On LASSO for Predictive Regression

Ji Hyung Lee* Zhentao Shi† Zhan Gao‡

Abstract

A typical predictive regression employs a multitude of potential regressors with various degrees of persistence while their signal strength in explaining the dependent variable is often low. Variable selection in such context is of great importance. In this paper, we explore the pitfalls and possibilities of the LASSO methods in this predictive regression framework with mixed degrees of persistence. With the presence of stationary, unit root and cointegrated predictors, we show that the adaptive LASSO maintains the consistent variable selection and the oracle property due to its penalty scheme that accommodates the system of regressors. On the contrary, conventional LASSO does not have this desirable feature as the penalty is imposed according to the marginal behavior of each individual regressor. We demonstrate this theoretical property via extensive Monte Carlo simulations, and evaluate its empirical performance for short- and long-horizon stock return predictability.

*Department of Economics, University of Illinois. 1407 W. Gregory Dr., 214 David Kinley Hall, Urbana, IL 61801, United States. Email: jihyung@illinois.edu.

†Department of Economics, the Chinese University of Hong Kong. 928 Esther Lee Building, the Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong SAR, China. Email: zhentao.shi@cuhk.edu.hk. Shi acknowledges the financial support from the Hong Kong Research Grants Council Early Career Scheme No.24614817.

‡Department of Economics, University of Southern California, Kaprielian Hall, 3620 South Vermont Avenue, Los Angeles, CA 90089, USA. Email: zhangao@usc.edu. The authors thank Mehmet Caner, Zongwu Cai, Yoosoon Chang, Changjin Kim, Zhipeng Liao, Tassos Magdalinos, Joon Park, Hashem Pesaran, Peter Phillips, Kevin Song, Jing Tao, Keli Xu, Jun Yu and seminar participants at Kansas, Indianda, Purdue, UBC, UW, Duke, KAEA, IPDC, AMES and IAAE conferences for helpful comments. We also thank Bonsoo Koo for sharing the data for empirical applications. All remaining errors are ours.

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 the severe test size distortion in the presence of highly persistent predictors coupled with regression endogeneity. When persistence and endogeneity loom large, the conventional inferential tools designed for stationary data can be misleading. Another major challenge in predictive regression is the low signal-to-noise ratio (SNR). Nontrivial regression coefficients representing predictability are hard to detect, because they are often contaminated by large estimation error. A large econometric literature is devoted to procedures for valid inference and improved prediction.

Machine learning arises in the era of big data. The advancement in machine learning techniques—driven by unprecedented abundance of data sources across many disciplines—offers opportunities for economic data analysis. Shrinkage methods, in particular, are increasingly popular for the econometric inference and prediction in view of its variable selection and regularization properties. The least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) has received much attention in the past two decades.

This paper studies LASSO methods in predictive regressions. The intrinsic low SNR in predictive regressions naturally calls for variable selection. 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 selection, since many of these variables *ex post* demonstrate little or no predictability. LASSO-type shrinkage methods are therefore attractive in the predictive regressions as they enable researchers to select the pertinent predictors and to exclude the irrelevant ones. However, time series regressors in predictive regressions have heterogeneous degrees of persistence. Some may exhibit short memory (e.g., Treasury bill), whereas others are highly persistent (e.g., most of financial/macro predictors). Moreover, a multitude of persistent predictors can be cointegrated. For example, the dividend payout ratio (DP ratio) is essentially a cointegrating residual between the dividend and price, and 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 yet been systematically investigated.

The performance of LASSO procedure crucially hinges on the choice of the tuning parameter. In this paper, we keep an agnostic view about the type of the regressors, and we examine whether a single tuning parameter can cope with the heterogeneous regressors. In particular, we explore the plain LASSO (Plasso, henceforth; Tibshirani, 1996), the standardized LASSO (Slasso. See below for the definition) and the adaptive LASSO (Alasso; 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 defies the conventional wisdom of the variable screening property of Plasso and Slasso.

The main contribution of this paper is unveiling Alasso's capability of handling a system of heterogeneous regressors in predictive regressions. With a proper choice of the tuning parameter, Alasso achieves the oracle property (Fan and Li, 2001), which implies optimal rate of convergence and consistent variable selection, despite the presence of a mixture of heterogeneous regressors. To our knowledge, this paper is the first to demonstrate these desirable properties of Alasso in nonstationary time series context. On the contrary, Plasso and Slasso impose penalty of the coefficients only according to the marginal behavior of each individual regressor. Thus in the cointegration system they fail to simultaneously achieve consistent coefficient estimation and variable screening.

We consider a linear process of the time series innovations, which encompasses a general ARMA structure arising in many practical applications, for example, the long-horizon return prediction. To focus on the distinctive feature of nonstationary time series, we adopt a simple asymptotic framework in which the number of regressors p being fixed and the number of time periods n passes to infinity. Our exploration in this paper paves a stepping stone toward the automated variable selection in a high-dimensional predictive regression with heterogeneously persistent regressors.

In Monte Carlo simulations, we examine the finite sample performance of Alasso in comparison with Plasso and Slasso, assessing their mean squared prediction errors and the variable selection success rates. These methods are further evaluated in a real data application of the widely used Welch and Goyal (2008) data to predict S&P500 stock return using 12 predictors. Alasso is shown to be robust in various estimation windows and prediction horizons.

Literature Review Since Tibshirani (1996)'s original paper of LASSO and Chen et al. (2001)'s basis pursuit, a variety of important extensions of LASSO have been proposed; for example Alasso (Zou, 2006) and elastic net (Zou and Hastie, 2005), to name a few. In econometrics, Caner (2009) and Caner and Zhang (2014) employ the LASSO-type procedures in GMM contexts. Belloni and Chernozhukov (2011), Belloni et al. (2012), Belloni et al. (2014), Belloni et al. (2018) develop new methodologies and uniform statistical theories for estimation and inference in a variety of microeconometric settings.

In comparison with the vast literature of LASSO in cross sectional regressions, shrinkage methods are less explored in time series models. Medeiros and Mendes (2016) study Alasso method in high-dimensional stationary time series models. Kock and Callot (2015) discuss LASSO in a vector autoregression (VAR) system. In the time series forecasting context, Inoue and Kilian (2008) 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 several machine learning estimators. Even fewer are 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 Alasso. In a vector error correction model (VECM), Liao and Phillips (2015) use Alasso for cointegration rank selection.

In predictive regressions, Kostakis et al. (2014), Lee (2016) and Phillips and Lee (2013, 2016) provide inferential procedures in the presence of multiple predictors with various degrees of persistence. Xu (2018) studies variable selection and inference with possible cointegration among the I(1) predictors. Koo et al. (2016) recently investigate the property of Plasso in predictive regressions. The last two papers are closely related to ours, while we investigate Alasso's variable selection in predictive regressions under mixed degrees of persistence.

Notation We use standard notations. We define $\|\cdot\|_1$ and $\|\cdot\|_2$ as the usual vector l_1-norm and l_2-norm respectively. $\Longrightarrow$ and $\xrightarrow{p}$ represent convergence in distribution and convergence in probability, respectively. All limit theory assumes $n \to \infty$ so we oftentimes omit this condition. $\asymp$ means of the same asymptotic order, and $\sim$ signifies "being distributed as" either exactly or asymptotically, depending on the contexts. We use $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$ for any $a, b \in \mathbb{R}$. The symbols $O(1)$ and $o(1)$ ($O_p(1)$ and $o_p(1)$) are (stochastically) asymptotically bounded or negligible quantities. For a generic set M, let $|M|$ be the cardinality of the set. For a generic vector $\theta = (\theta_j)_{j=1}^p$ with $p \geq |M|$, let $\theta_M = (\theta_j)_{j \in M}$ be the subvector of θ associated with the index set M.

The rest of the paper is organized as follows. Section 2 introduces the unit root regressors into a simple LASSO framework to clarify the idea. This model is substantially generalized in Section 3 to include I(0), I(1) and cointegrated regressors, and the asymptotic properties of Alasso are developed and compared with those of Plasso and Slasso. The theoretical results are explored through a set of empirically relevant simulation designs in Section 4. Last but not least, we examine the stock return regressions via these LASSO methods in Section 5.

2 LASSO Theory with Unit Roots

In this section, we study LASSO with p unit root regressors. To fix ideas, we investigate the asymptotic behavior of Alasso, Plasso, and Slasso under a simple nonstationary regression model. This model helps us understand the technical issues in LASSO arising from nonstationary predictors under the conventional choices of tuning parameters. Section 3 will generalize the model to include I(0), I(1) and cointegrated predictors altogether.

Assume the dependent variable y_i is generated from a linear model

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

where n is the sample size. The $p \times 1$ true coefficient $\beta_n^{*} = (\beta_{jn}^{*} = \beta_j^{0*}/n^{\delta_j})_{j=1}^{p}$, where $\beta_j^{0*} \in \mathbb{R}$ is a fixed constant independent of the sample size, and $\delta_j \in [0, 1)$. Without loss of generality, define $\delta_j = 0$ if $\beta_j^{0*} = 0$; thus β_{jn}^{*} varies with the sample size if $\beta_j^{0*} \neq 0$ and $\delta_j \in (0, 1)$. 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 to model the weak SNR in predictive regressions (Phillips and Lee, 2013; Timmermann and Zhu, 2017).[1] The $1 \times p$ regressor vector $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})$, with

$$x_{i\cdot} = x_{(i-1)\cdot} + e_{i\cdot} = \sum_{k=1}^{i} e_{k\cdot}, \tag{2}$$

where the innovation $e_{k\cdot} = (e_{k1}, ..., e_{kp})$. For simplicity, we assume the initial value $e_{0\cdot} = \mathbf{0}_p'$, and the following iid assumption on the innovations.

Assumption 2.1 *The vector of innovation $e_{i\cdot}$ and u_i follow the joint distribution*

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

where Σ is positive-definite.

The regression equation (1) can be equivalently written as

$$y = \sum_{j=1}^{p} x_j \beta_{jn}^{*} + u = X\beta_n^{*} + u, \tag{3}$$

where $y = (y_1, \ldots, y_n)'$ is the $n \times 1$ response vector, $u = (u_1, \ldots, u_n)'$, $x_j = (x_{j1}, \ldots, x_{jn})'$, and $X = [x_1, \ldots, x_p]$ is the $n \times p$ predictor matrix. This pure I(1) regressor model in (3) is a direct extension of the common predictive regression application with a single unit root predictor (e.g., DP

[1] Unlike Hirano and Wright (2017), we exclude $\delta_j = 1$ (Pitman drift) to eliminate the effect of nuisance parameters in the limit. Also see Remark 3.6 below for the related discussion and clarification.

ratio). The mixed roots case in Section 3 will be more relevant in practice when multiple predictors are present.

The literature focuses on the non-standard statistical inference caused by persistent regressors and weak signal. The asymptotic theory is usually confined to a reasonable number of candidate predictors, but not too many. Following the literature of predictive regressions, we consider the asymptotic framework in which p is fixed and the sample size $n \to \infty$.[2] This simple asymptotic framework allows us to focus on the contrast between the standard LASSO literature and the predictive regression environment that involves nonstatioary regressors.

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

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

whose asymptotic behavior is well understood (Phillips, 1987). Assumption 2.1 implies the following functional central limit theorem

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor nr \rfloor} \begin{pmatrix} e'_{k\cdot} \\ u_k \end{pmatrix} \Longrightarrow \begin{pmatrix} B_x(r) \\ B_u(r) \end{pmatrix} \equiv BM\left(\Sigma\right). \tag{4}$$

To represent the asymptotic distribution of the OLS estimator, let $u_i^+ = u_i - \Sigma_{eu}\Sigma_{ee}^{-1}e'_{i\cdot}$ then $\frac{1}{\sqrt{n}}\sum_{i=1}^{\lfloor nr \rfloor} u_i^+ \Longrightarrow B_{u^+}(r)$. By definition, $\mathrm{cov}\left(e_{i\cdot}, u_i^+\right) = \mathbf{0}'_p$ so that

$$\frac{X'u}{n} \Longrightarrow \zeta := \int_0^1 B_x(r)dB_{u^+}(r) + \Sigma_{eu}\Sigma_{ee}^{-1}\int_0^1 B_x(r)dB_x(r)',$$

which is the sum of a (mixed) normal random vector and a scaled nonstandard random vector. The OLS limit distribution is

$$n\left(\widehat{\beta}^{ols} - \beta_n^*\right) = \left(\frac{X'X}{n^2}\right)^{-1}\frac{X'u}{n} \Longrightarrow \Omega^{-1}\zeta, \tag{5}$$

where $\Omega := \int_0^1 B_x(r)B_x(r)'dr$.

In addition to the low SNR in predictive regressions, some true coefficients β_j^{0*} in (3) could be exactly zero so that the associated predictors are redundant in the regression. Let $M^* = \{j : \beta_j^{0*} \neq 0\}$ be the set of relevant regressors, and $M^{*c} = \{1, \ldots, p\} \backslash M^*$ be the set of redundant ones. If we had prior knowledge about M^*, ideally we would estimate the unknown parameter by OLS in the set M^*:

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

where $p^* = |M^*|$ is the number of relevant regressors. We call this the *oracle* estimator, and (5) implies that its asymptotic distribution is

$$n\left(\widehat{\beta}^{oracle} - \beta_n^*\right) \Longrightarrow \Omega_{M^*}^{-1}\zeta_{M^*},$$

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

The oracle information about M^* is infeasible in practice. It is well known that machine learning methods such as LASSO and its variants have built-in mechanism for variable screening. Next, we

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

study LASSO's asymptotic behavior in the predictive regression with these pure unit root regressors.

2.1 Adaptive LASSO with Unit Root Regressors

Alasso for (1) is defined as

$$\widehat{\beta}^{Alasso} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \widehat{\tau}_j |\beta_j|, \tag{6}$$

where the weight is $\widehat{\tau}_j = |\widehat{\beta}_j^{\text{init}}|^{-\gamma}$ for some initial estimator $\widehat{\beta}^{\text{init}}$, and λ_n and γ are the two tuning parameters. In practice, γ is often fixed at either 1 or 2, and λ_n is selected as the main tuning parameters. Throughout this paper, we discuss the case of a fixed $\gamma \geq 1$ and $\widehat{\beta}^{\text{init}} = \widehat{\beta}^{ols}$. Alasso enjoys the oracle property in regressions with weakly dependent regressors (Medeiros and Mendes, 2016).

The following Theorem 2.1 confirms that Alasso maintains the oracle property in the regression with unit root regressors. For a generic index set M, let $\widehat{M}_n = \{j : \widehat{\beta}_j^{Alasso} \neq 0\}$ be Alasso's estimated *active set*, and let $\overline{\delta} = \max_{j \leq p} \delta_j$.

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: $P\left(\widehat{M}_n = M^*\right) \to 1$.

(b) Asymptotic distribution of $\widehat{\beta}_{M^*}^{Alasso}$*:* $n(\widehat{\beta}^{Alasso} - \beta_n^*)_{M^*} \Longrightarrow \Omega_{M^*}^{-1}\zeta_{M^*}$.

Theorem 2.1 (a) shows that the estimated active set $\widehat{M}_n$ coincides with the true active set M^* with probability approaching to one. (b) indicates that Alasso's asymptotic distribution in the true active set is the same as the oracle estimator with known M^*.

In this nonstationary regression, the adaptiveness is maintained through the proper choice of $\widehat{\tau}_j = |\widehat{\beta}_j^{\text{ols}}|^{-\gamma}$. On the one hand, when the true coefficients are non-zero, $\widehat{\tau}_j$ delivers a penalty of a negligible order $\lambda_n n^{\gamma\overline{\delta}-1} = o(1)$, recovering the OLS limit theory. On the other hand, if the true coefficients are zero, $\widehat{\tau}_j$ imposes a heavier penalty of the order$\lambda_n n^{\gamma-1} \to \infty$, thereby achieving consistent variable selection. We generalize the intuition provided in (Zou, 2006, Remark 2) under the deterministic design to the setting with nonstationary regressors.

Remark 2.2 *In Theorem 2.1, we observe some interconnected rate conditions between* λ_n*,* $\overline{\delta}$ *and* γ*. To achieve the oracle asymptotic distribution in* M^**, 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 in* M^{*c}*. Consider the usual formulation of the tuning parameter* $\lambda_n = c_\lambda b_n n^{\frac{1}{2}}$*. If we substitute this* λ_n *into the restriction, we have*

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

When $\overline{\delta} = 1/2$ *(a balancing order for I(0)-I(1) regression) and* $\gamma = 1$*, the restriction is further simplified to* $b_n + \frac{1}{b_n n^{1/2}} \to 0$*. Any slowly shrinking sequence such as* $b_n = (\log\log n)^{-1}$*, which is commonly imposed in the adaptive LASSO literature in cross section regressions, satisfies this rate condition.*

Given the positive results about adaptive LASSO with unit root regressors, we continue to study Plasso and one of its simple variant, which we call Slasso.

2.2 Plain LASSO with Unit Roots

LASSO produces a parsimonious model as it tends to select the relevant variables. In this paper, Plasso is defined as

$$\widehat{\beta}^{Plasso} = \arg\min_{\beta\in\mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 . \tag{7}$$

Plasso is a special case of the penalized estimation in (6) with the weights $\widehat{\tau}_j$, $j = 1, \ldots, p$, fixed at unity. The following results characterize its asymptotic behavior with various choices of λ_n under unit root regressors. For exposition, we define a function D as

$$D(s, v, \beta) = \sum_{j=1}^{\dim(\beta)} s_j \left(v_j \mathrm{sgn}(\beta_j) I(\beta_j \neq 0) + |v_j| I(\beta_j = 0)\right)$$

for three generic vectors s, v, and β 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 $n(\widehat{\beta}^{Plasso} - \beta_n^*) \Longrightarrow \Omega^{-1}\zeta$.

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

$$n(\widehat{\beta}^{Plasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{v'\Omega v - 2v'\zeta + c_\lambda D(\mathbf{1}_p, v, \beta^{0*})\right\}.$$

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

$$\frac{n^2}{\lambda_n}(\widehat{\beta}^{Plasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{v'\Omega v + D(\mathbf{1}_p, v, \beta^{0*})\right\}$$

Remark 2.4 *Corollary 2.3 echoes, but is different from, Zou (2006, Section 2). Under the unit root regressor framework, (a) implies that the conventional tuning parameter $\lambda_n \asymp \sqrt{n}$ is too small for variable selection. With such a choice, the asymptotic distribution is the same as that of OLS. (b) shows that the additional term $c_\lambda D(\mathbf{1}_p, v, \beta^{0*})$ will affect the limit distribution when λ_n is enlarged to the order of n. In this case, the asymptotic distribution is given as the minimizer of a criterion function involving Ω, ζ and the true coefficient β^{0*}. Similar to LASSO in the cross sectional setting, there is no guarantee for consistent variable selection.*[3] *When we enlarge λ_n even further, (c) indicates that the convergence rate is slowed down to $\widehat{\beta}^{Plasso} - \beta_n^* = O_p(\lambda_n/n^2)$. Notice that the asymptotic distribution* $\arg\min_v \left\{v'\Omega v + D(\mathbf{1}_p, v, \beta^{0*})\right\}$ *is non-degenerate due to the randomness of Ω. This is in sharp contrast to Zou (2006)'s Lemma 3, which shows a degenerate asymptotic distribution as the Gram matrix there converges to a positive-definite non-random matrix. Regarding the implication for the variable selection, similar to Zou's insight, we conclude that the plain LASSO's variable selection is in general inconsistent with unit root regressors.*

[3] In this paper, we call it the *variable screening effect* if some estimated coefficient are shrunk to exactly zero (whether or not the true coefficients are zeros), instead of the *variable selection effect* (which means that an estimator asymptotically distinguishes those true zero coefficients from the non-zero ones).

2.3 Standardized LASSO with Unit Roots

Given that the usual choice of λ_n being too small for Plasso to conduct variable screening, one may consider a popular alternative. Plasso 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 in (7) so Plasso estimator does not change proportionally to $\widehat{\beta}_j^{Plasso}/c$. To keep the estimation scale-invariant to the choice of arbitrary unit of x_j, researchers often scale-standardize LASSO as

$$\widehat{\beta}^{Slasso} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j \, |\beta_j| \, . \tag{8}$$

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_j. In this paper, we call (8) the *standardized LASSO*. Such standardization is the default option for LASSO in many statistical packages, for example the `R` package `glmnet`.

Slasso is another special case of (6) with $\widehat{\tau}_j = \widehat{\sigma}_j$. When such a scale standardization is carried out with stationary and weakly dependent regressors, each $\widehat{\sigma}_j^2$ converges in probability to its population variance, which is finite. Standardization does not alter the convergence rate of the estimator. In contrast, when x_j has a unit root, from (4) we have

$$\frac{\widehat{\sigma}_j}{\sqrt{n}} = \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2} \Longrightarrow d_j = \sqrt{\int B_{x_j}^2(r)dr - \left(\int B_{x_j}(r)\, dr\right)^2} \tag{9}$$

so that $\widehat{\sigma}_j = O_p(\sqrt{n})$. As a result, it imposes a much heavier penalty on the associated coefficients than that of the stationary time series case. Adopting a standard argument for LASSO as in Knight and Fu (2000) and Zou (2006), we have the following asymptotic distribution for $\widehat{\beta}^{Slasso}$. Let $d = (d_1, \ldots, d_p)'$.

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

(a) If $\lambda_n/\sqrt{n} \to 0$, then $n(\widehat{\beta}^{Slasso} - \beta_n^*) \Longrightarrow \Omega^{-1}\zeta$.

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

$$n(\widehat{\beta}^{Slasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{v'\Omega v - 2v'\zeta + c_\lambda D(d, v, \beta^{0*})\right\}.$$

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

$$\frac{n^{3/2}}{\lambda_n}(\widehat{\beta}^{Slasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{v'\Omega v + D(d, v, \beta^{0*})\right\}.$$

Remark 2.6 *In Corollary 2.5 (a), the range of the tuning parameter to restore the OLS asymptotic distribution is much smaller than that in Corollary 2.3 (a), due to the magnitude of d. When we increase λ_n as in (b), the term $D(d, v, \beta^{0*})$ will generate the variable screening effect under the usual choice of tuning parameter $\lambda_n \asymp \sqrt{n}$. In contrast, its counterpart $D(\mathbf{1}_p, v, \beta^{0*})$ emerges in Corollary 2.3 when $\lambda_n \asymp n$. While the first argument of $D(\cdot, v, \beta^{0*})$ in Plasso is the unit vector $\mathbf{1}_p$, in Slasso it is replaced by the random vector d, which introduces an extra source of randomness in the variable screening. Again, in general Slasso has no mechanism to achieve consistent variable* *selection. An even large λ_n as in (c) only slows down the rate of convergence but does not help with variable selection.*

To summarize this section, in the regression with unit root predictors, Alasso retains the oracle property under the usual choice of the tuning parameter. For Plasso to screen variables, we need to lift the tuning parameter up to the order of n. For Slasso, although $\lambda_n \asymp \sqrt{n}$ is sufficient for variable screening, the sample variance of the nonstationary regressors carries 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 that exhibit 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 extend the model in Section 2 to allow I(0) and I(1) regressors with possible cointegration among those I(1) regressors. The LASSO theory in this section will provide a general guidance for multivariate predictive regressions in practice.

3.1 Model

In this model we introduce three types of predictors. Assume that the dependent variable y_i is generated from the linear model

$$y_i = \sum_{l=1}^{p_z} z_{il}\alpha_l^* + \sum_{l=1}^{p_c} x_{il}^c\phi_l^* + \sum_{l=1}^{p_x} x_{il}\beta_l^* + u_i = z_{i\cdot}\alpha^* + x_{i\cdot}^c\phi^* + x_{i\cdot}\beta^* + u_i, \tag{10}$$

for $i = 1, \ldots, n$, where $z_{i\cdot} = (z_{i1}, ..., z_{ip_z})$, $x_{i\cdot}^c = (x_{i1}^c, ..., x_{ip_c}^c)$, and $x_{i\cdot} = (x_{i1}, ..., x_{ip_x})$ represent the stationary, cointegrated, and unit root regressors, respectively, and $p = p_z + p_c + p_x$. Each $x_l = (x_{1l}, ..., x_{nl})'$, $l = 1, \ldots, p_x$, is a unit root predictor (initialized at zeros for simplicity) as that in Section 2. Regarding the cointegration system, let p_1 be the cointegration rank, so that $p_2 = p_c - p_1$ is the number of unit roots in the cointegration system. A triangular representation[4] governs the $1 \times p_c$ predictor $x_{i\cdot}^c$ as

$$\underset{p_1\times p_c}{A}\, x_{i\cdot}^{c\prime} = x_{1i\cdot}^{c\prime} - \underset{p_1\times p_2}{A_1}\, x_{2i\cdot}^{c\prime} = v_{1i\cdot}', \tag{11}$$
$$\Delta x_{2i\cdot}^c = v_{2i\cdot},$$

where $A = [I_{p_1}, -A_1]$ and $x_{i\cdot}^c = (x_{1i\cdot}^c, x_{2i\cdot}^c)$.

Given the data of sample size n, the empirical model is

$$y = Z\alpha^* + X^c\phi^* + X\beta^* + u = Z\alpha^* + X_1^c\phi_1^* + X_2^c\phi_2^* + X\beta^* + u, \tag{12}$$

where X_1 is the $n \times p_1$ matrix that stacks $x_{1i\cdot}$, $i = 1, \ldots, n$, and similarly X_2 is associated with $x_{2i\cdot}$. By stacking $v_{1i\cdot}, i = 1, \ldots, n$, the cointegration residual in (11), into an $n \times p_1$ matrix V_1, the above

[4]The triangular representation (Phillips, 1991) is a convenient and general representation of a cointegrated system. Xu (2018) recently used this structure in predictive regressions.

expression can be represented as

$$y = Z\alpha^* + V_1\phi_1^* + X_2^c\widetilde{\phi}_2^* + X\beta^* + u, \tag{13}$$

where $\widetilde{\phi}_2^* = \widetilde{\phi}_1^* + \phi_2^*$ and $\widetilde{\phi}_1^* = A_1'\phi_1^*$. The coefficients ϕ_1^* and $\widetilde{\phi}_2^*$ signify the effects of the cointegration residuals and the unit roots in the cointegration system, respectively. To ensure this model's validity, we keep $\widetilde{\phi}_2^*$ and β^* sufficiently small to maintain the stationarity of y. As an asymptotic mechanism, we use the same local-to-zero modeling as in Section 2 for the coefficients of the I(1) regressors, so ϕ_2^* offsets the non-zero full rank component $A_1'\phi_1^*$, leading to a small $\widetilde{\phi}_2^*$. We explicitly define $\alpha^* = \alpha^{0*}$ and $\phi_1^* = \phi_1^{0*}$ as two coefficients independent of the sample size, as they are associated with the stationary components Z and V_1 in (13). On the other hand, similar to the modeling approach in Section 2, for the unit root variables we define $\widetilde{\phi}_2^* = \widetilde{\phi}_{2n}^* = (\widetilde{\phi}_{2,l}^{0*}/n^{\delta_l})_{l=1}^{p_2}$ and $\beta^* = \beta_n^* = (\beta_l^{0*}/n^{\delta_l})_{l=1}^{p_x}$ as two local-to-zero coefficients, in which $\widetilde{\phi}_{2,l}^{0*}$ and β_l^{0*} are invariant with n while n^{δ_l} determines the rate of shrinking to zero when the sample size increases.

3.2 OLS theory with mixed roots

We assume a linear process for the innovation and cointegrating residual vectors. In contrast to the unrealistic iid assumption in Section 2, the linear process assumption is fairly general, including many practical dependent processes (stationary AR and MA processes, for example) as special cases. Let $v_{i\cdot} = (v_{1i\cdot}, v_{2i\cdot})$.

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

$$\underset{(p+1)\times 1}{\xi_i} := (z_{i\cdot}, v_{i\cdot}, e_{i\cdot}, u_i)' = F(L)\varepsilon_i = \sum_{k=0}^{\infty} F_k\varepsilon_{i-k},$$

$$\underset{(p+1)\times 1}{\varepsilon_i} = \left(\varepsilon_{i\cdot}^{(z)}, \varepsilon_{i\cdot}^{(v)}, \varepsilon_{i\cdot}^{(e)}, \varepsilon_{i}^{(u)}\right)' \sim \text{iid}\left(\mathbf{0}_{p+1}, \underset{(p+1)\times(p+1)}{\Sigma_\varepsilon} = \begin{pmatrix} \Sigma_{zz} & \Sigma_{zv} & \Sigma_{ze} & \mathbf{0} \\ \Sigma_{zv}' & \Sigma_{vv} & \Sigma_{ve} & \mathbf{0} \\ \Sigma_{ze}' & \Sigma_{ve}' & \Sigma_{ee} & \Sigma_{eu} \\ \mathbf{0}' & \mathbf{0}' & \Sigma_{eu}' & \Sigma_{uu} \end{pmatrix}\right),$$

where $F_0 = I_{p+1}$, $\sum_{k=0}^{\infty} k\|F_k\| < \infty$, $F(x) = \sum_{k=0}^{\infty} F_k x^k$ and $F(1) = \sum_{k=0}^{\infty} F_k > 0$.

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 the innovation of stationary or the cointegrated predictors.*

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

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

Define

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

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

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

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,

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

The Beveridge-Nelson decomposition and weak convergence to stochastic integral lead to

$$Z^{+\prime}u/\sqrt{n} \Longrightarrow \zeta_{z^+} \sim N\left(0, \Sigma_{uu}\Omega_{zz}^+\right) \tag{14}$$

$$X^{+\prime}u/n \Longrightarrow \zeta_{x^+} \sim \int B_{x^+}(r)dB_\varepsilon(r)'F_u(1)' + \Delta_{+u} \tag{15}$$

where the one-sided long-run covariance matrix $\Delta_{+u} = \sum_{h=0}^{\infty} \mathbb{E}\left(\widetilde{u}_i u_{i-h}\right)$ with $\widetilde{u}_i = (v_{2i\cdot}', e_{i\cdot})'$. Jointly these two components converge in distribution to a stable law

$$\begin{pmatrix} Z^{+\prime}u/\sqrt{n} \\ X^{+\prime}u/n \end{pmatrix} \Longrightarrow \zeta^+, \tag{16}$$

although it is difficult to express ζ^+ analytically.

We first study OLS, the initial estimator for Alasso. For notational convenience, define the predictor matrix

$$\underset{n\times p}{W} = \left[\underset{n\times p_z}{Z}, \underset{n\times p_c}{X^c}, \underset{n\times p_x}{X}\right] = \left[\underset{n\times p_z}{Z}, \underset{n\times p_1}{X_1^c}, \underset{n\times p_2}{X_2^c}, \underset{n\times p_x}{X}\right],$$

and the parameter $\underset{p\times 1}{\theta_n^*} = \left(\alpha^{0*\prime}, \phi_n^{*\prime}, \beta_n^{*\prime}\right)'$ where $\phi_n^* = \left(\phi_1^{0*\prime}, \phi_{2n}^{*\prime}\right)'$, so that

$$y = W\theta_n^* + u.$$

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 (10) satisfies Assumption 3.1, then*

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

where $\Omega^+ := \begin{pmatrix} \Omega_{zz}^+ & \mathbf{0} \\ \mathbf{0} & \Omega_{xx}^+ \end{pmatrix}$.

Remark 3.3 *Theorem 3.2 and the definition of ζ_{x^+} implies 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 $\widehat{\theta}_n^{ols} - \theta_n^* = O_p(\mathrm{diag}(R_n^{-1}))$. The rates of convergence of the OLS estimator helps to study the asymptotic behavior of Alasso. In principle, either the fully-modified OLS (Phillips and Hansen, 1990) or the canonical cointegrating regression estimator (Park, 1992) can be an initial estimator as well because the convergence rates are same. Note that we keep an agnostic view about the identities of the stationary, unit root and cointegrated regressors, so Theorem 3.2 cannot be used for statistical inference as we do not know which coefficients converges at the $n^{-1/2}$-rate and which at the n^{-1}-rate.*

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

3.3 Adaptive LASSO with mixed roots

Similarly to Section 2.1, we define Alasso under the system of (12) as

$$\widehat{\theta}^{Alasso} = \arg\min_{\theta \in \mathbb{R}^p} \|y - W\theta\|_2^2 + \lambda_n \sum_{j=1}^{p} \widehat{\tau}_j |\theta_j|, \tag{17}$$

where $\widehat{\tau}_j = |\widehat{\theta}_j^{ols}|^{-\gamma}$. The following theorem confirms that Alasso maintains the oracle property in the presence of stationary, unit root and cointegrated regressors. Let $\mathcal{I}_0 = \{1, \cdots, p_z\}$, $\mathcal{C}_1 = \{p_z + 1, \cdots, p_z + p_1\}$, $\mathcal{C}_2 = \{p_z + p_1 + 1, \ldots, p_z + p_c\}$ and $\mathcal{I}_1 = \{p_z + p_c + 1, \ldots, p\}$, which are the index sets associated with $z_{i\cdot}$, $x_{1i\cdot}^c$, $x_{2i\cdot}^c$ and $x_{i\cdot}$, respectively. We keep using $\widehat{M}_n = \{j : \widehat{\theta}_j^{Alasso} \neq 0\}$ as the estimated active set and $M^* = \{j : \theta_j^{0*} \neq 0\}$ as the true active set, where

$$\theta_j^{0*} = \begin{cases} \alpha_l^{0*}, & l = j,\ j \in \mathcal{I}_0 \\ \phi_{1,l}^{0*}, & l = j - p_z, j \in \mathcal{C}_1 \\ \widetilde{\phi}_{2,l}^{0*} 1\{\widetilde{\phi}_{1,l}^{0*} = 0\} + (\lim_{n\to\infty} \widetilde{\phi}_{2,ln}^{*} - \widetilde{\phi}_{1,l}^{0*}) 1\{\widetilde{\phi}_{1,l}^{0*} \neq 0\}, & l = j - p_z - p_1, j \in \mathcal{C}_2 \\ \beta_l^{0*}, & l = j - p_z - p_c, j \in \mathcal{I}_1, \end{cases}$$

The definition of θ_j^{0*} for $j \in \mathcal{C}_2$ ensures that it reflects the effect of $\phi_{2n}^* = \widetilde{\phi}_{2n}^* - \widetilde{\phi}_1^{0*}$. Let $\bar{\delta}$ be the biggest δ_l among $\left(\beta_{l0}^*/n^{\delta_l}\right)_{l=1}^{p_x}$ and $\left(\widetilde{\phi}_{2,l}^{0*}/n^{\delta_l}\right)_{l=1}^{p_2}$.

Theorem 3.4 *Suppose that the linear model (10) satisfies Assumption 3.1. If the tuning parameter λ_n is chosen such that $\lambda_n \to \infty$ and*

$$\frac{\lambda_n}{n^{(1/2)\wedge(1-\gamma\cdot\bar{\delta})}} + \frac{1}{\lambda_n n^{(\gamma-1)/2}} \to 0, \tag{18}$$

then, we have

(a) Variable selection consistency: $P\left(\widehat{M}_n = M^*\right) \to 1$.

(b) Asymptotic distribution of $\widehat{\theta}_{M^*}^{Alasso}$: $[R_n(\widehat{\theta}^{Alasso} - \theta_n^*)]_{M^*} \Longrightarrow \left(\Omega_{M^*}^+\right)^{-1} \zeta_{M^*}^+$.

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 (18) is reasonable because, (i) we choose $\gamma \geq 1$ in practice to prevent Alasso 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. A single rate of the tuning parameter λ_n in Theorem 3.4 is able to cope with all the stationary, unit root and cointegrated regressors.*

Remark 3.6 *Another related results in the literature are uniformly valid inference and forecasting after the LASSO model selection, see Belloni et al. (2015, 2018) or Hirano and Wright (2017), for example. These papers allow the so-called model selection error 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.*

Caner and Knight (2013) and Kock (2016) studied Alasso's rate adaptiveness in the pure autoregressive setting with iid error processes. In their cases, the potential nonstationary regressor is the first-order lagged dependent variable while other regressors are stationary. Therefore, the components of different convergence rates are known in advance. 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 regression 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 predictors with unknown orders of integration, we may not be able to sort them into different persistence categories in predictive regressions. Theorem 3.4 provides a simple condition leading to a desirable oracle property without requiring prior knowledge on the persistence of multivariate regressors.

Given the intensive study of LASSO-type estimation in recent years, Theorem 3.4 still comes as a surprise. With the cointegration system in the predictors, our result shows that Alasso not only adapts to the *marginal* behavior of single regressors, but also automatically adapts to the behavior of a *system* of regressors. Such adaptiveness to a cointegration system has not been explored in the literature, to the best of our knowledge. In contrast, the weight $\widehat{\tau}_j$ in (17) is a constant for Plasso or exploits merely the marginal variation of x_j for Slasso. As will be formally discussed in the following section, such weighting schemes are unable to tackle the cointegrated regressors. Since the cointegrated regressors are individually unit root processes, only when classified into a system can we form a linear combination of these unit root processes to produce a stationary time series. Alasso implicitly assigns appropriate level of penalty inside the cointegration system without knowing the identity of these variables.

3.4 Conventional LASSO with mixed roots

We now study the asymptotic theory of Plasso

$$\widehat{\theta}^{Plasso} = \arg\min_{\theta \in \mathbb{R}^p} \|y - W\theta\|_2^2 + \lambda_n \|\theta\|_1, \tag{19}$$

under the system of (12).

Corollary 3.7 *Suppose the linear model (10) satisfies Assumption 3.1.*

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

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

$$R_n(\widehat{\theta}^{Plasso} - \theta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega^+ v - 2v'\zeta^+ + c_\lambda \sum_{j \in \mathcal{I}_0 \cup \mathcal{C}_1} D\left(1, v_j, \theta_j^{0*}\right) \right\}$$

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

$$\begin{aligned} \frac{n}{\lambda_n}(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{I}_0 \cup \mathcal{C}_1} &\Longrightarrow \arg\min_v \left\{ v'\Omega_{zz}^+ v + \sum_{j \in \mathcal{I}_0 \cup \mathcal{C}_1} D\left(1, v_j, \theta_j^{0*}\right) \right\} \\ n(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{C}_2 \cup \mathcal{I}_1} &\Longrightarrow \left(\Omega_{xx}^+\right)^{-1} \zeta_{x^+}. \end{aligned}$$

Remark 3.8 *In Corollary 3.7(a), the tuning parameter is too small and the limit distribution of Plasso is equivalent to that of OLS; there is no variable screening effect. When the tuning parameter is raised to the case of (b), the term $D\left(1, v_j, \theta_j^{0*}\right)$ strikes variable screening. However, variable screening takes place only in the stationary part, while 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 Ω^+ is a block diagonal matrix, we can write down the marginal distributions*

$$\begin{aligned} \sqrt{n}(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{I}_0 \cup \mathcal{C}_1} &\Longrightarrow \arg\min_v \left\{ v'\Omega_{zz}^+ v - 2v'\zeta_{z^+} + c_\lambda \sum_{j \in \mathcal{I}_0 \cup \mathcal{C}_1} D\left(1, v_j, \theta_j^{0*}\right) \right\} \\ n(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{C}_2 \cup \mathcal{I}_1} &\Longrightarrow \left(\Omega_{xx}^+\right)^{-1} \zeta_{x^+}. \end{aligned}$$

It is clear that no variable screening is executed in the nonstatioary component. If we further increase the tuning parameter as in the case of (c), then the convergence rate of the I(0) part is dragged down by the large λ_n but still no variable screening in the I(1) part. Moreover, it implies inconsistency of $\widehat{\theta}^{Plasso}_{\mathcal{I}_0 \cup \mathcal{C}_1}$ if $\lambda_n/n \to c_\lambda \in (0,\infty)$.

The result in Corollary 3.7 reveals a major drawback of Plasso in the mixed root model. Since it has one single rate for the tuning parameter, it is not adaptive to these various types of predictors. There is no way for Plasso to achieve variable screening and consistent estimation simultaneously in the I(0) and I(1) predictors.

Let us now turn to Slasso, defined as

$$\widehat{\theta}^{Slasso} = \arg\min_{\theta \in \mathbb{R}^p} \|y - W\theta\|_2^2 + \lambda_n \sum_{j=1}^p \widehat{\sigma}_j \left|\theta_j\right|. \tag{20}$$

The stationary regressors are accompanied with $\widehat{\sigma}_j = O_p(1)$ for $j \in \mathcal{I}_0 \cup \mathcal{C}_1$, while the nonstationary regressors are coupled with $\widehat{\sigma}_j = O_p(\sqrt{n})$ for $j \in \mathcal{C}_2 \cup \mathcal{I}_1$. According to the following results on the asymptotic properties, Slasso suffers similar problems as Plasso does.

Corollary 3.9 *Suppose the linear model (10) satisfies Assumption 3.1.*

(a) If $\lambda_n \to 0$, then $R_n(\widehat{\theta}^{Slasso} - \theta_n^*) \Longrightarrow \left(\Omega^+\right)^{-1} \zeta^+$.

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

$$R_n(\widehat{\theta}^{Slasso} - \theta^*) \Longrightarrow \arg\min_v \left\{ v'\Omega^+ v - 2v'\zeta^+ + c_\lambda \sum_{j \in \mathcal{C}_1} D\left(d_j, v_j, \theta_j^{0*}\right) \right\}.$$

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

$$\begin{aligned} \frac{\sqrt{n}}{\lambda_n}(\widehat{\theta}^{Slasso} - \theta^*)_{\mathcal{C}_1} &\Longrightarrow \arg\min_v \left\{ v'\Omega^+_{\mathcal{C}_1} v + \sum_{j \in \mathcal{C}_1} D\left(d_j, v_j, \theta_j^{0*}\right) \right\} \\ (R_n)_{\mathcal{I}_0 \cup \mathcal{C}_2 \cup \mathcal{I}_1}(\widehat{\theta}^{Slasso} - \theta_n^*)_{\mathcal{I}_0 \cup \mathcal{C}_2 \cup \mathcal{I}_1} &\Longrightarrow (\Omega^+_{\mathcal{I}_0 \cup \mathcal{C}_2 \cup \mathcal{I}_1})^{-1} \zeta^+_{\mathcal{I}_0 \cup \mathcal{C}_2 \cup \mathcal{I}_1}, \end{aligned}$$

where $\Omega^+_M = [\Omega^+_{jj'}]_{j,j' \in M}$ for a generic index set M, and $(R_n)_M$ is similarly defined.

Remark 3.10 *The tuning parameter in Corollary 3.9(a) is $\sqrt{n}$-order smaller than that in Corollary 3.7(a) to produce the same asymptotic distribution as OLS. The distinction arises from the coefficients in the set $\mathcal{C}_1$ associated with the cointegration residuals. Their corresponding penalty terms have the multipliers $\widehat{\sigma}_j = O_p(\sqrt{n})$, instead of the desirable $O_p(1)$. In other words, the penalty level is too heavy for these parameters. The overwhelming penalty level incurs variable screening effect as soon as $\lambda_n = c_\lambda \in (0, \infty)$, as in (b). Moreover, (c) implies that for the consistency of $\widehat{\phi}_1$ the tuning parameter λ_n must be small enough in the sense $\lambda_n/\sqrt{n} \to 0$. In this case, no variable screening is possible for all other coefficients in the index set $\mathcal{I}_0 \cup \mathcal{C}_2 \cup \mathcal{I}_1$. If we further raise λ_n to $\lambda_n/\sqrt{n} \to c_\lambda \in (0, \infty)$, those $\widehat{\phi}_1$ will be inconsistent for ϕ_1^0. Again, simultaneously for all components there is no way for Slasso to achieve consistent parameter estimation and variable screening.*

To sum up this section, in the general model with various types of regressors, Alasso 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 mixed root model in this section. In contrast, Plasso using the single tuning parameter does not adapt to the different order of magnitudes of the stationary and nonstationary regressors. Slasso suffers from overwhelming penalties for those coefficients associated with the cointegrating residuals. Keeping an agnostic view about the persistence of the regressors, we recommend Alasso in the multivariate predictive regression with potential mixed roots.

4 Monte Carlo Simulation

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

4.1 Simulation Design

To echo the settings in Section 2 and 3, we consider two data generating processes (DGPs): one with pure unit root regressors and the other with mixed roots and cointegration. Appendix B.1 will include two more DGPs using lagged dependent variables as regressors.

DGP 1 (Pure unit roots). 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 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 $e_{i2} = e_{i1} - \nu_{1i}$ and $e_{i4} = e_{i3} - \nu_{2i}$, where ν_{1i} and ν_{2i} are independent AR(1) processes with the AR(1) coefficient 0.2. x_{i1} and x_{i2} are independent random walks as those in DGP 1.

As we develop our theory with regressors of fixed dimension, the OLS is a natural benchmark. Another benchmark is the oracle OLS, in which the true model is known and thus employed. This oracle OLS estimator is of course infeasible in reality. The sample sizes in our exercise range n = 40, 80, 120, 200, 400 and 800. For each simulation setting, we generate the data with 1000 burn-in periods and run 1000 replications for each sample size n.

In the simulations as well as the empirical application later, we do not penalize the intercept when implementing the shrinkage estimators. Each shrinkage estimator relies on its tuning parameter λ_n, which is the appropriate rate multiplied by a constant c_λ. We use 10-fold cross validation, where the sample is temporally ordered and then partitioned into 10 consecutive blocks, to guide the choice of c_λ. To make sure that the tuning parameter changes according to the rate we specify in the asymptotic theory, we set $n = 200$ and run an exploratory simulation for 100 times for each method that requires 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 replications. To set the tuning parameters in other sample sizes when $n \neq 200$, we multiply the constant c_λ that we have learned from $n = 200$ by the rates that our asymptotic theory suggests; that is, we multiply c_λ by $\sqrt{n}$ for Plasso or Slasso, and by $\sqrt{n}/\log(\log(n))$ for Alasso.

4.2 Performance Comparison

Table 1 reports the out-of-sample prediction accuracy in terms of the mean prediction squared error (MPSE), $E\left[(y_{T+1} - \widehat{y}_{T+1})^2\right]$. By the simulation design, the variance of the idiosyncratic error is 1, which is the unpredictable variation. Table 2 summarizes the variable screening performance. Recall that the set of relevant regressors as $M^* = \left\{j \in \{1,\ldots,p\} : \theta_j^* \neq 0\right\}$ and the estimated active set is $\widehat{M} = \left\{j \in \{1,\ldots,p\} : \widehat{\theta}_j \neq 0\right\}$. We define two *success rates* for variable screening:

$$SR_1 = \frac{1}{|M^*|} E\left[\left|\left\{j : j \in \widehat{M},\, j \in M^*\right\}\right|\right], \quad SR_2 = \frac{1}{|M^{*c}|} E\left[\left|\left\{j : j \in M^{*c},\, j \in \widehat{M}^c\right\}\right|\right].$$

Here SR_1 is the percentage of the correct selection in the active set, whereas SR_2 is the percentage of correct elimination of the zero coefficients. We also report the overall success rate of classification into zero coefficients and non-zero coefficients

$$SR = \frac{1}{p} E\left[\left| \{ j \in \{1, \ldots, p\} : I(\theta_j^* = 0) = I(\widehat{\theta}_j = 0) \right| \right],$$

where $I(\cdot)$ is the indicator function. These expectations are computed by the average in the 1000 simulation 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
	800	0.9855	0.9927	**0.9879**	0.9897	0.9896
DGP 2	40	1.2626	1.4900	1.3793	**1.3638**	1.4190
	80	1.1029	1.2156	**1.1903**	1.2055	1.2100
	120	1.0984	1.1640	**1.1463**	1.1565	1.1584
	200	1.1017	1.1523	**1.1241**	1.1386	1.1388
	400	0.9569	0.9722	**0.9606**	0.9662	0.9675
	800	1.0102	1.0172	**1.0125**	1.0145	1.0171

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

Table 2: Variable Screening

		SR			SR_1			SR_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
	800	**0.8874**	0.6040	0.6883	0.9983	**0.9998**	0.9993	**0.7765**	0.2083	0.3773
DGP 2	40	**0.6845**	0.5953	0.5541	0.8018	0.8933	**0.9525**	**0.5673**	0.2973	0.1558
	80	**0.7719**	0.6175	0.5773	0.9148	0.9835	**0.9895**	**0.6290**	0.2515	0.1650
	120	**0.8103**	0.6045	0.5796	0.9580	0.9943	**0.9963**	**0.6625**	0.2148	0.1630
	200	**0.8378**	0.5915	0.5834	0.9880	0.9990	**0.9993**	**0.6875**	0.1840	0.1675
	400	**0.8661**	0.5840	0.5959	0.9980	**1.0000**	**1.0000**	**0.7343**	0.1680	0.1918
	800	**0.8846**	0.5728	0.6111	**1.0000**	**1.0000**	**1.0000**	**0.7693**	0.1455	0.2223

Note: Bold numbers are for the best performance.

According to Table 1, the shrinkage methods always outperform OLS. Plasso and Slasso are slightly better in forecasting than adaptive LASSO in DGP 1 of pure unit root regressors. As the sample size increases to $n = 800$, Alasso eventually surpasses the conventional LASSO estimators. In DGP 2 of mixed regressors, the setting is more complicated for the conventional LASSO to navigate. Except for the smallest sample size $n = 40$, Alasso outperforms the competitors in terms of MPSE by a nontrivial edge.

The forecasting performance is associated with the variable screen effect, which is reported in Table 2. In DGP 1, Plasso and Slasso are not far behind Alasso. The advantage of Alasso in variable screening becomes more prominent in DGP 2 of mixed roots and cointegration. Alasso outperforms the competitors in SR, and the gain is large in SR_2. As sample size increases, all SR, SR_1 and SR_2 of Alasso increase in both DGPs, which supports the variable screening consistency of Alasso. The asymptotic theory suggests $\lambda_n \sim \sqrt{n}$ is too small for Plasso to eliminate 0 coefficients corresponding to I(1) regressors. Plasso and Slasso achieve high SR_1 at the cost of low SR_2, i.e. they tend to keep more active regressors even some of the selected ones are redundant. As the sample size increases, the difference in SR_1 among methods becomes negligible.[5]

In view of the results in Table 1, the clear effectiveness of Alasso's variable screening capability helps with forecast accuracy in the context of predictive regressions where many included regressors actually exhibit no predictive power.

5 Empirical Illustration

This section presents some empirical illustration on stock return predictability with the updated Welch and Goyal (2008) dataset. We focus on the potential improvement in terms of MPSE and some meaningful variable selection results using Alasso.

5.1 Data

As in Koo et al. (2016), we use the monthly Welch and Goyal (2008) data from January 1945 to December 2012. The dependent variable is *excess return*, defined as the difference between the continuously compounded return on the S&P 500 index and the three-month Treasury bill rate. The data include 12 financial and macroeconomic variables as predictors.[6]

Over the whole sample period, the excess return has an estimated AR(1) coefficient of 0.1494, indicating moderate persistence, similar to the long-term return of government bonds (ltr), stock variance (svar) and inflation (infl). The other 9 predictors exhibits high persistence, with AR(1) coefficients greater than 0.95. The mixture of stationary predictors and persistent predictors fits the mixed roots environment studied in this paper.

[5] In Table 2 Slasso has the lowest variable elimination success rate SR_2, whereas in asymptotics it imposes heavier penalty on coefficients of I(1) regressors than Plasso does due to the presence of $\widehat{\tau}_j = \widehat{\sigma}_j = O_p\left(\sqrt{n}\right)$ 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 on the in-sample MSE and hence favors c_λ achieving lower MPSE and adjusts c_λ in finite sample. 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 $c_\lambda^{Slasso} = c_\lambda^{Plasso}$, SR_2 of Slasso would become much higher.

[6] The predictors include the Dividend Price Ratio (dp), the difference between the log of the 12-month moving sum dividends and the log of the S&P 500 index; Dividend Yield (dy), the difference between the log of the 12-month moving sum dividends and the log of lagged the S&P 500 index; Earning Price Ratio (ep), the difference between the log of the 12-month moving sum earnings and the log of the S&P 500 index; Term Spread (tms), the difference between the long-term government bond yield and the Treasury Bill rate; Default Yield Spread (dfy), the difference between Moody's BAA and AAA-rated corporate bond yields; Default Return Spread (dfr), the difference between the returns of long-term corporate bonds and long-term government bonds; Book-to-Market Ratio (bm), the ratio of the book value to market value for the Dow Jones Industrial Average; Treasury Bill Rates (tbl), the 3-month Treasury Bill rates; Long-Term Return (ltr), the rate of returns of long-term government bonds; Net Equity Expansion (ntis), 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; Stock Variance (svar), the sum of the squared daily returns on the S&P 500 index; Inflation (infl), the log growth of the Consumer Price Index (all urban consumers).

Table 3: Mean Prediction Squared Error (MPSE)

	MPSE					Percentage relative to OLS				
h	OLS	RWwD	Alasso	Plasso	Slasso	OLS	RWwD	Alasso	Plasso	Slasso
					10-year rolling window					
$\frac{1}{12}$	0.00209	0.00188	0.00187	**0.00186**	0.00187	1.00000	0.90189	0.89399	**0.88889**	0.89541
$\frac{1}{4}$	0.00936	0.00663	**0.00615**	0.00834	0.00758	1.00000	0.70822	**0.65706**	0.89042	0.80928
$\frac{1}{2}$	0.01835	0.01644	**0.01316**	0.01608	0.01534	1.00000	0.89558	**0.71680**	0.87582	0.83553
1	0.03404	0.04292	**0.02882**	0.03084	0.02951	1.00000	1.26089	**0.84675**	0.90605	0.86718
2	0.07708	0.12968	**0.05398**	0.07248	0.06261	1.00000	1.68233	**0.70031**	0.94033	0.81229
3	0.20066	0.27608	**0.12125**	0.15875	0.17730	1.00000	1.37586	**0.60422**	0.79110	0.88356
					15-year rolling window					
$\frac{1}{12}$	0.00203	0.00196	**0.00182**	0.00186	0.00187	1.00000	0.96935	**0.89922**	0.91664	0.92465
$\frac{1}{4}$	0.00826	0.00692	**0.00605**	0.00656	0.00654	1.00000	0.83711	**0.73186**	0.79379	0.79109
$\frac{1}{2}$	0.02009	0.01714	**0.01548**	0.01846	0.01697	1.00000	0.85304	**0.77052**	0.91870	0.84451
1	0.03996	0.04449	0.03013	**0.02940**	0.03686	1.00000	1.11338	0.75411	**0.73572**	0.92240
2	0.05947	0.13694	**0.03887**	0.05240	0.05392	1.00000	2.30285	**0.65368**	0.88111	0.90664
3	0.11166	0.29198	**0.08014**	0.10578	0.11163	1.00000	2.61489	**0.71774**	0.94737	0.99971

Note: Bold numbers are for the best performance.

As recognized in the literature, the signal of persistent predictors may become stronger in long-horizon return prediction (Cochrane, 2009). In addition to the one-month-ahead short-horizon prediction, we also construct the long-horizon excess return as the sum of continuous compounded monthly excess return on the S&P 500 index,

$$\text{LongReturn}_i = \sum_{k=i}^{i+12\times h-1} \text{ExReturn}_k$$

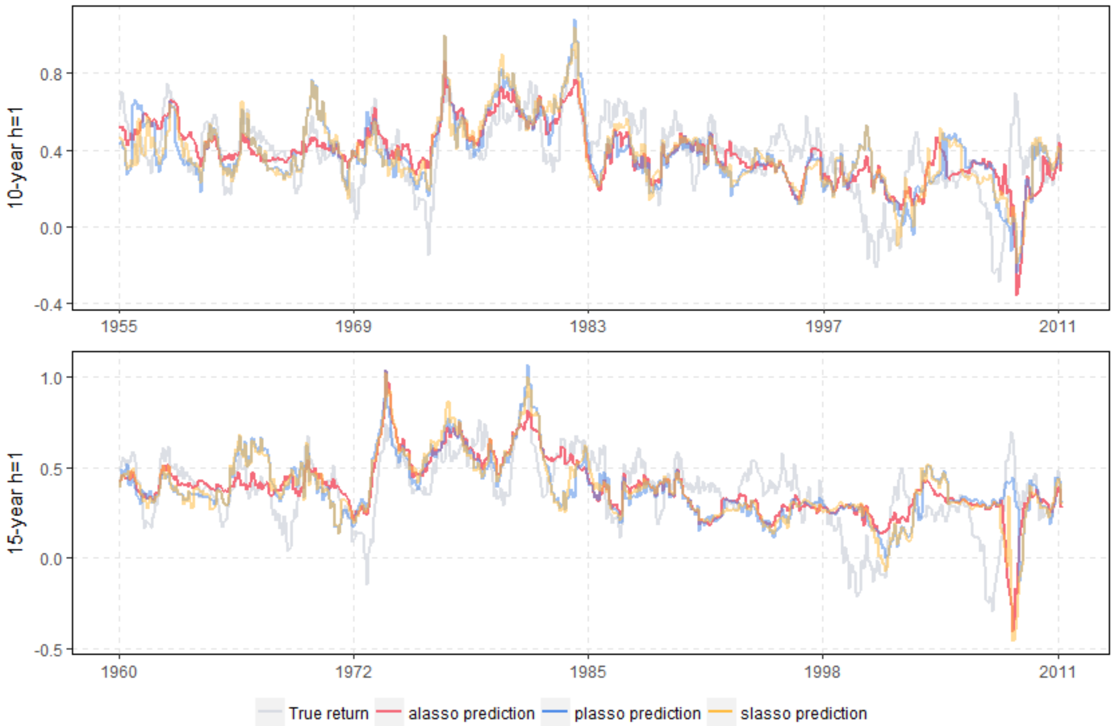
where h is the length of the forecasting horizon. $h = 1$ stands for one year. We choose $h = \frac{1}{12}, \frac{1}{4}, \frac{1}{2}, 1, 2, 3$ for illustration, so $h = \frac{1}{12}$ is one-month-ahead short-horizon prediction.

5.2 Performance Comparison

We apply the set of feasible forecasting methods as in Section 4 to forecast both short-horizon and long-horizon stock returns recursively with both 10-year and 15-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, $\widehat{y}_{n+1} = \frac{1}{n}\sum_i^n y_i$, as another benchmark. All variables are included in the predictive regression, to which Welch and Goyal (2008) refer as the *kitchen sink model*. 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 relative to that of OLS. The tuning parameter for shrinkage estimators are automatically determined by 10-fold cross-validation with consecutive partitions in each estimation window.

The forecasting performance results are summarized in Table 3. All three shrinkage methods can improve the OLS and RWwD benchmarks, and Alasso outperforms the others in most cases. In short-horizon ($h = \frac{1}{12}$) prediction with 10-year rolling window, Plasso is the marginal winner. As the signal accumulates in the long-horizon prediction, Alasso achieves smaller MPSE. The exceptional case with 15-year rolling window and $h = 1$ is due to that Alasso fails to track the recovery trend

Figure 1: True Return v.s. Predicted Return



after the financial crisis. As shown in Figure 1, Plasso provides better forecasts during the periods after the financial crisis whereas Alasso gives the opposite prediction. With 10-year rolling window, all three methods fail to provide sound forecasts after the financial crisis. The financial crisis is such an unprecedented event in the given sample time period. We use the case with 10-year rolling window and $h = 1$ as an example and plot the estimated coefficients in Figure 2. The unstable nature of the predictive model is clear. The shrinkage methods select different variables as the estimation windows roll over. Alasso eliminates more variables than Plasso and Slasso do and hence suggests more parsimonious models. Similar patterns appear in other cases.

In terms of prediction performance, it is known that eliminating irrelvant predictors is more important than including relevant ones, see Ploberger and Phillips (2003), for example. Inclusion of the irrelevant predictors could be detrimental in forecasting contexts, and including irrelvant I(1) predictors in predictive regression can be extremely harmful since stock returns are supposed to be stationary. In this sense, Alasso can provide a more conservative variable selection in predictive regression.

6 Conclusion

We explore a few popular LASSO-type procedures in the presence of stationary, nonstationary and cointegrated predictors. Similar issues raised by Zou (2006) are confirmed within this environment. Alasso is promising even with time series regression with mixed roots. In addition to the well known benefit of Alasso—ability to differentiate the penalty structure according to zero and non-zero regression coefficients, we show that the adaptiveness is extended to the order of integration of predictors—I(0), I(1) or cointegration. In this sense, Alasso is rate-adaptive to a system of multiple

Figure 2: Estimated Coefficients (10-year rolling window, $h = 1$)

predictors with various degrees of persistence, unlike the conventional LASSO that imposes the penalty based only on the marginal variation of each predictor.

Alasso saves the effort to sort out the predictors according to their degrees of persistence so we can be agnostic to the time series properties of predictors. The automatic penalty adjustment of Alasso guarantees consistent model selection and optimal rate of convergence. Such desirable properties may in practice improve the out-of-sample prediction performance under the complex predictive environment with a mixture of regressors.

To focus on the mixed root setting, we adopt the simplest asymptotic framework with fixed p and $n \to \infty$ to demonstrate the clear contrast amongst OLS, Alasso, Plasso and Slasso. This asymptotic framework is in line with the-state-of-art of the predictive regression study in financial econometrics (Kostakis et al., 2014; Phillips and Lee, 2016; Xu, 2018). On the other hand, a large number of potential regressors are available in the era of big data. The practice demands an extension to allow an infinite number of regressors in the limit. As the restricted eigenvalue condition (Bickel et al., 2009) is unsuitable in our context where the nonstationary part of the Gram matrix does not degenerate, in future research we will seek new technical apparatus to deal with the minimal eigenvalue of the Gram matrix of unit root processes in order to generalize the insight gleaned from the fixed-p asymptotics to diverging p.

References

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

Belloni, A. and V. Chernozhukov (2011). l_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics 39*(1), 82–130.

Belloni, A., V. Chernozhukov, D. Chetverikov, and Y. Wei (2018). Uniformly valid post-regularization confidence regions for many functional parameters in Z-estimation framework. *The Annals of Statistics 46*(6B), 3643–3675.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Belloni, A., V. Chernozhukov, and K. Kato (2015). Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika 102*(1), 77–94.

Belloni, A., V. Chernozhukov, and K. Kato (2018). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association* (in press), 1–33.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics 37*(4), 1705–1732.

Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory 25*(1), 270–290.

Caner, M. and K. Knight (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), 691–715.

Caner, M. and H. H. Zhang (2014). Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics 32*(1), 30–47.

Chen, S. S., D. L. Donoho, and M. A. Saunders (2001). Atomic decomposition by basis pursuit. *SIAM review 43*(1), 129–159.

Cochrane, J. H. (2009). *Asset pricing (revised edition)*. Princeton university press.

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

Hirano, K. and J. H. Wright (2017). Forecasting with model uncertainty: Representations and risk reduction. *Econometrica 85*(2), 617–643.

Inoue, A. and L. Kilian (2008). 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), 511–522.

Knight, K. and W. Fu (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics 28*(5), 1356–1378.

Kock, A. B. (2016). Consistent and conservative model selection with the adaptive LASSO in stationary and nonstationary autoregressions. *Econometric Theory 32*(1), 243–259.

Kock, A. B. and L. Callot (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics 186*(2), 325–344.

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

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

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

Lettau, M. and S. Ludvigson (2001). Consumption, aggregate wealth, and expected stock returns. *The Journal of Finance 56*(3), 815–849.

Liao, Z. and P. C. Phillips (2015). Automated estimation of vector error correction models. *Econometric Theory 31*(3), 581–646.

Medeiros, M. C. and E. F. Mendes (2016). l_1-regularization of high-dimensional time-series models with non-gaussian and heteroskedastic errors. *Journal of Econometrics 191*(1), 255–271.

Park, J. Y. (1992). Canonical cointegrating regressions. *Econometrica 60*(1), 119–143.

Phillips, P. C. (1987). Time series regression with a unit root. *Econometrica 55*(2), 277–301.

Phillips, P. C. (1991). Optimal inference in cointegrated systems. *Econometrica 59*(2), 283–306.

Phillips, P. C. and B. E. Hansen (1990). Statistical inference in instrumental variables regression with I(1) processes. *The Review of Economic Studies 57*(1), 99–125.

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

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

Phillips, P. C. B. and V. Solo (1992). Asymptotics for linear processes. *The Annals of Statistics 20*(2), 971–1001.

Ploberger, W. and P. C. Phillips (2003). Empirical limits for time series econometric models. *Econometrica 71*(2), 627–673.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Timmermann, A. and Y. Zhu (2017). Monitoring forecasting performance. *working paper*.

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

Xu, K.-L. (2018). Testing for return predictability with co-moving predictors of unknown form. *working paper*.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 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 around the true parameter β_n^*, and let

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

Define $\widehat{v}^{(n)} = n(\widehat{\beta}^{Alasso} - \beta_n^*)$. The fact that $\widehat{\beta}^{Alasso}$ minimizes (6) implies $\widehat{v}^{(n)} = \arg\min_v \Psi_n(v)$. Let

$$\begin{aligned} V_n(v) &= \Psi_n(v) - \Psi_n(0) \\ &= \|u - \frac{X'v}{n}\|^2 - \|u\|^2 + \lambda_n \left(\sum_{j=1}^{p} \widehat{\tau}_j |\beta_{jn}^* + \frac{v_j}{n}| - \sum_{j=1}^{p} \widehat{\tau}_j |\beta_{jn}^*| \right) \\ &= v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \lambda_n \sum_{j=1}^{p} \widehat{\tau}_j (|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|). \end{aligned} \tag{21}$$

The first and the second term in the right-hand side of (21) converge in distribution, $\frac{X'X}{n^2} \Longrightarrow \Omega$ and $\frac{u'X}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{i\cdot}' u_i \Longrightarrow \zeta$, by the functional central limit theorem (FCLT) and the continuous mapping theorem. The third term involves the weight $\widehat{\tau}_j = |\widehat{\beta}_j^{ols}|^{-\gamma}$ for each j. Since the OLS estimator $n\left(\widehat{\beta}^{ols} - \beta_n^*\right) \Longrightarrow \Omega^{-1}\zeta = O_p(1)$, we have

$$\widehat{\tau}_j = \left|\beta_{jn}^* + O_p\left(n^{-1}\right)\right|^{-\gamma} = |\beta_j^{0*}/n^{\delta_j} + O_p\left(n^{-1}\right)|^{-\gamma} \tag{22}$$

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

$$|\beta_{jn}^* + n^{-1}v_j| - |\beta_{jn}^*| = n^{-1}v_j \mathrm{sgn}(\beta_{jn}^*) = n^{-1}v_j \mathrm{sgn}(\beta_j^{0*}). \tag{23}$$

(22) and (23) now imply

$$\begin{aligned} \lambda_n \widehat{\tau}_j \cdot (|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|) &= \frac{\lambda_n}{n|\beta_j^{0*}/n^{\delta_j} + O_p\left(n^{-1}\right)|^{\gamma}} v_j \mathrm{sgn}(\beta_j^{0*}) = \frac{\lambda_n n^{\delta_j \cdot \gamma}}{n|\beta_j^{0*} + o_p\left(1\right)|^{\gamma}} v_j \mathrm{sgn}(\beta_j^{0*}) \\ &\leq \frac{\lambda_n n^{\gamma\bar{\delta}-1}}{|\beta_j^{0*} + o_p\left(1\right)|^{\gamma}} v_j \mathrm{sgn}(\beta_j^{0*}) = O_p\left(\lambda_n n^{\gamma\bar{\delta}-1}\right) = o_p\left(1\right) \end{aligned} \tag{24}$$

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

$$\lambda_n \widehat{\tau}_j \cdot n(|\beta_j^* + n^{-1}v_j| - |\beta_j^*|) = \frac{\lambda_n}{n|\widehat{\beta}_j^{ols}|^{\gamma}} |v_j| = \frac{\lambda_n n^{\gamma-1}}{|n\widehat{\beta}_j^{ols}|^{\gamma}} |v_j| = \frac{\lambda_n n^{\gamma-1}}{O_p\left(1\right)} |v_j| \to \infty \tag{25}$$

since $\lambda_n n^{\gamma-1} \to \infty$ and the OLS estimator is asymptotically non-degenerate. Thus we have $V_n(v) \Longrightarrow V(v)$ for every fixed v, where

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

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

$$\begin{pmatrix} v_{M^*} \\ v_{M^{*c}} \end{pmatrix} = \begin{pmatrix} \Omega_{M^*}^{-1}\zeta_{M^*} \\ \mathbf{0}_{|M^{*c}|} \end{pmatrix}.$$

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

$$\widehat{v}_{M^*}^{(n)} = n(\widehat{\beta}_{M^*}^{Alasso} - \beta_{M^*}^*) \Longrightarrow \Omega_{M^*}^{-1}\zeta_{M^*} \text{ and } \widehat{v}_{M^{*c}}^{(n)} \Longrightarrow \mathbf{0}_{|M^{*c}|}. \tag{26}$$

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

Next we show variable selection consistency. We have $P(M^* \subseteq \widehat{M}_n) \to 1$ immediately follows from the first part of (26), as $\widehat{v}_{M^*}^{(n)}$ converges in distribution to a non-degenerate continuous random variable. For those $j \in M^{*c}$, if the event $\{j \in \widehat{M}_n\}$ occurs, then the Karush-Kuhn-Tucker (KKT) condition entails

$$\frac{2}{n}x_j'(y - X\widehat{\beta}^{Alasso}) = \frac{\lambda_n \widehat{\tau}_j}{n}. \tag{27}$$

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

$$\frac{\lambda_n \widehat{\tau}_j}{n} = \frac{\lambda_n}{n|\widehat{\beta}_j^{ols}|^\gamma} = \frac{\lambda_n n^{\gamma-1}}{|n\widehat{\beta}_j^{ols}|^\gamma} = \frac{\lambda_n n^{\gamma-1}}{O_p(1)} \to \infty, \tag{28}$$

from the given rate condition. However, using $y = X\beta_n^* + u$ and (26) at the left-hand side of (27), we have

$$\begin{aligned} \frac{2}{n}x_j'(y - X\widehat{\beta}^{Alasso}) &= \frac{2}{n}x_j'(X\beta_n^* - X\widehat{\beta}^{Alasso} + u) = 2\left(\frac{x_j'X}{n^2}\right)n(\beta_n^* - \widehat{\beta}^{Alasso}) + 2\frac{x_j'u}{n} \\ &= 2\left(\frac{x_j'X}{n^2}\right)\left(\widehat{v}_{M^*}^{(n)} + \widehat{v}_{M^{*c}}^{(n)}\right) + 2\frac{x_j'u}{n} \Longrightarrow 2\Omega_{j\cdot}(\Omega_{M^*}^{-1}\zeta_{M^*} + o_p(1)) + 2\zeta_j, \end{aligned} \tag{29}$$

where $\Omega_{j\cdot}$ is the j-th row of Ω. In other words, the left-hand side of (27) remains a non-degenerate continuous random variable in the limit. For any $j \in M^{*c}$, the disparity of the two sides of the KKT condition implies

$$P(j \in \widehat{M}_n) = P\left(\frac{2}{n}x_j'(y - X\widehat{\beta}^{Alasso}) = \frac{\lambda_n \widehat{\tau}_j}{n}\right) \to 0.$$

That is, $P(M^{*c} \subseteq \widehat{M}_n^c) \to 1$ or equivalently $P(\widehat{M}_n \subseteq M^*) \to 1$. We thus conclude the variable selection consistency $P(\widehat{M}_n = M^*) \to 1$. ■

Proof. [Proof of Corollary 2.3] The proof is a simple variant of that of Theorem 2.1 by setting

$\widehat{\tau}_j = 1$ for all j. For Part(a), the counterpart of (21) 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,

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

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

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

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

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

$$V_n(v) = v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \frac{\lambda_n}{n}D\left(\mathbf{1}_p, v, \beta^{0*}\right) \Longrightarrow v'\Omega v - 2v'\zeta + c_\lambda D\left(\mathbf{1}_p, v, \beta^{0*}\right).$$

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

$$\widetilde{\Psi}_n(v) = \|Y - X\left(\beta^*_n + \frac{\lambda_n}{n^2}v\right)\|^2 + \lambda_n \sum_{j=1}^{p}|\beta^*_{jn} + \frac{\lambda_n}{n^2}v_j|,$$
$$\widetilde{V}_n(v) = \widetilde{\Psi}_n(v) - \widetilde{\Psi}_n(0) = \frac{\lambda_n^2}{n^4}v'(X'X)v - \frac{\lambda_n}{n^2}2u'Xv + \lambda_n \sum_{j=1}^{p}(|\beta^*_{jn} + \frac{\lambda_n}{n^2}v| - |\beta^*_{jn}|).$$

If $\beta^{0*}_j \neq 0$, in the limit $\frac{\lambda_n}{n^2}v$ is dominated by any $\beta^*_{jn} = \beta^{0*}_j/n^{\delta_j}$ given the rate $\frac{\lambda_n}{n^{2-\delta}} \to 0$. For a sufficiently large n,

$$\begin{aligned}
\widetilde{V}_n(v) &= \frac{\lambda_n^2}{n^2}v'(\frac{X'X}{n^2})v - \frac{\lambda_n}{n}2\left(\frac{u'X}{n}\right)v + \frac{\lambda_n^2}{n^2}D\left(\mathbf{1}_p, v, \beta^*_0\right) \\
&= \frac{\lambda_n^2}{n^2}\left[v'(\frac{X'X}{n^2})v - \frac{1}{\lambda_n/n}2\left(\frac{u'X}{n}\right)v + D\left(\mathbf{1}_p, v, \beta^*_0\right)\right] \\
&= \frac{\lambda_n^2}{n^2}\left[v'(\frac{X'X}{n^2})v + o_p(1) + D\left(\mathbf{1}_p, v, \beta^*_0\right)\right].
\end{aligned}$$

Notice that the scaled difference $\widehat{v}^{(n)} = \lambda_n^{-1}n^2(\widehat{\beta}^{lasso} - \beta^*_n)$ can be expressed as $\widehat{v}^{(n)} = \arg\min_v \widetilde{\Psi}_n(v)$. By the strict convexity of $\widetilde{V}_n(v)$ and $\widetilde{V}(v) = v'\Omega v + D\left(\mathbf{1}_p, v, \beta^*_0\right)$, we invoke the Convexity Lemma to obtain $\frac{n^2}{\lambda_n}(\widehat{\beta}^{Plasso} - \beta^*_n) \Longrightarrow \arg\min_v \widetilde{V}(v)$. ■

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

$$\Psi_n(v) = \|Y - X\left(\beta_n^* + \frac{v}{n}\right)\|^2 + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j |\beta_{jn}^* + \frac{v_j}{n}|,$$

$$V_n(v) = \Psi_n(v) - \Psi_n(0) = v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|).$$

When $\lambda_n/\sqrt{n} \to c_\lambda \geq 0$ and $\frac{\widehat{\sigma}_j}{\sqrt{n}} \Longrightarrow d_j$ as in (9), the penalty term

$$\lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j(|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|) = \frac{\lambda_n}{\sqrt{n}} D\left(\frac{\widehat{\sigma}}{\sqrt{n}}, v, \beta^{0*}\right) \Longrightarrow c_\lambda \sum_{j=1}^{p} D\left(d, v, \beta^{0*}\right)$$

where $\widehat{\sigma} = (\widehat{\sigma}_j)_{j=1}^{p}$. Part (b) follows by the same argument in the proof of Corollary 2.3(b), and Part (a) is simply the special case when $c_\lambda = 0$.

Part (c) is also 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

$$\widetilde{\Psi}_n(v) = \|Y - X\left(\beta_n^* + \frac{\lambda_n}{n^{3/2}}v\right)\|^2 + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j |\beta_{jn}^* + \frac{\lambda_n}{n^{3/2}}v_j|,$$

$$\widetilde{V}_n(v) = \widetilde{\Psi}_n(v) - \widetilde{\Psi}_n(0) = \frac{\lambda_n^2}{n^3} v'(X'X)v - \frac{\lambda_n}{n^{3/2}} 2u'Xv + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j(|\beta_{jn}^* + \frac{\lambda_n}{n^{3/2}}v_j| - |\beta_{jn}^*|).$$

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

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

so that

$$\begin{aligned}
\widetilde{V}_n(v) &= \frac{\lambda_n^2}{n}\left[v'(\frac{X'X}{n^2})v - \frac{1}{\lambda_n/\sqrt{n}} 2\left(\frac{u'X}{n}\right)v + D\left(\frac{\widehat{\sigma}}{\sqrt{n}}, v, \beta^{0*}\right)\right] \\
&= \frac{\lambda_n^2}{n}\left[v'(\frac{X'X}{n^2})v + D\left(\frac{\widehat{\sigma}}{\sqrt{n}}, v, \beta^{0*}\right) + o_p(1)\right].
\end{aligned}$$

Define $\widetilde{V}(v) = v'\Omega v + D\left(d, v, \beta^{0*}\right)$, and the conclusion follows. ■

A.2 Proofs in Section 3

Derivation of Eq.(15)

The columns in X^+ are a unit root processes with no cointegration relationship. Let the i-th row of X^+ be $\underset{(p_2+p_x)\times 1}{X_i^+} = [X_{2i\cdot}^c, X_{i\cdot}]'$. Using the component-wise BN decomposition, the scalar

$u_i = \underset{1\times(p+1)}{F_u(1)} \times \underset{(p+1)\times 1}{\varepsilon_i} - \triangle\widetilde{\varepsilon}_{ui}$. Thus we have

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

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

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

where Δ_{+u} is the corresponding submatrix of the one-sided long-run covariance and $u_{xi}^+ = X_i^+ - X_{i-1}^+$. Combining these results, we have (15).

Proof. [Proof of Theorem 3.2] The OLS estimator

$$\begin{aligned} R_n\left(\widehat{\theta}_n^{ols} - \theta_n^*\right) &= R_n\left(W'W\right)^{-1} W'u \\ &= R_n\left(R_nQ\right)^{-1} R_nQ\left(W'W\right)^{-1} Q'R_n\left(Q'R_n\right)^{-1} W'u \\ &= R_nQ^{-1}R_n^{-1}\left[R_n^{-1}Q'^{-1}W'WQ^{-1}R_n^{-1}\right]^{-1} R_n^{-1}Q'^{-1}W'u, \end{aligned} \tag{30}$$

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

$$\begin{aligned} WQ^{-1} &= [Z, X_1^c, X_2^c, X] \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & 0 & 0 \\ 0 & -A_1' & I_{p_2} & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix} \\ &= \left[Z, X_1^c - X_2^c A_1', X_2^c, X\right] = \left[Z, V_1, X_2^c, X\right], \end{aligned}$$

in which I(0) and I(1) components are separated. We have

$$R_n^{-1}Q'^{-1}W'WQ^{-1}R_n^{-1} = \begin{pmatrix} \frac{Z^{+\prime}Z^+}{n} & \frac{Z^{+\prime}X^+}{n^{3/2}} \\ \frac{Z^{+\prime}X^+}{n^{3/2}} & \frac{X^{+\prime}X^+}{n^2} \end{pmatrix} \Longrightarrow \begin{pmatrix} \Omega_{zz}^+ & 0 \\ 0 & \Omega_{xx}^+ \end{pmatrix} = \Omega^+. \tag{31}$$

Since

$$R_n^{-1}Q'^{-1}W'u = \begin{pmatrix} Z^{+\prime}u/\sqrt{n} \\ X^{+\prime}u/n \end{pmatrix} \Longrightarrow \zeta^+ \tag{32}$$

and

$$R_nQ^{-1}R_n^{-1} = \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & 0 & 0 \\ 0 & -A_1'/\sqrt{n} & I_{p_2} & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix} \to I_p, \tag{33}$$

the conclusion follows by substituting (31), (32) and (33) into (30). ∎

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

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

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

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

The first term

$$v' R_n^{-1} W' W R_n^{-1} v = v' \left(R_n^{-1} Q' R_n\right) \left(R_n^{-1} Q'^{-1} W'^{-1} R_n^{-1}\right) \left(R_n Q R_n^{-1}\right) v \Longrightarrow v' \Omega^+ v \tag{34}$$

by (31) and (33) as we have shown in the proof of Theorem 3.2. Similarly, the second term

$$2v' R_n^{-1} W' u = 2v' \left(R_n^{-1} Q' R_n\right) \left(R_n^{-1} Q'^{-1} W' u\right) \Longrightarrow 2v' \zeta^+. \tag{35}$$

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 $j \in \mathcal{I}_0 \cup \mathcal{C}_1$, the coefficients are invariant with the sample size. If $\theta_j^{0*} \neq 0$, we have $(|\theta_{jn}^* + n^{-1/2} v_j| - |\theta_{jn}^*|) = n^{-1/2} v_j \mathrm{sgn}(\theta_j^{0*})$, and

$$\lambda_n \widehat{\tau}_j \cdot (|\theta_{jn}^* + \frac{v_j}{\sqrt{n}}| - |\theta_{jn}^*|) = O_p\left(\lambda_n n^{-1/2}\right) = o_p(1).$$

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

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

  given the rate of λ_n.

- For $j \in \mathcal{C}_2 \cup \mathcal{I}_1$, the coefficient $\theta_{jn}^* = \theta_j^{0*}/n^{\delta_j}$ depending on n, where δ_j is the corresponding coefficient's diminishing rate to zero. If $\theta_j^{0*} \neq 0$, then θ_{jn}^* dominates $n^{-1} v_j$ in the limit. We have $(|\theta_{jn}^* + n^{-1} v_j| - |\theta_{jn}^*|) = n^{-1} v_j \mathrm{sgn}(\theta_j^{0*}) = n^{-1} v_j \mathrm{sgn}(\theta_j^{0*})$, and

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

  by the same derivation in (24). On the other hand, if $\theta_j^{0*} = 0$, then

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

according to the derivation in (25).

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

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

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

$$\widehat{v}_{M^*}^{(n)} = [R_n^{-1}(\widehat{\theta}^{Alasso} - \theta_n^*)]_{M^*} \Longrightarrow (\Omega_{M^*})\,\zeta_{M^*}^{+} \text{ and } \widehat{v}_{M^{*c}}^{(n)} \Longrightarrow \mathbf{0}_{|M^{*c}|}. \tag{36}$$

The first part of the above result establishes Theorem 3.4(b), and it also implies $P(M^* \subseteq \widehat{M}_n) \to 1$.

For $j \in M^{*c}$, if the event $\{j \in \widehat{M}_n\}$ occurs, then the KKT condition entails

$$2w_j'(y - W\widehat{\theta}^{Alasso}) = \lambda_n \widehat{\tau}_j. \tag{37}$$

We will invoke similar argument as in (28) and (29) to show the disparity of the two sides of the KKT condition. The left-hand side of (37) is the j-th element of the $p \times 1$ vector $2W'(y - W\widehat{\theta}^{Alasso})$.

If we pre-multiply the diagonal matrix $\frac{1}{2}R_n^{-1}$ to the vector $2W'(y - W\widehat{\theta}^{Alasso})$, we have

$$\begin{aligned} R_n^{-1}W'(y - W\widehat{\theta}^{Alasso}) &= R_n^{-1}W'\left(W\left(\theta_n^* - \widehat{\theta}^{Alasso}\right) + u\right) \\ &= R_n^{-1}W'WR_n^{-1}R_n\left(\theta_n^* - \widehat{\theta}^{Alasso}\right) + R_n^{-1}W'u \\ &= R_n^{-1}W'WR_n^{-1}\left(\widehat{v}_{M^*}^{(n)} + \widehat{v}_{M^{*c}}^{(n)}\right) + R_n^{-1}W'u \\ &\Longrightarrow \Omega^{+}O_p(1) + \zeta^{+} = O_p(1) \end{aligned}$$

where the last line follows by (34), (35) and (36). We thus verify the left-hand side of (37) is of $O_p(R_{jn})$ as it is the j-th row of the p-equation system.

Now if we multiply R_{jn}^{-1} to the right-hand side of (37), we have

$$\frac{1}{R_{jn}}\lambda_n\widehat{\tau}_j = \frac{\lambda_n}{R_{jn}|\widehat{\theta}_j^{ols}|^{\gamma}} = \frac{\lambda_n R_{jn}^{\gamma-1}}{|R_{jn}\widehat{\theta}_j^{ols}|^{\gamma}} = O_p\left(\lambda_n R_{jn}^{\gamma-1}\right) \to \infty$$

as $\gamma \geq 1$ and $\lambda_n \to \infty$. We thus verify the right-hand side of (37) is of order bigger than $O_p(R_{jn})$. It immediately follows that given the specified rate for λ_n, for any $j \in M^{*c}$ we have

$$P(j \in \widehat{M}_n) = P\left(2R_{jn}^{-1}w_j^{+\prime}(y - W\widehat{\theta}^{Alasso}) = R_{jn}^{-1}\lambda_n\widehat{\tau}_j\right) \to P\left(O_p(1) = \infty\right) = 0.$$

In other words, $P(M^{*c} \subseteq \widehat{M}_n^c) \to 1$ or equivalently $P(\widehat{M}_n \subseteq M^*) \to 1$. 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

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

In view of (34) and (35),

$$V_n(v) \Longrightarrow V(v) = v'\Omega^+ v - 2v'\zeta^+ + \lim_{n\to\infty}\left\{\frac{\lambda_n}{\sqrt{n}}\sum_{j\in\mathcal{I}_0\cup\mathcal{C}_1} D(1, v_j, \theta_j^{0*}) + \frac{\lambda_n}{n}\sum_{j\in\mathcal{C}_2\cup\mathcal{I}_1} D(1, v_j, \theta_j^{0*})\right\} \tag{38}$$

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.

Part (c) needs more subtle investigation. Define

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

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

$$\begin{aligned}\left(\frac{n}{\lambda_n^2}\right)\widetilde{V}_n(v) \quad &= \quad v'\left(R_n^{-1}W'WR_n^{-1}\right)v - 2v'R_n^{-1}W'u/\lambda_n \\ &\quad + \frac{n}{\lambda_n}\sum_{j\in\mathcal{I}_0\cup\mathcal{C}_1}(|\theta_{jn}^* + \frac{\lambda_n}{n}v_j| - |\theta_{jn}^*|) + \frac{n}{\lambda_n}\sum_{j\in\mathcal{I}_1\cup\mathcal{C}_2}(|\theta_{jn}^* + \frac{\lambda_n}{n^{3/2}}v_j| - |\theta_{jn}^*|).\end{aligned} \tag{39}$$

By the rate condition of λ_n, the second term $2v'R_n^{-1}W'u/\lambda_n = o_p(1)$. Given $v_j \neq 0$ and n large enough:

- If $j \in \mathcal{I}_0 \cup \mathcal{C}_1$, $\theta_{jn}^* = \theta_j^{0*}$ is invariant with n so that

$$\frac{n}{\lambda_n}|\theta_j^{0*} + \frac{\lambda_n}{n}v_j| - |\theta_j^{0*}|) = \frac{n}{\lambda_n}D\left(1, \frac{\lambda_n}{n}v_j, \theta_j^{0*}\right) = D\left(1, v_j, \theta_j^{0*}\right). \tag{40}$$

- If $j \in \mathcal{C}_2 \cup \mathcal{I}_1$, the coefficient $\theta_{jn}^* = \theta_j^{0*}/n^{\delta_j}$ may shrink faster than $\frac{\lambda_n}{n^{3/2}}$. The elementary inequality $||a+b| - |a|| \leq |b|\, I\left(|b| \geq |a|\right) + 3\,|b|\, I\left(|a| < |b|\right) \leq 3\,|b|$ for any $a, b \in \mathbb{R}$ guarantees

$$\left|\frac{n}{\lambda_n}\left(|\theta_{jn}^* + \frac{\lambda_n}{n^{3/2}}v_j| - |\theta_{jn}^*|\right)\right| \leq 3\frac{n}{\lambda_n}\left|\frac{\lambda_n}{n^{3/2}}v_j\right| = O\left(n^{-1/2}\right), \tag{41}$$

  which is dominated by $D\left(1, v_j, \theta_j^{0*}\right)$ in the limit if $v_j \neq 0$ and $\theta_j^{0*} \neq 0$.

Thus we have

$$\left(\frac{n}{\lambda_n^2}\right)\widetilde{V}_n(v) \Longrightarrow v'\Omega^+ v + \sum_{\mathcal{I}_0\cup\mathcal{C}_1} D\left(1, v_j, \theta_j^{0*}\right),$$

and the Convexity Lemma implies

$$\begin{aligned}\frac{n}{\lambda_n}(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{I}_0\cup\mathcal{C}_1} \quad &\Longrightarrow \quad \arg\min_v\left\{v'\Omega_{zz}^+ v + \sum_{j\in\mathcal{I}_0\cup\mathcal{C}_1} D\left(1, v_j, \theta_j^{0*}\right)\right\} \\ \frac{n^{3/2}}{\lambda_n}(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{C}_2\cup\mathcal{I}_1} \quad &\xrightarrow{p} \quad \mathbf{0}_{p_2+p_x}\end{aligned} \tag{42}$$

Next, define

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

where $\check{R}_n = \begin{pmatrix} \check{\lambda}_n^{-1}\sqrt{n}I_{p_z+p_1} & 0 \\ 0 & nI_{p_2+p_x} \end{pmatrix}$ and $\check{\lambda}_n$ is any sequence such that $\check{\lambda}_n/n \to 0$ and $\check{\lambda}_n/\lambda_n \to \infty$. Similar derivation shows that

$$\check{V}_n(v) \Longrightarrow \check{V}(v) = v_x'\Omega_{xx}^+ v_x - 2v_x'\zeta_{x^+}^+ + \lim_{n\to\infty}\frac{\check{\lambda}_n^2}{n}\left(v_z'\Omega_{zz}^+ v_z + \frac{\check{\lambda}_n}{\lambda_n}\sum_{j\in\mathcal{I}_0\cup\mathcal{C}_1} D(1, v_j, \theta_j^{0*})\right)$$

where $v_x = v_{\mathcal{I}_1\cup\mathcal{C}_2}$ and $v_x = v_{\mathcal{I}_0\cup\mathcal{C}_1}$. Since $\check{V}(v) \to \infty$ for any $v_z \neq \mathbf{0}_{p_z+p_1}$, we must have

$$\check{V}(v) = \begin{cases} v_x'\Omega_{xx}^+ v_x - 2v_x'\zeta_{x^+}^+, & \text{if } v_z = \mathbf{0}_{p_z+p_1}. \\ \infty, & \text{otherwise.} \end{cases}$$

It implies that

$$\begin{aligned} \frac{n}{\check{\lambda}_n}(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{I}_0\cup\mathcal{C}_1} &\xrightarrow{p} \mathbf{0}_{p_z+p_1} \\ n(\widehat{\theta}^{Plasso} - \theta_n^*)_{\mathcal{C}_2\cup\mathcal{I}_1} &\Longrightarrow \left(\Omega_{xx}^+\right)^{-1}\zeta_{x^+}. \end{aligned} \tag{43}$$

The conclusion follows by combining (42) and (43). ■

Proof. [Proof of Corollary 3.9] For Part (a) and (b), 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

$$V_n(v) = v'\left(R_n^{-1}W'WR_n^{-1}\right)v - 2vR_n^{-1}W'u + c_\lambda \cdot \sum_{j=1}^{p}\widehat{\sigma}_j(|\theta_{jn}^* + R_{jn}v_j| - |\theta_{jn}^*|). \tag{44}$$

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

- if $j \in \mathcal{I}_0$, the coefficient θ_j^{0*} is independent of n and

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

  since these indices are associated with the stationary variable Z and therefore $\widehat{\sigma}_j = O_p(1)$;

- if $j \in \mathcal{C}_1$, again θ_j^{0*} is independent of n and

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

  since these indices are associated with unit root processes in X_1^c and therefore $\frac{\widehat{\sigma}_j}{\sqrt{n}} \Longrightarrow d_j$;

- if $j \in \mathcal{C}_2 \cup \mathcal{I}_1$, for these unit root processes in X_2^c and X, similarly we have

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

The above analysis of the third term of (44) implies

$$V_n(v) \Longrightarrow V\left(v\right) = v'\left(R_n^{-1} W' W R_n^{-1} \right) v - 2v' R_n^{-1} W' u + c_\lambda \sum_{j \in \mathcal{C}_1} D\left(d_j, v_j, \theta_j^{0*} \right),$$

and the conclusion follows.

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

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

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

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

by the rate condition of λ_n. Again we study the last term. For $v_j \neq 0$ and a sufficiently large n:

- for $j \in \mathcal{I}_0$,

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

- for $j \in \mathcal{C}_1$,

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

- for $j \in \mathcal{C}_2 \cup \mathcal{I}_1$, the rate condition $\lambda_n / n^{(1-\bar{\delta}) \wedge 0.5} \to 0$ makes sure that $\theta^*_{jn} = \theta_j^{0*} / n^{\delta_j}$ dominates $\frac{\lambda_n}{n}$ in the limit, so that

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

We obtain $\frac{\widetilde{V}_n(v)}{\lambda_n^2} \Longrightarrow v'\Omega^+ v + \sum_{j\in\mathcal{C}_1} D\left(d_j, v_j, \theta_j^{0*}\right)$ and it follows that

$$\frac{R_n}{\lambda_n}(\widehat{\theta}^{Slasso} - \theta^*) \Longrightarrow \arg\min_v \left\{ v'\Omega^+ v + \sum_{j\in\mathcal{C}_1} D\left(d_j, v_j, \theta_j^{0*}\right) \right\}.$$

The above expression implies that

$$\begin{aligned} \frac{\sqrt{n}}{\lambda_n}(\widehat{\theta}^{Slasso} - \theta^*)_{\mathcal{C}_1} &\Longrightarrow \arg\min_v \left\{ v'\Omega^+_{\mathcal{C}_1} v + \sum_{j\in\mathcal{C}_1} D\left(d_j, v_j, \theta_j^{0*}\right) \right\} \\ (R_n)_{\mathcal{I}_0\cup\mathcal{C}_2\cup\mathcal{I}_1}(\widehat{\theta}^{Slasso} - \theta_n^*)_{\mathcal{I}_0\cup\mathcal{C}_2\cup\mathcal{I}_1} &\xrightarrow{p} \mathbf{0}_{p_z+p_2+p_x}. \end{aligned} \tag{45}$$

By essentially parallel argument as in the proof of Corollary 3.7 (c), we can introduce another $\check{\lambda}_n$ such that $\check{\lambda}_n/n \to 0$ and $\check{\lambda}_n/\lambda_n \to 0$ and derive that

$$\begin{aligned} \frac{R_n}{\check{\lambda}_n}(\widehat{\theta}^{Slasso} - \theta^*)_{\mathcal{C}_1} &\xrightarrow{p} \mathbf{0}_{p_1} \\ (R_n)_{\mathcal{I}_0\cup\mathcal{C}_2\cup\mathcal{I}_1}(\widehat{\theta}^{Slasso} - \theta_n^*)_{\mathcal{I}_0\cup\mathcal{C}_2\cup\mathcal{I}_1} &\Longrightarrow (\Omega^+_{\mathcal{I}_0\cup\mathcal{C}_2\cup\mathcal{I}_1})^{-1}\zeta^+_{\mathcal{I}_0\cup\mathcal{C}_2\cup\mathcal{I}_1}. \end{aligned} \tag{46}$$

The conclusion follows by collecting (45) and (46). ■

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 and Knight (2013) proposing to treat the unit root test as a model selection problem by regressing Δy_{i+1} on lags of y_i, we use 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

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

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

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

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 accommodate lags of predictors in predictive regressions, for example Medeiros and Mendes (2016). We employ the following DGP in which a stationary autoregression generates the dependent variable

$$y_{i+1} = \gamma^* + \rho^* y_i + \sum_{j=1}^{2} \phi^*_{jn} x^c_{ij} + \beta^*_{1n} x_i + \beta^*_{2n} x_{i-1} + \sum_{j=1}^{3} \left(\alpha^*_{j1} z_{ij} + \alpha^*_{j2} z_{i-1,j}\right) + u_{i+1}$$

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^c_{i1} and x^c_{i2} are generated by $x^c_{i2} = x^c_{i1} - \mu_i$ where x^c_{i1} 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.

Table 4: Mean Prediction Squared Error (MPSE)

	n	Oracle	OLS	Alasso	Plasso	Slasso
DGP 3	40	1.2041	1.6302	1.3955	1.4681	**1.3407**
	80	1.1000	1.2244	1.1504	1.1823	**1.1399**
	120	1.0703	1.1815	**1.1022**	1.1255	1.1084
	200	0.9686	0.9962	**0.9878**	0.9942	0.9917
	400	0.9971	1.0131	**0.9986**	1.0026	1.0023
	800	1.0085	1.0175	**1.0110**	1.0162	1.0134
DGP 4	40	1.3062	1.5539	1.5104	**1.4882**	1.5178
	80	1.2616	1.3047	1.2944	**1.2879**	1.2953
	120	1.0529	1.0945	**1.0783**	1.0873	1.0933
	200	1.0794	1.1202	**1.1003**	1.1083	1.1170
	400	1.0055	1.0177	**1.0110**	1.0139	1.0153
	800	1.0496	1.0537	**1.0504**	1.0535	1.0548

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

The results summarized in Table 4 and 5 are similar to that in DGP 2, which demonstrates the merits of Alasso in the presence of autoregression.

Table 5: Variable Screening

	n	SR Alasso	SR Plasso	SR Slasso	SR_1 Alasso	SR_1 Plasso	SR_1 Slasso	SR_2 Alasso	SR_2 Plasso	SR_2 Slasso
DGP 3	40	0.6662	0.5957	**0.6996**	0.8300	0.8558	**0.8918**	0.5024	0.3356	**0.5074**
	80	**0.6846**	0.5693	0.6753	0.8402	0.9092	**0.9220**	**0.5290**	0.2294	0.4286
	120	**0.6772**	0.5544	0.6599	0.8448	0.9256	**0.9278**	**0.5096**	0.1832	0.3920
	200	**0.6878**	0.5525	0.6515	0.8422	**0.9450**	0.9378	**0.5334**	0.1600	0.3652
	400	**0.6849**	0.5481	0.6307	0.8350	**0.9622**	0.9548	**0.5348**	0.1340	0.3066
	800	**0.7010**	0.5478	0.6270	0.8356	**0.9742**	0.9628	**0.5664**	0.1214	0.2912
DGP 4	40	**0.8188**	0.7446	0.6549	0.9449	0.9743	**0.9921**	**0.5983**	0.3428	0.0648
	80	**0.8547**	0.7330	0.6558	0.9691	0.9900	**0.9957**	**0.6545**	0.2833	0.0610
	120	**0.8649**	0.7273	0.6513	0.9684	0.9890	**0.9937**	**0.6838**	0.2693	0.0520
	200	**0.8773**	0.7210	0.6546	0.9673	0.9910	**0.9941**	**0.7198**	0.2485	0.0605
	400	**0.9053**	0.7137	0.6582	0.9711	0.9946	**0.9947**	**0.7900**	0.2223	0.0693
	800	**0.9242**	0.7124	0.6605	0.9703	**0.9943**	0.9934	**0.8435**	0.2190	0.0780

Note: Bold numbers are for the best performance.