPSE instead of higher variable screening correct ratio. For example, in DGP 1, $c_\lambda^{plasso} = 1.295$ whereas $c_\lambda^{slasso} = 0.265$ which is much smaller than c_λ^{plasso}. If we fix c_λ^{plasso} by cross-validation and let $c_\lambda^{slasso} = c_\lambda^{plasso}$, CR_2 of standardized LASSO would become higher. We leave this simulation result in Appendix B.2.

5 Empirical Application

To illustrate the performance of Adaptive Lasso in predictive regression and compare with that of other common approaches, this section presents an empirical study on stock return predictability with the updated Welch and Goyal (2008) dataset.[4]

5.1 Data

Following Welch and Goyal (2008) and extending the time span, we use the monthly data from January 1945 to December 2016, with the dependent variable, excess return, defined as the difference between the continuously compounded return on the S&P 500 index and the three-month Treasury bill rate[5]

$$\text{ExReturn}_i = \log\left(\text{index}_i/\text{index}_{i-1}\right) - \log\left(1 + \text{tbl}_i/12\right)$$

and 12 predictors[6].

Table 3 summarizes all the variables included and the first-order autoregression coefficient of each variable estimated with the whole sample period. The excess return has an estimated AR(1) coefficient of 0.0444, which indicates little persistence, similar to the default return spread (dfr), the long-term return of government bonds (ltr), stock variance (svar) and inflation (infl). The other predictors show high persistence, with AR(1) coefficients greater than 0.95. The mixture of stationary predictors and persistent predictors fits the mixed roots environment that we studied in previous Sections.

5.2 Performance Comparison

We apply the set of feasible forecasting methods as in Section 4 to forecast one-month-ahead excess returns recursively with both 20-year and 30-year rolling windows. In addition to OLS, we include the random walk with drift (RWwD), i.e. we take the historical average of the excess returns as the forecast, $\hat{y}_{n+1} = \frac{1}{n}\sum_i^n y_i$ as a benchmark . All variables are included in the predictive regression, which is referred to the kitchen sink model in Welch and Goyal (2008). The forecasting performance is evaluated based on the out-of-sample MPSE and *percentage* defined as the ratio of the MPSE of a particular method over that of OLS.

The results summarized in Table 4 shows that adaptive LASSO has the exactly same forecasting performance as historical average. Plain LASSO and standardized LASSO also

[4] Retrieved from http://www.hec.unil.ch/agoyal/

[5] In the raw data, the monthly variable tbl is the annualized rate of the three-month T-bill, which should be transformed to monthly terms to be compatible with the index return.

[6] Following Koo et al. (2016), we exclude the dividend payout ratio (de) and long-term yield (lty) from the 14 predictors in Welch and Goyal (2008) for comparison.

Table 3: Variables and AR(1) Coefficients

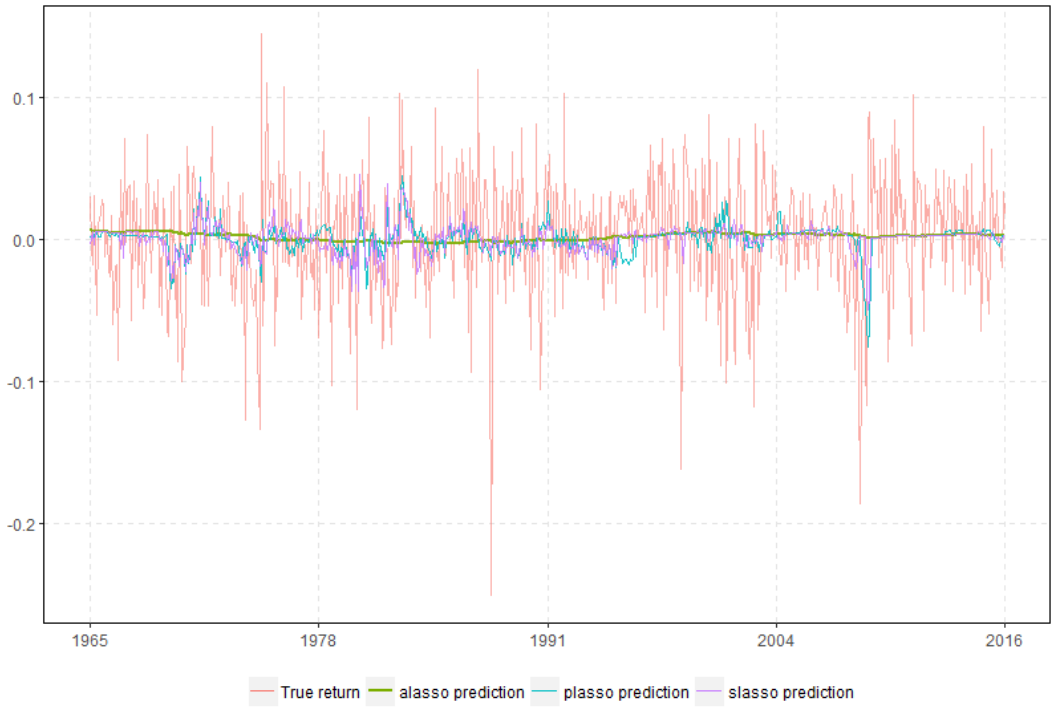
Predictor	Description	AR(1) Coef.
ExReturn	Excess Return: the difference between the continuously compounded return on the S&P 500 index and the three-month Treasury Bill rate	0.0444
dp	Dividend Price Ratio: the difference between the log of the 12-month moving sum dividends and the log of the S&P 500 index	0.9941
dy	Dividend Yield: the difference between the log of the 12-month moving sum dividends and the log of lagged the S&P 500 index	0.9941
ep	Earning Price Ratio: the difference between the log of the 12-month moving sum earnings and the log of the S&P 500 index	0.9904
tms	Term Spread: the difference between the long-term government bond yield and the Treasury Bill rate	0.9576
dfy	Default Yield Spread: the difference between Moody's BAA and AAA-rated corporate bond yields	0.9717
dfr	Default Return Spread: the difference between the returns of long-term corporate bonds and long-term government bonds	-0.0735
bm	Book-to-Market Ratio: the ratio of the book value to market value for the Dow Jones Industrial Average	0.9927
tbl	Treasury Bill Rates: the 3-month Treasury Bill rates	0.9905
ltr	Long-Term Return: the rate of returns of long-term government bonds	0.0500
ntis	Net Equity Expansion: the ratio of the 12-month moving sums of net issues by NYSE listed stocks over the total end-of-year market capitalization of NYSE stocks	0.9778
svar	Stock Variance: the sum of the squared daily returns on the S&P 500 index	0.4714
infl	Inflation: the log growth of the Consumer Price Index (all urban consumers)	0.4819

Table 4: Mean Prediction Squared Error (MPSE)

		RWwD	OLS	alasso	plasso	slasso
20-year	MPSE	**0.00191**	0.00209	**0.00191**	0.00195	0.00192
	Percentage	**0.91672**	1.00000	**0.91672**	0.93509	0.92155
30-year	MPSE	**0.00190**	0.00204	**0.00190**	0.00196	0.00193
	Percentage	**0.93319**	1.00000	**0.93319**	0.96115	0.94767

Note: Bold numbers are for the best performance.

Figure 1: True Monthly Return vs Predicted Return (20-Year Rolling Windows)



improves OLS but performs worse than adaptive LASSO. Figure 5.2 plots the true monthly return and predicted returns. Adaptive LASSO provides relatively conservative forecasts whereas plain LASSO and standardized LASSO are more aggressive.

As documented in Welch and Goyal (2008), historical average sets a benchmark of stock return forecasting that is hard to beat, which implies weak signals and quick mean reversion in returns. There is also no consensus in the literature that any of the predictors has significant predictive power. As a result, it is not surprising that adaptive LASSO eliminates all predictors from the model and performs identically to historical average. If we believe that none of 12 predictors in the model are relevant "enough" in the true DGP, adaptive LASSO achieves variable screening successfully. Hirano and Wright (2017) also report the similar conclusion on stock return prediction application.

5.3 Long-horizon Return Forecasting

As recognized in the literature, the signal of persistent predictors may become stronger in long-horizon return prediction; see Cochrane (2009). We use the monthly Welch and Goyal (2008) data set from January 1945 to December 2016 and construct the long-horizon excess

Table 5: Mean Prediction Squared Error (MPSE)

	MPSE				Percentage relative to OLS			
	OLS	alasso	plasso	slasso	OLS	alasso	plasso	slasso
20-years								
$h=1$	0.0589	**0.0277**	0.0419	0.0417	1.0000	**0.4696**	0.7116	0.7078
$h=2$	0.1149	**0.0628**	0.0851	0.0852	1.0000	**0.5466**	0.7405	0.7414
$h=3$	0.1859	**0.1044**	0.1404	0.1634	1.0000	**0.5618**	0.7555	0.8791
$h=4$	0.2927	**0.1502**	0.2235	0.2302	1.0000	**0.5133**	0.7636	0.7866
$h=5$	0.2611	**0.2056**	0.2431	0.2626	1.0000	**0.7875**	0.9310	1.0056
$h=6$	0.4148	**0.2727**	0.3341	0.3466	1.0000	**0.6573**	0.8054	0.8357
30-year								
$h=1$	0.0551	**0.0264**	0.0319	0.0316	1.0000	**0.4782**	0.5779	0.5730
$h=2$	0.1095	**0.0610**	0.0866	0.0814	1.0000	**0.5571**	0.7906	0.7432
$h=3$	0.1267	**0.0983**	0.1099	0.1187	1.0000	**0.7758**	0.8676	0.9365
$h=4$	0.1642	**0.1215**	0.1326	0.1490	1.0000	**0.7399**	0.8077	0.9074
$h=5$	0.2185	**0.1535**	0.2307	0.2100	1.0000	**0.7025**	1.0555	0.9611
$h=6$	0.3905	**0.1731**	0.3627	0.3064	1.0000	**0.4434**	0.9288	0.7847

return as the sum of continuous compounded monthly excess return on the S&P 500 index,

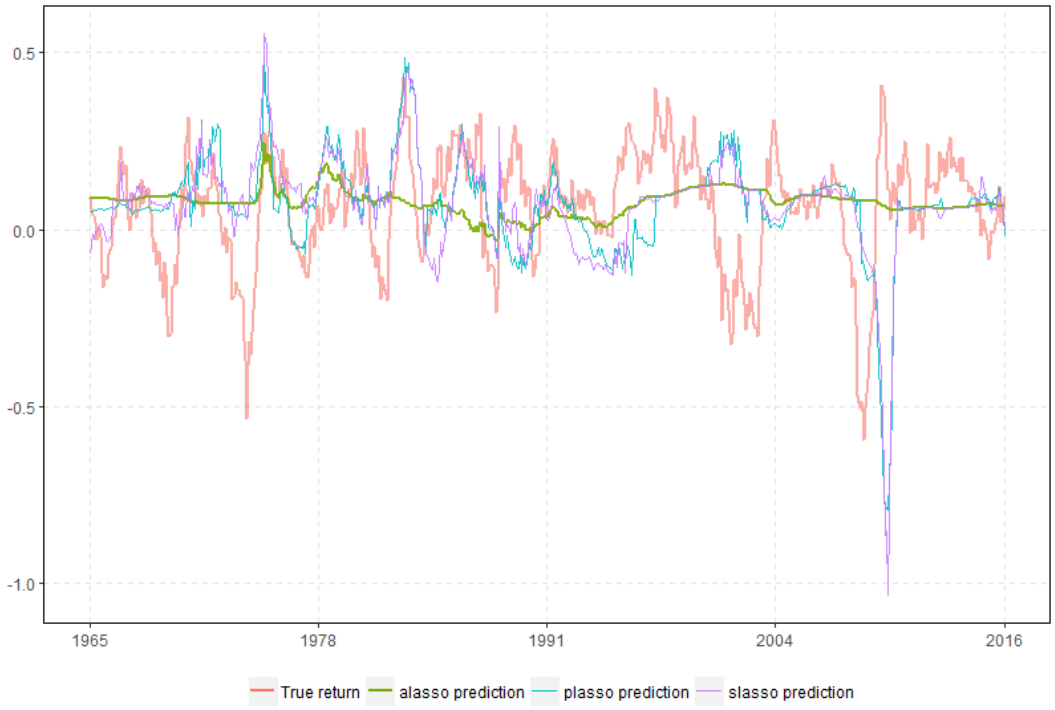
$$\text{ExReturn}_i = \log\left(\text{index}_i/\text{index}_{i-1}\right), \quad \text{LongReturn}_i = \sum_{k=i}^{i+12\times h-1} \text{ExReturn}_k$$

where h is the length of the forecasting horizon in terms of year and ranges from $h = 1, 2, \cdots, 6$.

We use the 12 predictors in Section 5.1 to predict the long-horizon return recursively with both 20-year and 30-year rolling windows. Since LongReturn_i has overlapping horizons, it is not reasonable to compute the historical average as a forecast, but we include all the other forecasting methods for comparison. The forecasting performance is summarized in Table 5.3, and Figure 2 plots the true long-horizon return and predicted values with 20-year rolling window and $h = 1$.

Adaptive Lasso outperforms the others uniformly in all cases, improving OLS up to more than 50%. According to Figure 2, the long-horizon return fluctuates with lower frequency than the monthly return. During certain periods, the predictors seem to demonstrate predictive power. Plain LASSO and standardized LASSO may track the fluctuation of returns but may also generate totally opposite forecasts. Adaptive LASSO demonstrates its merits in variable screening to select a proper subset of predictors within the rolling window.

Figure 2: Long-horizon Return vs Predicted Return (20-Year Rolling Windows, $h = 1$)



References

[1] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C., 2012. "Sparse models and methods for optimal instruments with an application to eminent domain." *Econometrica*, 80(6), pp.2369-2429.

[2] Belloni, A. and Chernozhukov, V., 2011. "L1-penalized quantile regression in high-dimensional sparse models." *The Annals of Statistics*, 39(1), pp.82-130.

[3] Belloni, A., Chernozhukov, V., Chetverikov, D. and Wei, Y., 2015. Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework. arXiv preprint arXiv:1512.07619.

[4] 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), pp.608-650.

[5] Bickel, P. J., Ritov, Y. A., & Tsybakov, A. B. (2009). Simultaneous analysis of LASSO and Dantzig selector. *The Annals of Statistics*, 1705-1732.

[6] Caner, M., 2009. "Lasso-type GMM estimator." *Econometric Theory*, 25(1), pp.270-290.

[7] Caner, M. and Knight, K., 2013. "An alternative to unit root tests: Bridge estimators differentiate between nonstationary versus stationary models and select optimal lag." Journal of Statistical Planning and Inference, 143(4), pp.691-715.

[8] Caner, M. and Zhang, H.H., 2014. Adaptive elastic net for generalized methods of moments. Journal of Business & Economic Statistics, 32(1), pp.30-47.

[9] Cochrane, J. H. (2009). Asset pricing (*revised edition*). Princeton University Press.

[10] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association, 96(456), 1348-1360.

[11] Hirano, K, and J. H. Wright. "Forecasting with Model Uncertainty: Representations and Risk Reduction." *Econometrica* 85.2 (2017): 617-643.

[12] Inoue, A., and Lutz K. "How useful is bagging in forecasting economic time series? A case study of US consumer price inflation." Journal of the American Statistical Association 103.482 (2008): 511-522.

[13] Kock, A.B., 2016. Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions. Econometric Theory, 32(1), pp.243-259.

[14] Kock, A.B. and Callot, L., 2015. Oracle inequalities for high dimensional vector autoregressions. Journal of Econometrics, 186(2), pp.325-344.

[15] Koo, B., Anderson, H. M., Seo, M. H., & Yao, W. (2016). High-dimensional predictive regression in the presence of cointegration, working paper.

[16] Kostakis, A., Magdalinos, T., & Stamatogiannis, M. P. (2014). Robust econometric inference for stock return predictability. *The Review of Financial Studies*, 28(5), 1506-1553.

[17] Lee, J. H. (2016). Predictive quantile regression with persistent covariates: IVX-QR approach. *Journal of Econometrics*, 192(1), 105-118.

[18] Liao, Z., & Phillips, P. C. (2015). Automated estimation of vector error correction models. *Econometric Theory*, 31(3), 581-646.

[19] Medeiros, M. C., & Mendes, E. F. (2016). L1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1), 255-271.

[20] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2), 186-199.

[21] Phillips, P. C. (1987). Time series regression with a unit root. *Econometrica: Journal of the Econometric Society*, 277-301.

[22] Phillips, P.C., 1991. Optimal inference in cointegrated systems. Econometrica: Journal of the Econometric Society, pp.283-306.

[23] Phillips, P.C., 2005. Automated discovery in econometrics. Econometric Theory, 21(1), pp.3-20.

[24] Phillips, P. C., & Lee, J. H. (2013). Predictive regression under various degrees of persistence and robust long-horizon regression. *Journal of Econometrics*, 177(2), 250-264.

[25] Phillips, P. C., & Lee, J. H. (2016). Robust econometric inference with mixed integrated and mildly explosive regressors. *Journal of Econometrics*, 192(2), 433-450.

[26] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

[27] Timmermann, A., & Zhu, Y. (2017). Monitoring forecasting performance. Working paper.

[28] Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4), 1455-1508.

[29] Xu, K. L. (2017). Testing for Return Predictability with Co-moving Predictors of Unknown Form, working paper.

[30] Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.

[31] Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), pp.301-320.

A Technical Appendix

A.1 Proofs in Section 2

Proof. [Proof of Theorem 2.1] We modify the proof of Zou (2006, Theorem 2). Let $\beta_n = \beta_n^* + n^{-1}v$ be a perturbation from the original parameter β_n^*, and let

$$\Psi_n(v) = \|Y - \sum_{j=1}^{p} x_j(\beta_{jn}^* + \frac{v_j}{n})\|^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j|\beta_{jn}^* + \frac{v_j}{n}|.$$

Define $\hat{v}^{(n)} = n(\hat{\beta}^{alasso} - \beta_n^*)$. Since $\hat{\beta}^{alasso}$ is the minimizer of (5), $\hat{v}^{(n)} = \arg\min_v \Psi_n(v)$. Let

$$\begin{aligned} V_n(v) &= \Psi_n(v) - \Psi_n(0) \\ &= \|u - \frac{X'v}{n}\|^2 - \|u\|^2 + \lambda_n \left(\sum_{j=1}^{p} \hat{w}_j|\beta_{jn}^* + \frac{v_j}{n}| - \sum_{j=1}^{p} \hat{w}_j|\beta_{jn}^*| \right) \\ &= v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \lambda_n \sum_{j=1}^{p} \hat{w}_j(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|). \end{aligned} \tag{17}$$

By FCLT and the continuous mapping theorem, the first term and the second term of (17) converge in distribution, $\frac{X'X}{n^2} \Longrightarrow \Omega$ and $\frac{u'X}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{i\cdot}' u_i \Longrightarrow W$, respectively. We thus focus on the third term.

The third term involves the weight for each j, $\hat{w}_j = |\widehat{\beta}_j^{ols}|^{-\gamma}$. Since the OLS estimator $n\left(\widehat{\beta}^{ols} - \beta_n^*\right) \Longrightarrow \Omega^{-1}W = O_p(1)$, we have

$$\hat{w}_j = \left|\beta_{jn}^* + O_p\left(n^{-1}\right)\right|^{-\gamma} = |\beta_{0j}^*/n^{\delta_j} + O_p\left(n^{-1}\right)|^{-\gamma}, \tag{18}$$

for all j. If $\beta_{0j}^* \neq 0$, as the β_{jn}^* dominates $n^{-1}v_j$ for a large n,

$$(|\beta_{jn}^* + n^{-1}v_j| - |\beta_{jn}^*|) = n^{-1}v_j \mathrm{sgn}(\beta_{jn}^*) = n^{-1}v_j \mathrm{sgn}(\beta_{0j}^*). \tag{19}$$

(18) and (19) now imply

$$\begin{aligned} \lambda_n \hat{w}_j \cdot (|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|) &= \frac{\lambda_n}{|\beta_{0j}^*/n^{\delta_j} + O_p\left(n^{-1}\right)|^{\gamma} n} v_j \mathrm{sgn}(\beta_{0j}^*) = \frac{\lambda_n n^{\delta_j \cdot \gamma - 1}}{|\beta_{0j}^* + o_p\left(1\right)|^{\gamma}} v_j \mathrm{sgn}(\beta_{0j}^*) \\ &\leq \frac{\lambda_n n^{\gamma\bar{\delta} - 1}}{|\beta_{0j}^* + o_p\left(1\right)|^{\gamma}} v_j \mathrm{sgn}(\beta_{0j}^*) = O_p\left(\lambda_n n^{\gamma\bar{\delta}-1}\right) = o_p\left(1\right), \end{aligned} \tag{20}$$

by the given rate of λ_n. On the other hand, if $\beta_{0j}^* = 0$, then $(|\beta_{jn}^* + n^{-1}v_j| - |\beta_{jn}^*|) = n^{-1}|v_j|$. For any fixed $v_j \neq 0$,

$$\lambda_n \hat{w}_j \cdot n(|\beta_j^* + n^{-1}v_j| - |\beta_j^*|) = \frac{\lambda_n}{n|\widehat{\beta}_j^{ols}|^\gamma}|v_j| = \frac{\lambda_n n^{\gamma-1}}{|n\widehat{\beta}_j^{ols}|^\gamma}|v_j| = \frac{\lambda_n n^{\gamma-1}}{O_p(1)}|v_j| \to \infty. \tag{21}$$

as $\lambda_n n^{\gamma-1} \to \infty$. Thus we have $V_n(v) \Longrightarrow V(v)$ for every fixed v, where

$$V(v) = \begin{cases} v'\Omega v - 2v'W, & \text{if } v_{A^{*c}} = 0 \\ \infty, & \text{otherwise.} \end{cases}$$

Both $V_n(v)$ and $V(v)$ are strictly convex in v, and $V(v)$ is uniquely minimized at

$$\left(v_{A^*} = \Omega_{A^*}^{-1}W_{A^*}, v_{A^{*c}} = 0\right).$$

Applying the Convexity Lemma (Pollard, 1991), we have

$$\hat{v}_{A^*}^{(n)} = n(\hat{\beta}_{A^*}^{alasso} - \beta_{A^*}^*) \Longrightarrow \Omega_{A^*}^{-1}W_{A^*} \text{ and } \hat{v}_{A^{*c}}^{(n)} \Longrightarrow 0. \tag{22}$$

The first part of the above result establishes Theorem 2.1(b).

Next we show variable selection consistency. We have $P(A^* \subseteq A_n) \to 1$ immediately follows from the first part of (22) as $\hat{v}_{A^*}^{(n)}$ converges in distribution to a non-degenerate continuous random variable. For those $j \notin A^*$, if the event $\{j \in A_n\}$ occurs, then the KKT optimality condition entails

$$\frac{2}{n}x_j'(y - X\hat{\beta}^{alasso}) = \frac{\lambda_n \hat{w}_j}{n}. \tag{23}$$

Notice that on the right-hand side of the KKT condition

$$\frac{\lambda_n \hat{w}_j}{n} = \frac{\lambda_n}{n|\widehat{\beta}_j^{ols}|^\gamma} = \frac{\lambda_n n^{\gamma-1}}{|n\widehat{\beta}_j^{ols}|^\gamma} = \frac{\lambda_n n^{\gamma-1}}{O_p(1)} \to \infty, \tag{24}$$

from the given rate condition. However, looking at the left-hand side of (23), using $y = X\beta_n^* + u$ and (22) we have

$$\begin{aligned} \frac{2}{n}x_j'(y - X\hat{\beta}^{alasso}) &= \frac{2}{n}x_j'(X\beta_n^* - X\hat{\beta}^{alasso} + u) = 2\left(\frac{x_j'X}{n^2}\right)n(\beta_n^* - \hat{\beta}^{alasso}) + 2\frac{x_j'u}{n} \\ &= 2\left(\frac{x_j'X}{n^2}\right)\left(\hat{v}_{A^*}^{(n)} + \widehat{v}_{A^{*c}}^{(n)}\right) + 2\frac{x_j'u}{n} \\ &\Longrightarrow 2\Omega_{\cdot j} \cdot (\Omega_{A^*}^{-1}W_{A^*} + o_p(1)) + 2W_j. \end{aligned} \tag{25}$$

In other words, the left-hand side of (23) remains as a non-degenerate continuous random

variable in the limit. For any $j \in A^{*c}$, the disparity of the two sides of the KKT condition implies $P(j \in A_n) = P\left(\frac{2}{n}x_j'(y - X\hat{\beta}^{alasso}) = \frac{\lambda_n \hat{w}_j}{n}\right) \to 0$. That is, $P(A^{*c} \subseteq A_n^c) \to 1$ or equivalently $P(A_n \subseteq A^*) \to 1$. We thus conclude the variable selection consistency $P(A_n = A^*) \to 1$. ■

Proof. [Proof of Corollary 2.3] The proof is a simple variant of that of Theorem 2.1 by setting $\widehat{w}_j = 1$ for all j. For Part(a), the counterpart of (17) is

$$V_n(v) = v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \lambda_n \sum_{j=1}^{p}(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|).$$

For a fixed v_j and a sufficiently large n,

$$\lambda_n(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|) = \frac{\lambda_n v_j}{n}\text{sgn}(\beta_{0j}^*) = O\left(\frac{\lambda_n}{n}\right), \text{ if } \beta_{0j}^* \neq 0;$$
$$\lambda_n(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|) = \lambda_n\frac{|v_j|}{n} = O\left(\frac{\lambda_n}{n}\right), \text{ if } \beta_{0j}^* = 0.$$

Since $\lambda_n/n \to 0$, the effect of the penalty term is negligible. We have $V_n(v) \Longrightarrow V(v)$ for every fixed v, and furthermore $V_4(v) = v'\Omega v - 2v'W$. Due to the strict convexity of $V_n(v)$ and $V(v)$, the Convexity Lemma implies

$$n\left(\widehat{\beta}^{lasso} - \beta_n^*\right) = \hat{v}^{(n)} \Longrightarrow \Omega^{-1}W.$$

In other words, the LASSO estimator has the same asymptotic distribution of the OLS estimator.

For Part (b), as $\lambda_n/n \to c_\lambda \in (0, \infty)$, the effect of the penalty emerges as

$$V_n(v) = v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \frac{\lambda_n}{n}D(1_p, v, \beta_0^*) \Longrightarrow v'\Omega v - 2v'W + c_\lambda D(\mathbf{1}_p, v, \beta_0^*).$$

The conclusion of the statement again follows by the Convexity Lemma.

For Part (c), we define a new perturbation $\beta_n = \beta_n^* + \frac{\lambda_n}{n^2}v$, and

$$\Psi_n(v) = \|Y - X\left(\beta_n^* + \frac{\lambda_n}{n^2}v\right)\|^2 + \lambda_n \sum_{j=1}^{p}|\beta_{jn}^* + \frac{\lambda_n}{n^2}v_j|,$$
$$V_n(v) = \Psi_n(v) - \Psi_n(0) = \frac{\lambda_n^2}{n^4}v'(X'X)v - \frac{\lambda_n}{n^2}2u'Xv + \lambda_n \sum_{j=1}^{p}(|\beta_{jn}^* + \frac{\lambda_n}{n^2}v| - |\beta_{jn}^*|).$$

Given the rate $\frac{\lambda_n}{n^{2-\delta}} \to 0$, $\frac{\lambda_n}{n^2}v$ is dominated by any $\beta_{jn}^* = \beta_{0j}^*/n^{\delta_j}$ if $\beta_{0j}^* \neq 0$ in the limit. So,

for a sufficiently large n,

$$\begin{aligned} V_n(v) &= \frac{\lambda_n^2}{n^2} v'(\frac{X'X}{n^2})v - \frac{\lambda_n}{n} 2\left(\frac{u'X}{n}\right)v + \frac{\lambda_n^2}{n^2} D\left(\mathbf{1}_p, v, \beta_0^*\right) \\ &= \frac{\lambda_n^2}{n^2}\left[v'(\frac{X'X}{n^2})v - \frac{1}{\lambda_n/n} 2\left(\frac{u'X}{n}\right)v + D\left(\mathbf{1}_p, v, \beta_0^*\right)\right] \\ &= \frac{\lambda_n^2}{n^2}\left[v'(\frac{X'X}{n^2})v + o_p(1) + D\left(\mathbf{1}_p, v, \beta_0^*\right)\right]. \end{aligned}$$

Notice that the scaled difference $\hat{v}^{(n)} = \lambda_n^{-1} n^2(\hat{\beta}^{lasso} - \beta_n^*)$ can be expressed as $\hat{v}^{(n)} = \arg\min_v \Psi_n(v)$. By the strict convexity of $V_n(v)$ and $V(v) = v'\Omega v + D\left(\mathbf{1}_p, v, \beta_0^*\right)$, we invoke the Convexity Lemma to obtain $\frac{n^2}{\lambda_n}(\hat{\beta}^{lasso} - \beta_n^*) \Longrightarrow \arg\min_v V(v)$. ■

Proof. [Proof of Corollary 2.5] The standardized LASSO differs from the plain LASSO by setting the weight $\widehat{w}_j = \widehat{\sigma}_j$. For Part (a), we use the perturbation $\beta_n = \beta_n^* + n^{-1}v$, and

$$\begin{aligned} \Psi_n(v) &= \|Y - X\left(\beta_n^* + \frac{v}{n}\right)\|^2 + \lambda_n \sum_{j=1}^p \hat{\sigma}_j |\beta_{jn}^* + \frac{v_j}{n}|, \\ V_n(v) &= \Psi_n(v) - \Psi_n(0) = v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \lambda_n \sum_{j=1}^p \hat{\sigma}_j(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|). \end{aligned}$$

When $\lambda_n/\sqrt{n} \to c_\lambda \geq 0$ and $\frac{\widehat{\sigma}_j}{\sqrt{n}} \Longrightarrow d_j$ as in (8), the penalty term

$$\lambda_n \sum_{j=1}^p \hat{\sigma}_j(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|) = \frac{\lambda_n}{\sqrt{n}} D\left(\frac{\hat{\sigma}}{\sqrt{n}}, v, \beta_0^*\right) \Longrightarrow c_\lambda \sum_{j=1}^p D\left(d, v, \beta_0^*\right)$$

where $\widehat{\sigma} = (\widehat{\sigma}_j)_{j=1}^p$. Part (a) follows by the same argument in the proof of Corollary 2.3(b).

Part (b) here is similar to the proof of Corollary 2.3(c) by introducing a new perturbation $\beta_n = \beta_n^* + \frac{\lambda_n}{n^{3/2}}v$, and

$$\begin{aligned} \Psi_n(v) &= \|Y - X\left(\beta_n^* + \frac{\lambda_n}{n^{3/2}}v\right)\|^2 + \lambda_n \sum_{j=1}^p \hat{\sigma}_j |\beta_{jn}^* + \frac{\lambda_n}{n^{3/2}} v_j|, \\ V_n(v) &= \Psi_n(v) - \Psi_n(0) = \frac{\lambda_n^2}{n^3} v'(X'X)v - \frac{\lambda_n}{n^{3/2}} 2u'Xv + \lambda_n \sum_{j=1}^p \hat{\sigma}_j(|\beta_{jn}^* + \frac{\lambda_n}{n^{3/2}} v_j| - |\beta_{jn}^*|). \end{aligned}$$

Given the rate $\lambda_n/n^{\frac{3}{2}-\bar{\delta}} \to 0$, for a sufficiently large n we have

$$\lambda_n \hat{\sigma}_j \left(|\beta_{jn}^* + \frac{\lambda_n}{n^{3/2}} v_j| - |\beta_{jn}^*| \right) = \lambda_n D \left(\hat{\sigma}_j, \frac{\lambda_n}{n^{3/2}} v_j, \beta_{0j}^* \right) = \frac{\lambda_n^2}{n} D \left(\frac{\hat{\sigma}_j}{\sqrt{n}}, v_j, \beta_{0j}^* \right),$$

so that

$$\begin{aligned} V_n(v) &= \frac{\lambda_n^2}{n} \left[v'(\frac{X'X}{n^2})v - \frac{1}{\lambda_n/\sqrt{n}} 2 \left(\frac{u'X}{n} \right) v + D \left(\frac{\widehat{\sigma}}{\sqrt{n}}, v, \beta_0^* \right) \right] \\ &= \frac{\lambda_n^2}{n} \left[v'(\frac{X'X}{n^2})v + D \left(\frac{\widehat{\sigma}}{\sqrt{n}}, v, \beta_0^* \right) + o_p(1) \right]. \end{aligned}$$

Define $V(v) = v'\Omega v + D\left(d, v, \beta_0^*\right)$, and the conclusion follows by the same Convexity argument.

■

A.2 Proofs in Section 3

Proof. [Proof of Theorem 3.2] The OLS estimator

$$\begin{aligned} R_n \left(\widehat{\theta}_n^{ols} - \theta_n^* \right) &= R_n \left(W'W \right)^{-1} W'u \\ &= R_n \left(R_n Q \right)^{-1} R_n Q \left(W'W \right)^{-1} Q' R_n \left(Q' R_n \right)^{-1} W'u \\ &= R_n Q^{-1} R_n^{-1} \left[R_n^{-1} Q'^{-1} W'W Q^{-1} R_n^{-1} \right]^{-1} R_n^{-1} Q'^{-1} W'u, \end{aligned} \tag{26}$$

where $Q = \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & 0 & 0 \\ 0 & \underset{(p_2 \times p_1)}{A_1'} & I_{p_2} & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix}$. This Q is chosen so that

$$WQ^{-1} = [Z, X_1^c, X_2^c, X] \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & 0 & 0 \\ 0 & -A_1' & I_{p_2} & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix} = \left[Z, X_1^c - X_2^c A_1', X_2^c, X \right] = [Z, v_1, X_2^c, X],$$

in which I(0) and I(1) components are separated. To keep the notations concise, let $[Z, v_1] := Z^+$ and $[X_2^c, X] := X^+$. We have

$$R_n^{-1} Q'^{-1} W'W Q^{-1} R_n^{-1} = \begin{pmatrix} \frac{Z^{+\prime} Z^+}{n} & \frac{Z^{+\prime} X^+}{n^{3/2}} \\ \frac{Z^{+\prime} X^+}{n^{3/2}} & \frac{X^{+\prime} X^+}{n^2} \end{pmatrix} \Longrightarrow \begin{pmatrix} \Omega_{zz}^+ & 0 \\ 0 & \Omega_{xx}^+ \end{pmatrix} := \Omega^+. \tag{27}$$

Let the i-th column of X^+ be $\underset{(p_2+p_x)\times 1}{X_i^+} = [X_{2i}^{c\prime}, x_{i\cdot}]'$, which is a unit root vector with no cointegration relationship. Using the component-wise BN decomposition, the scalar $u_i = \underset{1\times(p+1)}{F_u(1)} \times \underset{(p+1)\times 1}{\varepsilon_i} - \triangle\tilde{\varepsilon}_{ui}$. Thus we have

$$\underset{(p_2+p_x)\times 1}{\frac{1}{n}X^{+\prime}u} = \frac{1}{n}\sum_{i=1}^{n} X_i^+ u_i' = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^+ \varepsilon_i'\right) F_u(1)' - \frac{1}{n}\sum_{i=1}^{n} X_i^+ \triangle\tilde{\varepsilon}_{ui}.$$

On the right-hand side of the above equation, $\frac{1}{n}\sum_{i=1}^{n} X_i^+ \varepsilon_i' \Longrightarrow \int B^+(r)dB_\varepsilon(r)'$, and summation by parts implies

$$\frac{1}{n}\sum_{i=1}^{n} X_i^+ \triangle\tilde{\varepsilon}_{ui} = \frac{1}{n}\sum_{i=1}^{n} u_{xi}^+ \tilde{\varepsilon}_{ui} + o_p(1) \xrightarrow{p} \underset{(p_2+p_x)\times 1}{\Delta_{+u}}$$

where Δ_{+u} is the corresponding submatrix of the one-sided long-run covariance Δ defined in (12,). Combining these results, we have

$$X^{+\prime}u/n \Longrightarrow \int B^+(r)dB_\varepsilon(r)'F_u(1)' + \Delta_{+u} := \xi_{X^+},$$

and furthermore

$$R_n^{-1}Q'^{-1}W'u = \begin{pmatrix} Z^{+\prime}u/\sqrt{n} \\ X^{+\prime}u/n \end{pmatrix} \Longrightarrow \begin{pmatrix} \xi_{Z^+} \\ \xi_{X^+} \end{pmatrix} := \xi^+. \tag{28}$$

Finally,

$$R_n Q^{-1} R_n^{-1} = \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & \frac{A_1}{\sqrt{n}} & 0 \\ 0 & 0 & I_{p_2} & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix} \to I_p. \tag{29}$$

The conclusion follows by substituting (27), (28) and (29) into (26). ■

Proof. [Proof of Theorem 3.4] The basic idea of this proof is close to that of Theorem 2.1, but there are some delicacy in the details. Let $\theta_n = \theta_n^* + R_n^{-1}v$ be a perturbation from θ_n^*, and

$$\Psi_n(v) = \|Y - \sum_{j=1}^{p} x_j(\theta_{jn}^* + R_{jn}^{-1}v_j)\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left|\beta_{jn}^* + R_{jn}^{-1}v_j\right|$$

where $R_{jn} = (R_n)_{jj}$ is the j-th diagonal element of R_n. Define

$$\begin{aligned} V_n(v) &= \Psi_n(v) - \Psi_n(0) = \left\|u - R_n^{-1}W'v\right\|_2^2 - \|u\|_2^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left(|\theta_{jn}^* + R_{jn}^{-1}v_j)| - |\theta_{jn}^*|\right) \\ &= v'R_n^{-1}W'WR_n^{-1}v - 2v'R_n^{-1}W'u + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left(|\theta_{jn}^* + R_{jn}^{-1}v_j)| - |\theta_{jn}^*|\right). \end{aligned}$$

The first term

$$v'R_n^{-1}W'WR_n^{-1}v = v'\left(R_n^{-1}Q'R_n\right)\left(R_n^{-1}Q'^{-1}W'^{-1}R_n^{-1}\right)\left(R_nQR_n^{-1}\right)v \Longrightarrow v'\Omega^+v \tag{30}$$

by (27) and (29) as we have shown in the proof of Theorem 3.2. Similarly, the second term

$$2v'R_n^{-1}W'u = 2v'\left(R_n^{-1}Q'R_n\right)\left(R_n^{-1}Q'^{-1}W'u\right) \Longrightarrow 2v'\xi^+. \tag{31}$$

We focus on the third term. Theorem 3.2 and Remark 3.3 have shown the OLS estimator $\widehat{\theta}_j^{ols} - \theta_{jn}^* = O_p\left(R_{jn}^{-1}\right)$ for each j. Given any fixed $v_j \neq 0$ and a sufficiently large n:

- For $j \in \{1, \ldots, p_z + p_1\}$, the coefficients are invariant with the sample size. If $\theta_{j0}^* \neq 0$, we have $(|\theta_{jn}^* + n^{-1/2}v_j| - |\theta_{jn}^*|) = n^{-1/2}v_j\text{sgn}(\theta_{0j}^*)$, and

  $$\lambda_n\hat{w}_j \cdot (|\alpha_j^{+*} + \frac{v_j}{\sqrt{n}}| - |\alpha_j^{+*}|) = O_p\left(\lambda_n n^{-1/2}\right) = o_p(1)$$

  If $\theta_{j0}^* = 0$, we have

  $$\lambda_n\hat{w}_j \cdot (|\theta_{jn}^* + n^{-1/2}v_j| - |\theta_{jn}^*|) = \frac{\lambda_n n^{\frac{\gamma-1}{2}}}{O_p(1)}|v_j| = O_p\left(\lambda_n n^{(\gamma-1)/2}\right) \to \infty$$

  given the rate of λ_n.

- For $j \in \{p_z + p_1 + 1, \ldots, p\}$, the coefficient $\theta_{jn}^* = \theta_{0j}^*/n^{\delta_j}$ depending on n. If $\theta_{0j}^* \neq 0$, then θ_{jn}^* dominates $n^{-1}v_j$ in the limit. We have $(|\theta_{jn}^* + n^{-1}v_j| - |\theta_{jn}^*|) = n^{-1}v_j\text{sgn}(\theta_{0j}^*) = n^{-1}v_j\text{sgn}(\theta_{0j}^*)$, and

  $$\lambda_n\hat{w}_j \cdot (|\theta_{jn}^* + \frac{v_j}{n}| - |\theta_{jn}^*|) = O_p\left(\lambda_n n^{\gamma\bar{\delta}-1}\right) = o_p(1)$$

  by the same derivation in (20). On the other hand, if $\theta_{0j}^* = 0$, then

  $$\lambda_n\hat{w}_j \cdot (|\theta_{jn}^* + n^{-1}v_j| - |\theta_{jn}^*|) = \frac{\lambda_n n^{\gamma-1}}{O_p(1)}|v_j| = O_p\left(\lambda_n n^{\gamma-1}\right) \to \infty,$$

according to the derivation in (21).

The above analysis indicates $V_n(v) \Longrightarrow V(v)$ for every fixed v, where

$$V(v) = \begin{cases} v'\Omega^+ v - 2v'\xi^+, & \text{if } v_{A^{+*c}} = 0. \\ \infty, & \text{otherwise.} \end{cases}$$

Let $\hat{v}^{(n)} = R_n^{-1}(\hat{\theta}^{alasso} - \theta_n^*)$. By the same argument about the strict convexity of $V_n(v)$ and $V(v)$, we have

$$\hat{v}^{(n)}_{A^{+*}} = R_n^{-1}(\hat{\theta}^{alasso} - \theta_n^*)_{A^{+*}} \Longrightarrow \left(\Omega^+_{A^{+*}}\right) \xi^+_{A^{+*}} \text{ and } \hat{v}^{(n)}_{A^{+*c}} \Longrightarrow 0. \tag{32}$$

The first part of the above result establishes Theorem 3.4(b), and it also implies $P(A^* \subseteq A_n) \to 1$.

For $j \notin A^{+*}$, if the event $\{j \in A_n^+\}$ occurs, then the KKT condition entails

$$\frac{2}{n} x_j^{+\prime}(y - W\hat{\theta}^{alasso}) = \frac{\lambda_n \hat{w}_j}{n} \text{ or } \frac{2}{\sqrt{n}} z_j^{+\prime}(y - W\hat{\theta}^{alasso}) = \frac{\lambda_n \hat{w}_j}{\sqrt{n}}. \tag{33}$$

We will invoke similar argument as in (24) and (25) to show the disparity of the two sides of the KKT condition, but the rates are different for the Z^+ part and the X^+ part:

- If $j \in \{1, \ldots, p_z + p_1\}$, the right-hand side of (33) is $\frac{\lambda_n \hat{w}_j}{\sqrt{n}} = \frac{\lambda_n n^{(\gamma-1)/2}}{|\sqrt{n}\hat{\theta}_j^{+ols}|^\gamma} = O_p(\lambda_n n^{(\gamma-1)/2})$, where as the left-hand side is

$$\begin{aligned} \frac{2}{\sqrt{n}} z_j^{+\prime}(y - W\hat{\theta}^{alasso}) &= \frac{2}{\sqrt{n}} z_j^{+\prime}(W\theta_n^* - W\hat{\theta}^{alasso} + u) \\ &= 2\left(\frac{z_j^{+\prime} W}{\sqrt{n}} R_n^{-1}\right) R_n(\theta_n^* - \hat{\theta}^{alasso}) + 2\frac{z_j^{+\prime} u}{\sqrt{n}} \\ &= 2\left(\frac{z_j^{+\prime} W}{\sqrt{n}} R_n^{-1}\right) \left(\hat{v}^{(n)}_{A^{+*}} + \hat{v}^{(n)}_{A^{+*c}}\right) + 2\frac{z_j^{+\prime} u}{\sqrt{n}} \\ &\Longrightarrow 2\Omega_{\cdot j} \cdot \left(\left(\Omega^+_{A^{+*}}\right)^{-1} \xi^+_{A^{+*}} + o_p(1)\right) + O_p(1), \end{aligned}$$

  which converges in distribution to a non-degenerate continuous random variable.

- If $j \in \{p_z + p_1 + 1, \ldots, p\}$, the right-hand side of the KKT condition is $\frac{\lambda_n \hat{w}_j}{n} = \frac{\lambda_n n^{\gamma-1}}{|n\hat{\theta}_j^{+ols}|^\gamma} =$

$O_p(\lambda_n n^{\gamma-1})$, whereas the left-hand side

$$\begin{aligned}\frac{2}{n}x_j^{+\prime}(y - W\hat{\theta}^{alasso}) &= \frac{2}{n}x_j^{+\prime}(W\theta_n^* - W\hat{\theta}^{alasso} + u) \\ &= 2\left(\frac{x_j^{+\prime}W}{n}R_n^{-1}\right)R_n(\theta_n^* - \hat{\theta}^{alasso}) + 2\frac{x_j^{+\prime}u}{n} \\ &= 2\left(\frac{x_j^{+\prime}W}{n}R_n^{-1}\right)\left(\hat{v}_{A^{+*}}^{(n)} + \widehat{v}_{A^{+*c}}^{(n)}\right) + 2\frac{x_j^{+\prime}u}{n} \\ &\Longrightarrow 2\Omega_{\cdot j}\cdot\left(\left(\Omega_{A^{+*}}^{+}\right)^{-1}\xi_{A^{+*}}^{+} + o_p(1)\right) + O_p\left(1\right)\end{aligned}$$

remains a non-degenerate continuous random variable asymptotically.

Given the specified rate for λ_n, for any $j \in A^{+*c}$ we have

$$P(j \in A_n^+) = P\left(\frac{2}{n}x_j^{+\prime}(y - W\hat{\theta}^{alasso}) = \frac{\lambda_n\hat{w}_j}{n} \text{ or } \frac{2}{\sqrt{n}}z_j^{+\prime}(y - W\hat{\theta}^{alasso}) = \frac{\lambda_n\hat{w}_j}{\sqrt{n}}\right) \to 0.$$

In other words, $P(A^{+*c} \subseteq A_n^{+c}) \to 1$ or equivalently $P(A_n^+ \subseteq A^{+*}) \to 1$. We therefore confirm the variable selection consistency. ■

Proof. [Proof of Corollary 3.7] For Part (a) and (b), let $\theta_n = \theta_n^* + R_n^{-1}v$ for some $v \in \mathbb{R}^p$. Define

$$V_n(v) = v'\left(R_n^{-1}W'WR_n^{-1}\right)v - 2vR_n^{-1}W'u + \lambda_n\sum_{j=1}^{p}(|\theta_{jn}^* + R_{nj}v_j| - |\theta_{jn}^*|).$$

The limiting behavior of the first and the second terms are derived in (30) and (31). Since Ω^+ is block diagonal, the sample criterion function has a nice separation in the limit,

$$V_n(v) \Longrightarrow V(v) = V_{x^+}(v_{x^+}) + V_{z^+}(v_{z^+}), \tag{34}$$

where $v_{z^+} = (v_j)_{j=1}^{p_z+p_1}$, $v_{x^+} = (v_j)_{j=p_z+p_1+1}^{p}$, and

$$\begin{aligned}V_{x^+}(v_{x^+}) &= v_{x^+}'\Omega_{xx}^{+}v_{x^+} - 2v_{x^+}'\xi_X^{+} + \lim_{n\to\infty}\frac{\lambda_n}{n}\cdot D\left(\mathbf{1}_{p_{x^+}+p_2}, v_{x^+}, \beta_0^{+*}\right) \\ V_{z^+}(v_{z^+}) &= v_{z^+}'\Omega_{zz}^{+}v_{z^+} - 2v_{z^+}'\xi_Z^{+} + \lim_{n\to\infty}\frac{\lambda_n}{\sqrt{n}}\cdot D\left(\mathbf{1}_{p_z+p_1}, v_z, \alpha_0^{+*}\right).\end{aligned}$$

In (34) the limit $V\left(v\right)$ is separable into two convex parts, implying

$$\min_{v\in\mathbb{R}^p}V(v) = \min_{v_{z^+}\in\mathbb{R}^{p_z+p_1}}V_{z^+}(v_{z^+}) + \min_{v_{x^+}\in\mathbb{R}^{p_x+p_2}}V_{x^+}(v_{x^+}).$$

Invoking the Convexity Lemma for both parts we obtain Part (a) and (b) by the same argument as in the proof of Corollary 2.3(a) and (b).

Part (c) needs more subtle investigation. Define

$$\tilde{V}_n(v) = v' \left(\tilde{R}_n^{-1} W' W \tilde{R}_n^{-1} \right) v - 2v \tilde{R}_n^{-1} W' u + \lambda_n \sum_{j=1}^{p} (|\theta_{jn}^* + \tilde{R}_{nj}^{-1} v_j| - |\theta_{jn}^*|)$$

where $\tilde{R}_n = \frac{\sqrt{n}}{\lambda_n} R_n$. Multiply n/λ_n^2 on both sides,

$$\left(\frac{n}{\lambda_n^2} \right) \tilde{V}_n(v) \tag{35}$$

$$= v' \left(R_n^{-1} W' W R_n^{-1} \right) v - 2v'_{x^+} \frac{X^{+\prime} u}{\lambda_n \sqrt{n}} - 2v'_{z^+} \frac{Z^{+\prime} u}{\lambda_n} \tag{36}$$

$$+ \frac{n}{\lambda_n} \sum_{j=1}^{p_z+p_1} (|\theta_{jn}^* + \frac{\lambda_n}{n} v_j| - |\theta_{jn}^*|) + \frac{n}{\lambda_n} \sum_{j=p_z+p_1+1}^{p} (|\theta_{jn}^* + \frac{\lambda_n}{n^{3/2}} v_j| - |\theta_{jn}^*|). \tag{37}$$

$$= v' \left(R_n^{-1} W' W R_n^{-1} \right) v + + \frac{n}{\lambda_n} \sum_{j=1}^{p_z+p_1} (|\theta_{jn}^* + \frac{\lambda_n}{n} v_j| - |\theta_{jn}^*|) + \frac{n}{\lambda_n} \sum_{j=p_z+p_1+1}^{p} (|\theta_{jn}^* + \frac{\lambda_n}{n^{3/2}} v_j| - |\theta_{jn}^*|) + o_p(1), \tag{38}$$

from the given rate condition of λ_n. Given $v_j \neq 0$ and n large enough:

- If $j \in \{1, \ldots, p_z + p_1\}$ we have

$$\frac{n}{\lambda_n} |\theta_{jn}^* + \frac{\lambda_n}{n} v_j| - |\theta_{jn}^*|) = \frac{n}{\lambda_n} D \left(\mathbf{1}_{p_z+p_1}, \frac{\lambda_n}{n} v_j, \theta_{0j}^* \right) = D \left(\mathbf{1}_{p_z+p_1}, v_j, \theta_{0j}^* \right) \tag{39}$$

  as $\theta_{jn}^* = \theta_{0j}^*$ is invariant with n.

- If $j \in \{p_z + p_1 + 1, \ldots, p\}$, the coefficient $\theta_{jn}^* = \theta_{0j}^* / n^{\delta_j}$ and may shrink faster than $\frac{\lambda_n}{n^{3/2}}$. The inequality $||a + b| - |a|| \leq |b| \, I(|b| \geq |a|) + 3 \, |b| \, I(|a| < |b|) \leq 3 \, |b|$ for any $a, b \in \mathbb{R}$ implies

$$\left| \frac{n}{\lambda_n} \left(|\theta_{jn}^* + \frac{\lambda_n}{n^{3/2}} v_j| - |\theta_{jn}^*| \right) \right| \leq 3 \frac{n}{\lambda_n} \left| \frac{\lambda_n}{n^{3/2}} v_j \right| = O \left(n^{-1/2} \right). \tag{40}$$

(39) and (40) show that the first term in the second line of (37) asymptotically dominates the second term. Thus

$$\left(\frac{n}{\lambda_n^2} \right) \tilde{V}_n(v) \Longrightarrow v'_{x^+} \Omega_{xx}^+ v_{x^+} + \left[v'_{z^+} \Sigma_{zz}^+ v_{z^+} + D \left(\mathbf{1}_{p_z+p_1}, v_{z^+}, \alpha^{+*} \right) \right].$$

The above inequality indicates that the limiting behavior of the components associated with Z^+ and the components associated with X^+ are separable. Invoking the Convexity Lemma

for both parts, we obtain Part (c). ■

Proof. [Proof of Corollary 3.9] For Part (a) when $\lambda_n = c_\lambda \in [0, \infty)$, let $\theta_n = \theta_n^* + R_n^{-1} v$ for some fixed $v \in \mathbb{R}^p$. Let

$$V_n(v) = v' \left(R_n^{-1} W' W R_n^{-1} \right) v - 2 v R_n^{-1} W' u + c_\lambda \cdot \sum_{j=1}^{p} \widehat{\sigma}_j (|\theta_{jn}^* + R_{jn} v_j| - |\theta_{jn}^*|).$$

For $v_j \neq 0$ and a sufficiently large n:

- if $j \in \{1, \ldots, p_z\}$,

$$\widehat{\sigma}_j \left(|\theta_{jn}^* + \frac{v_j}{\sqrt{n}}| - |\theta_{jn}^*| \right) = D \left(\widehat{\sigma}_j, \frac{v_j}{\sqrt{n}}, \theta_{0j}^* \right) = D \left(O_p(1), O\left(\frac{1}{\sqrt{n}} \right), \theta_{0j}^* \right) \xrightarrow{p} 0$$

  as the index is associated with the stationary variable Z and therefore $\widehat{\sigma}_j = O_p(1)$;

- if $j \in \{p_z + 1, \ldots, p_z + p_1\}$,

$$\begin{aligned} \widehat{\sigma}_j \left(|\theta_{jn}^* + \frac{v_j}{\sqrt{n}}| - |\theta_{jn}^*| \right) = D \left(\widehat{\sigma}_j, \frac{v_j}{\sqrt{n}}, \theta_{0j}^* \right) &= D \left(\frac{\widehat{\sigma}_j}{\sqrt{n}}, v_j, \theta_{0j}^* \right) \\ \Longrightarrow D \left(d_j, v_j, \theta_{0j}^* \right) &= O_p(1) \end{aligned}$$

  as the index is associated with unit root processes in X_1^c and therefore $\frac{\widehat{\sigma}_j}{\sqrt{n}} \Longrightarrow d_j$;

- if $j \in \{p_z + p_1 + 1, \ldots, p\}$,

$$\begin{aligned} \widehat{\sigma}_j \left(|\theta_{jn}^* + \frac{v_j}{n}| - |\theta_{jn}^*| \right) = D \left(\widehat{\sigma}_j, \frac{v_j}{n}, \theta_{0j}^* \right) &= D \left(\frac{\widehat{\sigma}_j}{\sqrt{n}}, \frac{v_j}{\sqrt{n}}, \theta_{0j}^* \right) \\ &= D \left(O_p(1), O\left(\frac{1}{\sqrt{n}} \right), \theta_{0j}^* \right) \xrightarrow{p} 0 \end{aligned}$$

  as $\frac{\widehat{\sigma}_j}{\sqrt{n}} \Longrightarrow d_j = O_p(1)$ for these regressors.

The above analysis of the third term implies

$$V_n(v) \Longrightarrow V(v) = v' \left(R_n^{-1} W' W R_n^{-1} \right) v - 2 v R_n^{-1} W' u + c_\lambda \sum_{j=p_z+1}^{p_z+p_1} D \left(d_j, v_j, \theta_{0j}^* \right),$$

and the conclusion follows.

For Part (b), let $\tilde{R}_n = R_n / \lambda_n$ and $\theta_n = \theta_n^* + \tilde{R}_n^{-1} v$ for some $v \in \mathbb{R}^p$. Define

$$\tilde{V}_n(v) = v' \left(\tilde{R}_n^{-1} W' W \tilde{R}_n^{-1} \right) v - 2 v \tilde{R}_n^{-1} W' u + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j (|\theta_{jn}^* + \tilde{R}_{jn}^{-1} v_j| - |\theta_{jn}^*|).$$

Multiply $1/\lambda_n^2$ on both sides,

$$\begin{aligned}\frac{\tilde{V}_n(v)}{\lambda_n^2} &= v'\left(R_n^{-1}W'WR_n^{-1}\right)v - 2v'_{x^+}\frac{X^{+\prime}u}{\lambda_n n} - 2v'_{z^+}\frac{Z^{+\prime}u}{\lambda_n\sqrt{n}} + \frac{1}{\lambda_n}\sum_{j=1}^{p}\widehat{\sigma}_j(|\theta^*_{jn}+\tilde{R}^{-1}_{jn}v_j| - |\theta^*_{jn}|) \\ &= v'\left(R_n^{-1}W'WR_n^{-1}\right)v + \frac{1}{\lambda_n}\sum_{j=1}^{p}\widehat{\sigma}_j(|\theta^*_{jn}+\tilde{R}^{-1}_{jn}v_j| - |\theta^*_{jn}|) + o_p(1).\end{aligned}$$

from the given rate condition of λ_n. Again we study the last term. By the same reasoning as in Part (a), for $v_j \neq 0$ and a sufficiently large n we have:

- if $j \in \{1,\ldots,p_z\}$,

$$\frac{1}{\lambda_n}\widehat{\sigma}_j\left(|\theta^*_{jn}+\frac{\lambda_n}{\sqrt{n}}v_j| - |\theta^*_{jn}|\right) = \frac{1}{\lambda_n}D\left(\widehat{\sigma}_j, \frac{\lambda_n}{\sqrt{n}}v_j, \theta^*_{0j}\right) = D\left(\widehat{\sigma}_j, \frac{v_j}{\sqrt{n}}, \theta^*_{0j}\right) \xrightarrow{p} 0;$$

- if $j \in \{p_z+1,\ldots,p_z+p_1\}$,

$$\begin{aligned}\frac{1}{\lambda_n}\widehat{\sigma}_j\left(|\theta^*_{jn}+\frac{\lambda_n}{\sqrt{n}}v_j| - |\theta^*_{jn}|\right) &= \frac{1}{\lambda_n}D\left(\widehat{\sigma}_j, \frac{\lambda_n}{\sqrt{n}}v_j, \theta^*_{0j}\right) = D\left(\frac{\widehat{\sigma}_j}{\sqrt{n}}, v_j, \theta^*_{0j}\right) \\ &= D\left(d_j, v_j, \theta^*_{0j}\right) = O_p\left(1\right);\end{aligned}$$

- if $j \in \{p_z+p_1+1,\ldots,p\}$, the rate condition $\lambda_n/n^{(1-\bar{\delta})\wedge 0.5} \to 0$ makes sure that $\theta^*_{jn} = \theta^*_{0j}/n^{\delta_j}$ dominates $\frac{\lambda_n}{n}$ so that

$$\frac{1}{\lambda_n}\widehat{\sigma}_j\left(|\theta^*_{jn}+\frac{\lambda_n}{n}v_j| - |\theta^*_{jn}|\right) = D\left(\widehat{\sigma}_j, \frac{v_j}{n}, \theta^*_{0j}\right) = D\left(\frac{\widehat{\sigma}_j}{\sqrt{n}}, \frac{v_j}{\sqrt{n}}, \theta^*_{0j}\right) \xrightarrow{p} 0.$$

We obtain $\frac{\tilde{V}_n(v)}{\lambda_n^2} \Longrightarrow v'\Omega^+v + \sum_{j=p_z+p_1+1}^{p_z+p_1} D\left(d_j, v_j, \theta^*_{0j}\right)$ and the conclusion follows. ■

B Additional Simulations

B.1 More DGPs

In this Section, we include two more DGPs to examine the forecasting performance and variable screening in the presence of autoregression.

DGP 3 (Unit-root autoregression). Motivated by Caner (2013) proposing to treat the unit root test as a model selection problem by regressing Δy_{i+1} on lags of y_i, we come

up with the following DGP that extends their setting by including stationary regressors. The dependent variable is generated from a unit-root autoregression $y_{i+1} = y_i + \beta_{1n}^* x_i + \beta_{2n}^* x_{i-1} + \sum_{j=1}^{6} \alpha_j^* z_{ij} + u_i$, where x_i is a random walk. The stationary regressors $Z_i = (z_{ij})_{j=1}^{6}$ follow a stationary VAR(2) borrowed from Koo et al. (2016, Section 5.1) [7]. We include lag terms of y_i as regressors. In the predictive regression, we use $\Delta y_{i+1} = y_{t+1} - y_t$ as the dependent variable, and the regression equation is

$$\Delta y_{i+1} = \phi_{1n}^* y_i + \phi_{2n}^* y_{i-1} + \beta_{1n}^* x_i + \beta_{2n}^* x_{i-1} + \sum_{j=1}^{6} \alpha_j^* z_{ij} + u_{i+1}$$

where $(\phi^*,\, \beta^*,\, \alpha^*) = \left(0,\, 0,\, \frac{1}{\sqrt{n}},\, \frac{1}{\sqrt{n}},\, 1,\, 1,\, 1,\, 0,\, 0,\, 0\right)$. Notice that y_i and y_{i-1} are inactive cointegrated regressors and this DGP also employs mixed roots and cointegration.

DGP 4 (Stationary autoregression). In addition to including lags of y_i, it is also a common practice to include lags of predictors in predictive regressions, for example Medeiros and Mendes (2016). We propose the following DGP in which a stationary autoregression generates the dependent variable

$$y_{i+1} = \gamma^* + \rho^* y_i + \sum_{j=1}^{2} \phi_{jn}^* x_{ij}^c + \beta_{1n}^* x_i + \beta_{2n}^* x_{i-1} + \sum_{j=1}^{3} \left(\alpha_{j1}^* z_{ij} + \alpha_{j2}^* z_{i-1,j}\right) + u_{i+1}$$

where $\gamma^* = 0.3$, $(\rho^*, \phi^*, \beta^*, \alpha_1^*, \alpha_2^*, \alpha_3^*) = \left(0.4,\, 0.75,\, -0.75,\, \frac{1.5}{\sqrt{n}},\, 0.6,\, 0.4,\, 0.8,\, 0,\, 0,\, 0\right)$. The cointegrated x_{i1}^c and x_{i2}^c are generated by $x_{i2}^c = x_{i1}^c - \mu_i$ where x_{i1}^c is a random walk and μ_i is a stationary AR(1) process with AR(1) coefficient 0.4. x_i follows a random walk. z_{i1},z_{i2} and z_{i3} are three independent AR(1) processes with AR(1) coefficients 0.5,0.2 and 0.2, respectively.

The results summarized in Table 6 and 7 are similar to that in DGP 2, which demonstrates the merits of adaptive LASSO in the presence of autoregression.

B.2 Standardized LASSO

We determine c_λ^{alasso} and c_λ^{plasso} as in Section 4 and let $c_\lambda^{plasso} = c_\lambda^{slasso}$. The results are summarized in Table 8 and **??**. The CR_2 of standardized LASSO is much higher than that of plain LASSO, which is consistent to what the asymptotic theory suggests.

[7] For completeness, the VAR(2) is $Z_i = A_{z1} Z_{i-1} + A_{z2} Z_{i-2} + v_t$, where

$$A_{z1} = \begin{pmatrix} 0 & 0 & 0 & 0.4 & 0 & 0 \\ 0.29 & 0.12 & 0 & 0 & 1.31 & 0.04 \\ 1.25 & -0.24 & 0 & 0 & -0.21 & 0.04 \\ 0.03 & 1.16 & 0 & 0 & 0.07 & 0.01 \\ 0.27 & -0.07 & 0 & 0 & 0.08 & 1.25 \\ 0 & 0 & 0.4 & 0 & 0 & 0 \end{pmatrix} \text{ and } A_{z2} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ -0.28 & -0.07 & 0 & 0 & -0.35 & -0.02 \\ -0.26 & 0.24 & 0 & 0 & 0.19 & -0.05 \\ -0.02 & -0.16 & 0 & 0 & -0.07 & 0.01 \\ -0.23 & 0.03 & 0 & 0 & -0.13 & -0.31 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Table 6: Mean Prediction Squared Error (MPSE)

	n	Oracle	OLS	alasso	plasso	slasso
DGP 3	40	1.2041	1.6302	1.4117	1.4454	**1.3456**
	80	1.1000	1.2244	1.1518	1.1808	**1.1411**
	120	1.0703	1.1815	**1.1025**	1.1260	1.1093
	200	0.9686	0.9962	**0.9879**	0.9943	0.9918
	400	0.9971	1.0131	**0.9985**	1.0028	1.0023
DGP 4	40	1.3636	1.5981	1.5684	**1.5368**	1.5839
	80	1.1238	1.2192	**1.1947**	1.2057	1.2207
	120	1.0682	1.1014	**1.0944**	1.1003	1.0980
	200	0.9654	1.0068	**0.9875**	1.0002	1.0055
	400	1.0484	1.0590	**1.0505**	1.0539	1.0587

Table 7: Variable Screening

		CR			CR_1			CR_2		
	n	alasso	plasso	slasso	alasso	plasso	slasso	alasso	plasso	slasso
DGP 3	40	0.6499	0.5755	**0.6843**	0.8396	0.8732	**0.9014**	0.4602	0.2778	**0.4672**
	80	**0.6772**	0.5642	0.6650	0.8444	0.9142	**0.9244**	**0.5100**	0.2142	0.4056
	120	**0.6758**	0.5513	0.6556	0.8462	0.9276	**0.9280**	**0.5054**	0.1750	0.3832
	200	**0.6903**	0.5529	0.6539	0.8410	**0.9450**	0.9378	**0.5396**	0.1608	0.3700
	400	**0.6917**	0.5518	0.6421	0.8330	**0.9596**	0.9550	**0.5504**	0.1440	0.3292
DGP 4	40	**0.7955**	0.7246	0.6453	0.9460	0.9759	**0.9941**	**0.5320**	0.2850	0.0347
	80	**0.8423**	0.7235	0.6495	0.9676	0.9897	**0.9966**	**0.6230**	0.2578	0.0420
	120	**0.8595**	0.7235	0.6475	0.9670	0.9900	**0.9931**	**0.6715**	0.2570	0.0425
	200	**0.8858**	0.7249	0.6494	0.9726	0.9921	**0.9953**	**0.7340**	0.2573	0.0440
	400	**0.9110**	0.7199	0.6530	0.9697	0.9934	**0.9947**	**0.8083**	0.2413	0.0550

Note: Bold numbers are for the best performance.

B.3 Cross-validation For Each Sample Size

In this section, we include the simulation results in which tuning parameters are determined separately for each sample size n. In detail, for each sample size $n = 40$, 80, 120, 200 and 400, we run an exploratory simulation for 100 times for each method that needs a tuning parameter. In each replication, we use the 10-fold cross-validation to obtain $c_\lambda^{(1)}, \ldots, c_\lambda^{(100)}$. Then we fix $c_\lambda = \text{median}\left(c_\lambda^{(1)}, \ldots, c_\lambda^{(100)}\right)$ in the full-scale 1000-time simulation for the corresponding sample size.

The results are summarized in Table 10 and 11.

Table 8: Mean Prediction Squared Error (MPSE)

	n	Oracle	OLS	alasso	plasso	slasso
DGP 1	40	1.2103	1.4715	1.3583	**1.1882**	1.2918
	80	1.1840	1.3185	1.2565	**1.2066**	1.2497
	120	1.0926	1.1698	1.1531	**1.1330**	1.1888
	200	1.0865	1.1580	1.1353	**1.1279**	1.1705
	400	0.9646	0.9836	**0.9797**	0.9837	1.0357
DGP 2	40	1.1513	1.3277	**1.2409**	1.2452	1.2219
	80	1.0694	1.1413	**1.1203**	1.1255	1.1499
	120	1.0595	1.0794	1.0816	**1.0799**	1.1025
	200	1.0789	1.1322	**1.1028**	1.1196	1.1287
	400	1.0010	1.0205	**1.0133**	1.0175	1.0623

Note: Bold numbers are for the best performance among all the feasible estimators.

Table 9: Variable Screening

			CR			CR_1			CR_2	
	n	alasso	plasso	slasso	alasso	plasso	slasso	alasso	plasso	slasso
DGP 1	40	0.5884	**0.6375**	0.6248	**0.7565**	0.6478	0.4783	0.4203	0.6273	**0.7713**
	80	0.6590	0.6741	**0.6929**	**0.8348**	0.8273	0.6183	0.4833	0.5210	**0.7675**
	120	0.7119	0.6806	**0.7196**	0.8765	**0.9025**	0.6880	0.5473	0.4588	**0.7513**
	200	**0.7628**	0.6731	0.7564	0.9368	**0.9668**	0.7763	0.5888	0.3795	**0.7365**
	400	**0.8381**	0.6383	0.8024	0.9830	**0.9940**	0.8693	0.6933	0.2825	**0.7355**
DGP 2	40	**0.6776**	0.5805	0.6095	0.8265	**0.9160**	0.8060	**0.5288**	0.2450	0.4130
	80	**0.7576**	0.5953	0.6394	0.9205	**0.9788**	0.8628	**0.5948**	0.2118	0.4160
	120	**0.7928**	0.5910	0.6531	0.9615	**0.9928**	0.8893	**0.6240**	0.1893	0.4170
	200	**0.8359**	0.5909	0.6811	0.9873	**0.9990**	0.9193	**0.6845**	0.1828	0.4430
	400	**0.8798**	0.5844	0.6903	0.9990	**1.0000**	0.9415	**0.7605**	0.1688	0.4390

Note: Bold numbers are for the best performance.

Table 10: Mean Prediction Squared Error (MPSE)

	n	Oracle	OLS	alasso	plasso	slasso
DGP 1	40	1.1877	1.4438	1.2719	**1.1757**	1.2230
	80	1.1848	1.3102	1.2774	**1.2408**	1.2617
	120	1.0976	1.1882	1.1819	1.1546	**1.1538**
	200	1.0339	1.0998	1.0862	1.0726	**1.0689**
	400	0.9969	1.0064	1.0055	**1.0029**	1.0077
DGP 2	40	1.1724	1.3307	1.2201	1.2117	**1.1995**
	80	1.0566	1.1821	1.1239	**1.1440**	1.1507
	120	1.0684	1.1652	**1.1216**	1.1413	1.1505
	200	0.9853	1.0038	**0.9975**	1.0010	1.0030
	400	1.0312	1.0583	**1.0476**	1.0529	1.0547
DGP 3	40	1.2789	1.6580	**1.4524**	1.5658	1.4699
	80	1.2047	1.4169	**1.3019**	1.3467	1.3172
	120	1.0876	1.1277	1.1247	1.1292	**1.1141**
	200	1.0847	1.1326	**1.1140**	1.1234	1.1204
	400	0.9786	1.0052	**0.9924**	0.9957	**0.9924**
DGP 4	40	1.3061	1.5686	1.4959	**1.4768**	1.5228
	80	1.1456	1.2225	**1.2077**	1.2248	1.2293
	120	1.1745	1.2071	**1.1942**	1.1988	1.2057
	200	1.0360	1.0627	**1.0431**	1.0486	1.0558
	400	1.0769	1.0865	**1.0825**	1.0835	1.0876

Note: Bold numbers are for the best performance among all the feasible estimators.

Table 11: Variable Screening

		CR			CR_1			CR_2		
	n	alasso	plasso	slasso	alasso	plasso	slasso	alasso	plasso	slasso
DGP 1	40	**0.6341**	0.6303	0.6319	0.4238	**0.6888**	0.5983	**0.8445**	0.5718	0.6655
	80	**0.7214**	0.6779	0.6915	0.6490	**0.8003**	0.7938	**0.7938**	0.5555	0.5893
	120	**0.7520**	0.6666	0.6956	0.8190	0.9118	**0.9120**	**0.6850**	0.4215	0.4793
	200	**0.7870**	0.6878	0.7053	0.9263	**0.9683**	0.9603	**0.6478**	0.4073	0.4503
	400	**0.8274**	0.6531	0.6726	0.9838	**0.9950**	0.9928	**0.6710**	0.3113	0.3525
DGP 2	40	**0.7191**	0.6448	0.6369	0.6273	**0.7068**	0.6625	**0.8110**	0.5828	0.6113
	80	**0.8039**	0.6809	0.6506	0.8543	**0.9023**	0.7335	**0.7535**	0.4595	0.5678
	120	**0.8209**	0.6690	0.6455	0.9580	**0.9850**	0.9693	**0.6838**	0.3530	0.3218
	200	**0.8199**	0.6264	0.6141	0.9930	**0.9983**	0.9973	**0.6468**	0.2545	0.2310
	400	**0.8404**	0.6254	0.5964	0.9975	**1.0000**	**1.0000**	**0.6833**	0.2508	0.1928
DGP 3	40	**0.7240**	0.5644	0.6847	0.7860	0.8850	**0.9036**	**0.6620**	0.2438	0.4658
	80	**0.7154**	0.5452	0.6691	0.8144	0.9122	**0.9162**	**0.6164**	0.1782	0.4220
	120	**0.7049**	0.5454	0.6741	0.8282	**0.9316**	0.9248	**0.5816**	0.1592	0.4234
	200	**0.7184**	0.5508	0.6766	0.8250	**0.9458**	0.9384	**0.6118**	0.1558	0.4148
	400	**0.7023**	0.5601	0.7013	0.8328	**0.9504**	0.9442	**0.5718**	0.1698	0.4584
DGP 4	40	**0.7747**	0.6959	0.6635	0.9557	0.9826	**0.9851**	**0.4580**	0.1943	0.1008
	80	**0.8416**	0.7297	0.6673	0.9684	0.9899	**0.9930**	**0.6198**	0.2745	0.0972
	120	**0.8663**	0.7513	0.6649	0.9707	0.9881	**0.9904**	**0.6835**	0.3368	0.0952
	200	**0.8973**	0.7427	0.6619	0.9727	0.9924	**0.9934**	**0.7653**	0.3058	0.0818
	400	**0.8920**	0.7326	0.6538	0.9697	0.9921	**0.9937**	**0.7560**	0.2785	0.0590

Note: Bold numbers are for the best performance.