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 LASSO methods in this predictive regression framework with mixed degrees of persistence. In the presence of stationary, unit root and cointegrated predictors, we show that the adaptive LASSO asymptotically breaks cointegrated groups although it cannot wipe out all inactive cointegrating variables. This new finding motivates a simple but novel post-selection adaptive LASSO, which we call the *twin adaptive LASSO* (TAlasso), to fix variable selection inconsistency. TAlasso's penalty scheme accommodates the system of heterogeneous regressors, and it recovers the well-known oracle property that implies variable selection consistency and optimal rate of convergence for all three types of regressors. In contrast, conventional LASSO fails to attain coefficient estimation consistency and variable screening in all components simultaneously, since its penalty is imposed according to the marginal behavior of each individual regressor only. We demonstrate the theoretical properties via extensive Monte Carlo simulations. These LASSO-type methods are applied to evaluate short- and long-horizon predictability of S&P 500 excess return.

Key words: cointegration, machine learning, nonstationary time series, shrinkage estimation, variable selection
JEL code: C22, C53, C61

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

†(Corresponding author) 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, Indiana, Purdue, UBC, UW, Duke, KAEA, IPDC, AMES and IAAE conferences for helpful comments. We also thank Bonsoo Koo for sharing the data for the empirical application. All remaining errors are ours.

1 Introduction

Predictive regressions are used extensively in empirical macroeconomics and finance. A leading example is the stock return regression for which predictability has long been a fundamental goal. The first central econometric issue in these models is severe test size distortion in the presence of highly persistent predictors coupled with regression endogeneity. When persistence and endogeneity loom large, the conventional inferential apparatus designed for stationary data can be misleading. Another major challenge in predictive regressions is the low signal-to-noise ratio (SNR). Nontrivial regression coefficients signifying predictability are hard to detect because they are often contaminated by large estimation error. Vast econometric literature is devoted to procedures for valid inference and improved prediction.

Advancement in machine learning techniques—driven by an unprecedented abundance of data sources across many disciplines—offers opportunities for economic data analysis. In the era of big data, shrinkage methods are becoming increasingly popular in econometric inference and prediction thanks to their variable selection and regularization properties. In particular, the *least absolute shrinkage and selection operator* (LASSO; Tibshirani, 1996) has received intensive study in the past two decades.

This paper investigates 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 ones. The more variables the researcher attempts to include, the greater is the need for a data-driven routine for variable selection, since many of these variables *ex post* demonstrate little or no predictability due to the competitive nature of the market. LASSO methods are therefore attractive in predictive regressions as they enable researchers to identify pertinent predictors and exclude irrelevant ones. However, time series in predictive regressions carry heterogeneous degrees of persistence. Some may exhibit short memory (e.g., Treasury bill), whereas others are highly persistent (e.g., most financial/macro predictors). Moreover, a multitude of persistent predictors can be cointegrated. For example, the dividend price 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 among consumption, asset holdings and labor income.

Given the difficulty to classify time series predictors, we keep an agnostic view on the types of the regressors and examine whether LASSO can cope with heterogeneous regressors. We explore the *plain LASSO* (Plasso, henceforth; Tibshirani, 1996), the *standardized LASSO* (Slasso, where the l_1-penalty is multiplied by the sample standard deviation of each regressor) and the *adaptive LASSO* (Alasso; Zou, 2006) with three categories of predictors: short memory (I(0)) regressors, cointegrated regressors, and non-cointegrated unit root (I(1)) regressors.

In this paper, we use the term *variable selection consistency* if a shrinkage method's estimated zeros exactly coincide with the true zero coefficients when the sample size is sufficiently large, and we call *variable screening effect* the phenomenon that an estimator coerces some coefficients to exactly zero. In cross-sectional regressions, Plasso and Slasso achieve coefficient estimation consistency as well as variable screening while Alasso further enjoys variable selection consistency (Zou, 2006). The heterogeneous time series regressors challenge the conventional wisdom. We find that neither Plasso nor Slasso maintains variable screening and consistent coefficient estimation in all components simultaneously, because Plasso imposes equal penalty weight regardless the nature of the regressor while Slasso's penalty only adjusts according to the scale variation of individual regressors but omits their connection via cointegration.

The main contribution of this paper is unveiling and enhancing Alasso in this mixed root context. As we consider a low-dimensional regression, we use OLS as Alasso's initial estimator. Alasso, with a proper choice of the tuning parameter, consistently selects the pure I(0) and I(1) variables, but it may over-select inactive cointegrating variables. The convergence rate of the OLS estimator is $\sqrt{n}$ for the cointegrated variables, and the resulting Alasso penalty weight is too small to eliminate the true inactive cointegrated groups as each cointegrating time series behaves as a unit root process individually. Nevertheless, as is formally stated in Theorem 3.5, with probability approaching one (wpa1) variables forming an inactive cointegrating group cannot all survive Alasso's selection.

Alasso's partially positive result suggests a straightforward remedy to reclaim the desirable oracle property: simply run a second-round Alasso among the variables selected by the first-round Alasso. Because the first-round Alasso has broken the chain that links variables in an inactive cointegration group, these over-selected ones become pure I(1) in the second round model. In the post-selection OLS, the speed of convergence of these variables is boosted from the slow $\sqrt{n}$-rate to the fast n-rate, and then the second-round Alasso can successfully suppress all the inactive coefficients wpa1. We call this post-selection Alasso procedure *twin adaptive LASSO*, and abbreviate it as TAlasso. TAlasso achieves the oracle property (Fan and Li, 2001), which implies the optimal rate of convergence and consistent variable selection under the presence of a mixture of heterogeneous regressors. To our knowledge, this paper is the first to establish these desirable properties in a nonstationary time series context. Notice that the name TAlasso distinguishes it from the *post-selection double LASSO* (Belloni et al., 2014). Double LASSO is named after double inclusion of selected variables in cross-sectional data to correct LASSO's shrinkage bias for uniform statistical inference. In contrast, our TAlasso is double exclusion in predictive regressions with mixed roots. TAlasso gives a second chance to remove remaining inactive variables missed by the first-round Alasso. We do not deal with uniform statistical inference in this paper.

When developing asymptotic theory, we consider a linear process of time series innovations, which encompasses a general ARMA structure arising in many practical applications, e.g. 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 is fixed and the number of time periods n passes to infinity. Our exploration in this paper paves a stepping stone toward automated variable selection in high-dimensional predictive regressions.

In Monte Carlo simulations, we examine in various data generating processes (DGP) the finite sample performance of Alasso/TAlasso in comparison to Plasso/Slasso, assessing their mean squared prediction errors and their variable selection success rates. TAlasso is much more capable in picking out the correct model, which helps with accurate prediction. These LASSO methods are further evaluated in a real data application of the widely used Welch and Goyal (2008) data to predict S&P 500 stock return using 12 predictors. These 12 predictors, to be detailed in Section 5 and plotted in Figure 1, highlight the necessity of considering the three types of heterogeneous regressors. Alasso/TAlasso is shown to be robust in various estimation windows and prediction horizons and attains stronger performance in forecasting.

Literature Review Since Tibshirani (1996)'s original paper of LASSO and Chen, Donoho, and Saunders (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). In econometrics, Caner (2009), Caner and Zhang (2014), Shi (2016), Su, Shi, and Phillips (2016) and Kock and Tang (2019) employ LASSO-type procedures in cross-sectional and panel data models. Belloni and Chernozhukov (2011), Belloni, Chen, Chernozhukov, and Hansen (2012), Belloni, Chernozhukov, and Hansen (2014), Belloni, Chernozhukov, Chetverikov, and Wei (2018) develop methodologies and uniform inferential theories in a variety of microeconometric settings.

Ng (2013) surveys the variable selection methods in predictive regressions. In comparison with the extensive literature in cross-sectional environment, theoretical properties of shrinkage methods are less explored in time series models in spite of great empirical interest in macro and financial applications (Chinco et al., 2019; Giannone et al., 2018; Gu et al., 2019). Medeiros and Mendes (2016) study Alasso in high-dimensional stationary time series. Kock and Callot (2015) discuss LASSO in a vector autoregression (VAR) system. In time series forecasting, 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 independently and identically distributed (iid) orthonormalized predictors to study the risk properties of several machine learning estimators. Papers on LASSO with nonstationary data are even fewer. Caner and Knight (2013) discuss the bridge estimator, a generalization of LASSO, for the augmented Dicky-Fuller test in autoregression, and under the same setup Kock (2016) studies Alasso.

In predictive regressions, Kostakis, Magdalinos, and Stamatogiannis (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. Under a similar setting with high-dimensional I(0) and I(1) regressors and one cointegration group, Koo, Anderson, Seo, and Yao (2019) investigate Plasso's variable estimation consistency as well as the non-standard asymptotic distribution. In a vector error correction model (VECM), Liao and Phillips (2015) use Alasso for cointegration rank selection. While LASSO methods' numerical performance is demonstrated by Smeekes and Wijler (2018) via simulations and empirical examples, we are the first who systematically explore the theory concerning variable selection under mixed regressor persistence.

Notation We use standard notations. We define $\|\cdot\|_1$ and $\|\cdot\|$ as the usual vector l_1- and l_2-norms, respectively. The arrows $\Longrightarrow$ and $\stackrel{p}{\to}$ represent weak convergence and convergence in probability, respectively. $\asymp$ means of the same asymptotic order, and $\sim$ signifies "being distributed as" either exactly or asymptotically, depending on the context. The symbols $O(1)$ and $o(1)$ ($O_p(1)$ and $o_p(1)$) denote (stochastically) asymptotically bounded or negligible quantities. For a generic set M, let $|M|$ be its cardinality. 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. I_p is the $p \times p$ identity matrix, and $I(\cdot)$ is the indicator function.

The rest of the paper is organized as follows. Section 2 introduces 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/TAlasso are developed and compared with those of Plasso/Slasso. The theoretical results are confirmed through a set of empirically motivated simulation designs in Section 4. Finally, we examine 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 simplistic nonstationary regression model. This model helps us understand the technical issues in LASSO arising from nonstationary predictors under 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 is $\beta_n^* = (\beta_{jn}^* = \beta_j^{0*}/\sqrt{n})_{j=1}^p$, where $\beta_j^{0*} \in \mathbb{R}$ is a fixed constant independent of the sample size. β_{jn}^* remains zero regardless of the sample size if $\beta_j^{0*} = 0$, while it varies with n if $\beta_j^{0*} \neq 0$. 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, 2015). Phillips and Lee (2013) and Timmermann and Zhu (2017) specify the local-to-zero parameter as β_j^{0*}/n^δ for some $\delta \in (0, 1)$ but for simplicity here we fix $\delta = 0.5$, the knife-edge balanced case. The $1 \times p$ regressor vector $x_{i\cdot} = (x_{i1}, \ldots, x_{ip})$ follows a pure unit root process

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

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

Assumption 2.1 *The vector of innovations $e_{i\cdot}$ and u_i are generated from*

$$(e_{i\cdot}, u_i)' \sim iid \ \ (0, \Sigma),$$

where $\Sigma = \begin{pmatrix} \Sigma_{ee} & \Sigma_{eu} \\ \Sigma_{eu}' & \sigma_u^2 \end{pmatrix}$ *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, $u = (u_1, \ldots, u_n)'$, $x_j = (x_{1j}, \ldots, x_{nj})'$, 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 ratio). The mixed roots case in Section 3 will be more realistic in practice when multiple predictors are present.

The literature has been focusing on the non-standard statistical inference caused by persistent regressors and weak signals. The asymptotic theory is usually confined to a small number of candidate predictors, but not many. Following the literature of predictive regressions, we consider the asymptotic framework in which p is fixed and the sample size $n \to \infty$. This simple asymptotic framework allows us to concentrate on the contrast between the standard LASSO in iid settings and the predictive regression involving nonstationary regressors.

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

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

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

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

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

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

which is the sum of a (mixed) normal random vector and a non-standard 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, where the associated predictors would be redundant in the regression. Let $M^* = \{j : \beta_j^{0*} \neq 0\}$ be the index set of relevant regressors, $p^* = |M^*|$, 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 parameters by OLS in the set M^* only:

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

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 are useful for variable screening. Next, we study LASSO's asymptotic behavior in predictive regressions with these pure unit root regressors.

2.1 Adaptive LASSO with Unit Root Regressors

Alasso for (1) is defined as

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

where the weight $\widehat{\tau}_j = |\widehat{\beta}_j^{init}|^{-\gamma}$ for some initial estimator $\widehat{\beta}_j^{init}$, 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 primary tuning parameters. We discuss the case of a fixed $\gamma \geq 1$ and $\widehat{\beta}^{init} = \widehat{\beta}^{ols}$.

While 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 regressions with unit root regressors. Let $\widehat{M}^{Alasso} = \{j : \widehat{\beta}_j^{Alasso} \neq 0\}$ be Alasso's estimated active set.

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*

$$\frac{\lambda_n}{n^{1-0.5\gamma}} + \frac{n^{1-\gamma}}{\lambda_n} \to 0,$$

then

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

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

Theorem 2.1 (a) shows that the estimated active set $\widehat{M}^{Alasso}$ coincides with the true active set M^* wpa1, in other words $\widehat{\beta}_j^{Alasso} \neq 0$ if $j \in M^*$ and $\widehat{\beta}_j^{Alasso} = 0$ if $j \in M^{*c}$. (b) indicates that Alasso's asymptotic distribution in the true active set is as if the oracle M^* is known.

In this nonstationary regression, the adaptiveness is maintained through the proper choice of $\widehat{\tau}_j = |\widehat{\beta}_j^{ols}|^{-\gamma}$. On the one hand, when the true coefficient is nonzero, $\widehat{\tau}_j$ delivers a penalty of a negligible order $\lambda_n n^{0.5\gamma-1} \to 0$, recovering the OLS limit theory. On the other hand, if the true coefficient is zero, $\widehat{\tau}_j$ imposes a heavier penalty of the order $\lambda_n n^{\gamma-1} \to \infty$, thereby achieving consistent variable selection. The intuition in Zou (2006, Remark 2) under deterministic design is generalized in our proof to the setting with nonstationary regressors.

Remark 2.2 *Consider the usual formulation of the tuning parameter $\lambda_n = c_\lambda b_n \sqrt{n}$, where c_λ is a constant, and b_n is a decreasing sequence to zero. Substituting this λ_n into the restriction, we have*

$$\frac{b_n}{n^{0.5(1-\gamma)}} + \frac{n^{0.5-\gamma}}{b_n} \to 0.$$

When $\gamma = 1$, the restriction is further simplified to $b_n + \frac{1}{b_n\sqrt{n}} \to 0$. A slowly shrinking sequence, such as $b_n = (\log\log n)^{-1}$ which is commonly imposed in the Alasso literature in cross section regressions, satisfies this rate condition.

Given the positive results about Alasso with unit root regressors, we continue to study Plasso and Slasso.

2.2 Plain LASSO with Unit Roots

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

$$\widehat{\beta}^{Plasso} = \arg\min_{\beta} \left\{ \|y - X\beta\|_2^2 + \lambda_n \|\beta\|_1 \right\}. \tag{7}$$

Plasso is a special case of the penalized estimation $\min_\beta \left\{ \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^p \widehat{\tau}_j |\beta_j| \right\}$ 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 $D : \mathbb{R}^3 \mapsto \mathbb{R}$ as $D(s, v, \beta) = s\left[v \cdot \mathrm{sgn}(\beta) I(\beta \neq 0) + |v| I(\beta = 0)\right].$

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

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

(c) If $\lambda_n/n \to \infty$, and $\lambda_n/n^{3/2} \to 0$,

$$\frac{1}{\lambda_n} n^2(\widehat{\beta}^{Plasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega v + \sum_{j=1}^p D(1, v_j, \beta_j^{0*}) \right\}.$$

Remark 2.4 *Corollary 2.3 echoes, but is different from, Zou (2006, Section 2). With unit root regressors, (a) implies that the conventional tuning parameter $\lambda_n \asymp \sqrt{n}$ is too small for variable screening. Such a choice produces an asymptotic distribution the same as that of OLS. (b) shows that the additional term $c_\lambda \sum_{j=1}^p D(1, v_j, \beta_j^{0*})$ affects 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 Plasso in the cross-sectional setting, there is no guarantee for consistent variable selection. 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 $\arg\min_v \left\{ v'\Omega v + \sum_{j=1}^p D(1, v_j, \beta_j^{0*}) \right\}$ is non-degenerate due to the randomness of Ω. This is in sharp contrast to Zou (2006)'s Lemma 3, where the Plasso estimator there degenerates to a constant. The distinction arises because in our context Ω is a non-degenerate distribution, whereas in Zou (2006)'s iid setting the counterpart of Ω is a non-random matrix.*

2.3 Standardized LASSO with Unit Roots

Plasso is scale-variant in the sense that if we change the unit of x_j by multiplying it with a nonzero 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

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

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 and $\bar{x}_j$ is the sample average. In this paper, we call (8) the *standardized LASSO*, and abbreviate it as Slasso. Such standardization is the default option for LASSO in many statistical packages, for example the `R` package `glmnet`. Notice that Alasso is also scale-invariant with $\widehat{\beta}^{init} = \widehat{\beta}^{ols}$ and $\gamma = 1$.

Slasso is another special case of $\min_\beta \left\{ \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^p \widehat{\tau}_j |\beta_j| \right\}$ 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. Standardization does not alter the convergence rate of the estimator. In contrast, when x_j has a unit root, from (4) we have

$$\frac{\widehat{\sigma}_j}{\sqrt{n}} = \sqrt{\frac{1}{n^2}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} \Longrightarrow \sqrt{\int_0^1 B_{x_j}^2(r)dr - \left(\int_0^1 B_{x_j}(r)\, dr\right)^2} := d_j, \tag{9}$$

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. Adopting a standard argument for LASSO as in Knight and Fu (2000) and Zou (2006), we have the following asymptotic distribution for $\widehat{\beta}^{Slasso}$.

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

$$n(\widehat{\beta}^{Slasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega v - 2v'\zeta + c_\lambda \sum_{j=1}^{p} D(d_j, v_j, \beta_j^{0*}) \right\}.$$

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

$$\frac{n^{3/2}}{\lambda_n}(\widehat{\beta}^{Slasso} - \beta_n^*) \Longrightarrow \arg\min_v \left\{ v'\Omega v + \sum_{j=1}^{p} D(d_j, v_j, \beta_j^{0*}) \right\}.$$

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 $\widehat{\sigma}_j$. When we increase λ_n as in (b), the term $\sum_{j=1}^{p} D(d_j, v_j, \beta_j^{0*})$ will generate the variable screening effect under the usual choice of tuning parameter $\lambda_n \asymp \sqrt{n}$. In contrast, its counterpart $\sum_{j=1}^{p} D(1, v_j, \beta_j^{0*})$ emerges in Corollary 2.3 when $\lambda_n \asymp n$. While the first argument of $D(\cdot, v_j, \beta_j^{0*})$ in Plasso is 1, in Slasso it is replaced by the random d_j which introduces an extra source of uncertainty in variable screening. Again, Slasso has no mechanism to secure consistent variable selection. A larger λ_n as in (c) 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 raise the tuning parameter λ_n up to the order of n. For Slasso, although $\lambda_n \asymp \sqrt{n}$ is sufficient for variable screening, $\widehat{\sigma}_j$ affects variable screening with extra randomness.

The unit root regressors are shown to alter the asymptotic properties of the 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/or cointegrated. In the following section, we will discuss the theoretical properties of LASSO under a mixed persistence environment.

3 LASSO Theory with Mixed Roots

We extend the model in Section 2 to accommodate I(0) and I(1) regressors with possible cointegration among the latter. The LASSO theory in this section will provide a general guidance for multivariate predictive regressions in practice.

3.1 Model

We introduce three types of predictors into the model. Suppose a $1 \times p_c$ cointegrated system $x_{i\cdot}^c = (x_{i1}^c, ..., x_{ip_c}^c)$ has cointegration rank p_1 so that $p_2 = p_c - p_1$ is the number of unit roots in this cointegration system. Let $x_{i\cdot}^c$ admit a triangular representation

$$
\begin{aligned}
\underset{p_1 \times p_c}{A} x_{i\cdot}^{c\prime} = x_{1i\cdot}^{c\prime} - \underset{p_1 \times p_2}{A_1} x_{2i\cdot}^{c\prime} &= \underset{p_1 \times 1}{v_{1i\cdot}^{\prime}}, \quad (10) \\
\Delta x_{2i\cdot}^c &= v_{2i\cdot},
\end{aligned}
$$

where $A = [I_{p_1}, -A_1]$, $x_{i\cdot}^c = (\underset{1\times p_1}{x_{1i\cdot}^c}, \underset{1\times p_2}{x_{2i\cdot}^c})$ and the vector $v_{1i\cdot}$ is the cointegrating residual. The triangular representation (Phillips, 1991) is a convenient and general form of a cointegrated system, and Xu (2018) has recently used this structure in predictive regressions.

Assume that y_i is generated from the linear model

$$
y_i = \sum_{l=1}^{p_z} z_{il}\alpha_l^* + \sum_{l=1}^{p_1} v_{1il}\phi_{1l}^* + \sum_{l=1}^{p_x} x_{il}\beta_l^* + u_i = z_{i\cdot}\alpha^* + v_{1i\cdot}\phi_1^* + x_{i\cdot}\beta^* + u_i. \quad (11)
$$

Each time series $x_l = (x_{1l}, ..., x_{nl})'$ is a unit root process (initialized at zeros for simplicity) as in Section 2, and each z_l is a stationary regressor. This is an *infeasible* equation since the cointegrating residual $v_{1i\cdot}$ is unobservable without *a priori* knowledge about the cointegration relationship. What we observe is the vector $x_{i\cdot}^c$ that contains cointegration groups. Substituting (10) into (11) we obtain a *feasible* regression equation

$$
y_i = z_{i\cdot}\alpha^* + x_{1i\cdot}^c\phi_1^* + x_{2i\cdot}^c\phi_2^* + x_{i\cdot}\beta^* + u_i = z_{i\cdot}\alpha^* + x_{i\cdot}^c\phi^* + x_{i\cdot}\beta^* + u_i \quad (12)
$$

where $\phi_2^* = -A_1'\phi_1^*$ and $\phi^* = (\phi_1^{*\prime}, \phi_2^{*\prime})'$. Stacking the sample of n observations, the infeasible and the feasible regressions can be written as

$$
\begin{aligned}
y &= Z\alpha^* + V_1\phi_1^* + X\beta^* + u \quad (13) \\
&= Z\alpha^* + X_1^c\phi_1^* + X_2^c\phi_2^* + X\beta^* + u \\
&= Z\alpha^* + X^c\phi^* + X\beta^* + u, \quad (14)
\end{aligned}
$$

where the variable V_1 is the $n \times p_1$ matrix that stacks $(v_{1i\cdot})_{i=1}^n$, and X_1^c, X_2^c, X^c, Z and X are defined similarly. We explicitly define $\alpha^* = \alpha^{0*}$ and $\phi^* = \phi^{0*}$ as two coefficients independent of the sample size, which are associated with Z and X^c, respectively. To keep the two sides of (11) balanced, as in Section 2 we specify a local-to-zero sequence $\beta^* = \beta_n^* = (\beta_l^{0*}/\sqrt{n})_{l=1}^{p_x}$ where β_l^{0*} is invariant to the sample size.

3.2 OLS theory with mixed roots

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

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

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

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

where $F_0 = I_{p+1}$, $\sum_{k=0}^{\infty} k \|F_k\| < \infty$, $F(x) = \sum_{k=0}^{\infty} F_k x^k$ *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}. On the other hand, 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 matrix associated with the innovation vector, where $F(1) = \left(F_z'(1), F_v'(1), F_e'(1), F_u(1)\right)'$. Moreover, define the sum of one-sided autocovariance as $\Lambda = \sum_{h=1}^{\infty} \mathbb{E}\left(\xi_i \xi_{i-h}'\right)$, and $\Delta = \Lambda + \mathbb{E}\left(\xi_i \xi_i'\right)$. We use the functional law (Phillips and Solo, 1992) under Assumption 3.1 to derive

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

We "extend" the I(0) regressors as $Z^+ = [Z, V_1]$ and the I(1) regressors as $X^+ := [X_2^c, X]$. Let the $(p+1)\times(p+1)$ matrix

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

according to the explicit form of Σ_ε. Then the left-top $p \times p$ submatrix of Ω, which we denote as

$[\Omega]_{p\times p}$, can be represented in a conformable manner as

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

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

$$Z^{+\prime}u/\sqrt{n} \implies \zeta_{z^+} \sim N\left(0, \Sigma_{uu}\Omega_{zz}^{+}\right) \tag{15}$$

$$X^{+\prime}u/n \implies \zeta_{x^+} \sim \int_0^1 B_{x^+}(r)dB_{\varepsilon}(r)'F_u(1) + \Delta_{+u} \tag{16}$$

where the one-sided long-run covariance matrix $\Delta_{+u} = \sum_{h=0}^{\infty} \mathbb{E}\left(\tilde{u}_i u_{i-h}\right)$ with $\tilde{u}_i = (v_{2i\cdot}, e_{i\cdot})'$. See Appendix Section A.2 for the derivation of (16).

We first study the asymptotic behavior of OLS, the initial estimator for Alasso. For rotational convenience, define the observed predictor matrix $\underset{n\times p}{W} = [\underset{n\times p_z}{Z}, \underset{n\times p_c}{X^c}, \underset{n\times p_x}{X}]$ and the associated true coefficients $\theta_n^* = \left(\alpha^{0*\prime}, \phi_1^{0*\prime}, \phi_2^{0*\prime}, \beta_n^{*\prime}\right)'$ where $\phi_2^{0*} = -A_1'\phi_1^{0*}$. We establish the asymptotic distribution of the OLS estimator

$$\widehat{\theta}^{ols} = (W'W)^{-1}W'y.$$

To state the result, we define a diagonal normalizing matrix $R_n = \begin{pmatrix} \sqrt{n}\cdot I_{p_z+p_1} & 0 \\ 0 & n\cdot I_{p_2+p_x} \end{pmatrix}$ and a rotation matrix $Q = \begin{pmatrix} I_{p_z} & 0 & 0 & 0 \\ 0 & I_{p_1} & 0 & 0 \\ 0 & A_1' & I_{p_2} & 0 \\ 0 & 0 & 0 & I_{p_x} \end{pmatrix}$.

Theorem 3.2 *If the linear model (12) satisfies Assumption 3.1, then*

$$R_nQ\left(\widehat{\theta}^{ols} - \theta_n^*\right) = \begin{pmatrix} \sqrt{n}(\widehat{\alpha}^{ols} - \alpha^{0*}) \\ \sqrt{n}(\widehat{\phi}_1^{ols} - \phi_1^{0*}) \\ n(A_1'\widehat{\phi}_1^{ols} + \widehat{\phi}_2^{ols}) \\ n(\widehat{\beta}^{ols} - \beta_n^*) \end{pmatrix} \implies \left(\Omega^{+}\right)^{-1}\zeta^{+}. \tag{17}$$

where $\Omega^{+} = \begin{pmatrix} \Omega_{zz}^{+} & 0 \\ 0 & \Omega_{xx}^{+} \end{pmatrix}$ *and* ζ^{+} *is the limiting distribution of* $\begin{pmatrix} Z^{+\prime}u/\sqrt{n} \\ X^{+\prime}u/n \end{pmatrix}$.

Remark 3.3 *In the rotated coordinate system, (17) and the definition of* ζ_{x^+} *imply 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 as* $Q\left(\widehat{\theta}^{ols} - \theta_n^*\right) = O_p(\mathrm{diag}(R_n^{-1}))$.

Remark 3.4 *Since we keep an agnostic view about the identities of the stationary, unit root and cointegrated regressors, Theorem 3.2 is not useful for statistical inference as we do not know which coefficients converge at $\sqrt{n}$-rate and which at n-rate. (17) shows that with the help of the rotation Q the estimators $\widehat{\phi}_1^{ols}$ and $\widehat{\phi}_2^{ols}$ are tightly connected in the sense $A_1'\widehat{\phi}_1^{ols}+\widehat{\phi}_2^{ols}=O_p\left(n^{-1}\right)$. Without the rotation that is unknown in practice, the OLS components associated with the stationary and the cointegration system converge at $\sqrt{n}$ rate whereas only those associated with the pure unit root converge at n-rate:*

$$\begin{pmatrix}\sqrt{n}(\widehat{\theta}^{ols}-\theta_n^{0*})_{\mathcal{I}_0\cup\mathcal{C}}\\ n(\widehat{\theta}^{ols}-\theta_n^{0*})_{\mathcal{I}_1}\end{pmatrix}=\begin{pmatrix}\sqrt{n}(\widehat{\alpha}^{ols}-\alpha^{0*})\\ \sqrt{n}(\widehat{\phi}_1^{ols}-\phi_1^{0*})\\ \sqrt{n}(\widehat{\phi}_2^{ols}-\phi_2^{0*})\\ n(\widehat{\beta}^{ols}-\beta_n^{*})\end{pmatrix}\Longrightarrow\begin{pmatrix}I_{p_z}&0&0&0\\0&I_{p_1}&0&0\\0&-A_1'&0&0\\0&0&0&I_{p_x}\end{pmatrix}\left(\Omega^{+}\right)^{-1}\zeta^{+}. \tag{18}$$

Although in isolation each variable in the cointegration system appears as a unit root time series, the presence of the cointegration makes those in the group highly correlated and the high correlation reduces their rate of convergence from n to $\sqrt{n}$. This effect is analogous to the deterioration of convergence rate for nearly perfectly collinear regressors.

3.3 Adaptive LASSO with mixed roots

Similarly to Section 2.1, we define Alasso under (14) as

$$\widehat{\theta}^{Alasso}=\arg\min_{\theta}\left\{\|y-W\theta\|_2^2+\lambda_n\sum_{j=1}^{p}\widehat{\tau}_j|\theta_j|\right\}, \tag{19}$$

where $\widehat{\tau}_j=|\widehat{\theta}_j^{ols}|^{-\gamma}$.[1] The literature on Alasso has established the oracle property in many models. Caner and Knight (2013) and Kock (2016) study Alasso's rate adaptiveness in a 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.

Surprisingly, Theorem 3.5 will show that in the mixed root model Alasso's oracle inequality holds only partially but not for all regressors. To discuss variable selection in this context, we introduce the following notations. We partition the index set of all regressors $\mathcal{M}=\{1,\ldots,p\}$ into four components $\mathcal{I}_0$ (I(0) variables, associated with $z_{i\cdot}$), $\mathcal{C}_1$ (associated with $x_{1i\cdot}^c$), $\mathcal{C}_2$ (associated with $x_{2i\cdot}^c$) and $\mathcal{I}_1$ (I(1) variables, associated with $x_{i\cdot}$). Let $\mathcal{C}=\mathcal{C}_1\cup\mathcal{C}_2$ and $\mathcal{I}=\mathcal{I}_0\cup\mathcal{I}_1$. Let $M^*=\left\{j:\theta_j^{0*}\neq 0\right\}$ be

[1] 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 as those of OLS.

the true active set for the feasible representation, and $\widehat{M}^{Alasso} = \{j : \widehat{\theta}_j^{Alasso} \neq 0\}$ be the estimated active set. Next, let $\mathcal{M}_Q = \mathcal{I} \cup \mathcal{C}_1$ be the set of coordinates that are invariant to the rotation in Q. Similarly, let $M_Q^* = M^* \cap \mathcal{M}_Q$ be the active set in the infeasible regression equation (11), and $M_Q^{*c} = \mathcal{M} \backslash M_Q^*$ be the corresponding inactive set. The DGP (13) obviously implies $\mathcal{C}_2 \cap M^* = \emptyset$ and $\mathcal{C}_2 \subseteq M_Q^{*c}$. For a generic set $M \subseteq \mathcal{M}$, let $\mathrm{CoRk}\,(M)$ be the cointegration rank of the variables in M.

Theorem 3.5 *Suppose that the linear model (12) satisfies Assumption 3.1, and for all the coefficients $j \in \mathcal{C}_2$ we have*

$$\phi_{2j}^{0*} \neq 0 \quad \text{if} \quad \sum_{l=1}^{p_1} \left| A_{1lj} \phi_{1l}^{0*} \right| \neq 0. \tag{20}$$

If the tuning parameter λ_n is chosen such that $\lambda_n \to \infty$ and

$$\frac{\lambda_n}{n^{1-0.5\gamma}} + \frac{n^{0.5(1-\gamma)}}{\lambda_n} \to 0, \tag{21}$$

then

(a) Consistency and asymptotic distribution:

$$(R_n Q(\widehat{\theta}^{Alasso} - \theta_n^*))_{M_Q^*} \Longrightarrow (\Omega^+_{M_Q^*})^{-1} \zeta^+_{M_Q^*} \tag{22}$$

$$(R_n Q(\widehat{\theta}^{Alasso} - \theta_n^*))_{M_Q^{*c}} \xrightarrow{p} 0. \tag{23}$$

(b) Partial variable selection consistency:

$$P\left(M^* \cap \mathcal{I} = \widehat{M}^{Alasso} \cap \mathcal{I}\right) \to 1, \tag{24}$$

$$P\left((M^* \cap \mathcal{C}) \subseteq (\widehat{M}^{Alasso} \cap \mathcal{C})\right) \to 1, \tag{25}$$

$$P\left(\mathrm{CoRk}(M^*) = \mathrm{CoRk}(\widehat{M}^{Alasso})\right) \to 1. \tag{26}$$

Remark 3.6 *Condition (20) is an extra assumption that rules out the pathological case that some nonzero elements in $A_{1lj}\phi_{1l}^{0*}$, $l = 1, \ldots, p_1$, happen to exactly cancel out one another and render ϕ_{2j}^{0*} to be inactive. In other words, it makes sure that if an x_j^c is involved in more than one active cointegration group, it must be active in (12). This condition holds in general, as it is violated only under very specific configuration of ϕ_1^{0*} and A_1. For instance, in the demonstrative example in Table 1, the condition breaks down if x_7^c's true coefficient $\phi_7^{0*} = \spadesuit_{71} \times \bigstar_1 + \spadesuit_{72} \times \bigstar_2 = 0$.*

Were we informed of the oracle about the true active variables in M^* and the cointegration matrix A_1, we would transform the cointegrated variables into cointegrating residuals, discard the inactive ones and then run OLS. That is, ideally we would conduct estimation with variables in M_Q^* only. Such an oracle OLS shares the same asymptotic distribution as the Alasso counterpart in (22).

Table 1: Diagram of a cointegrating system in predictors

$(\phi_1^{0*}, \phi_2^{0*})$	$\bigstar_1$	$\bigstar_2$	0	0	0	$\clubsuit_6$	$\clubsuit_7$	0	0	coint.	
$(\mathcal{C}_1, \mathcal{C}_2)$	x_1^c	x_2^c	x_3^c	x_4^c	x_5^c	x_6^c	x_7^c	x_8^c	x_9^c	resid.	ϕ_1^{0*}
	1	0	0	0	0	$\spadesuit_{61}$	$\spadesuit_{71}$	0	0	v_1	$\bigstar_1$
	0	1	0	0	0	0	$\spadesuit_{72}$	0	0	v_2	$\bigstar_2$
$(I_{p_1}, -A_1')$	0	0	1	0	0	0	0	$\spadesuit_{83}$	0	v_3	0
	0	0	0	1	0	$\spadesuit_{64}$	0	$\spadesuit_{84}$	$\spadesuit_{94}$	v_4	0
	0	0	0	0	1	0	$\spadesuit_{75}$	0	0	v_5	0

Note: The diagram represents a cointegration system of 9 variables $(x_j^c)_{j=1}^9$ of cointegrating rank 5. The last column represents the coefficients in ϕ_1^{0*}, with $\bigstar$ as a nonzero entry. In the matrix $-A_1'$, $\spadesuit$ is nonzero, and the ones in gray cells are irrelevant to the value $\phi_2^{0*} = -A_1'\phi_1^{0*}$ no matter zero or not. The first row displays the coefficients ϕ_1^{0*} (the same as in the last column, and $\bigstar$ for non-zeros) and ϕ_2^{0*} ($\clubsuit$ for non-zeros). In this example, $\left(x_j^c\right)_{j=1,2,6,7}$ are in the active set M^*. There are two active cointegration groups (x_1^c, x_6^c, x_7^c) and (x_2^c, x_7^c), and three inactive cointegration groups (x_3^c, x_8^c), $(x_4^c, x_6^c, x_8^c, x_9^c)$ and (x_5^c, x_7^c).

Outside the active set M_Q^*, (23) shows that all other (transformed) variables in M_Q^{*c} consistently converge to zero.

Remark 3.7 *The variable selection results in Theorem 3.5 are novel and interesting. (24) shows variable selection is consistent for the pure I(0) and I(1) variables, which is in line with the well-known oracle property of Alasso. However, instead of confirming the oracle property, (25) pronounces that in the cointegration set* $\mathcal{C}$ *the selected* $\widehat{M}^{Alasso}$ *asymptotically contains the true active ones in* M^* *but Alasso may over-select inactive ones. The omission stems from the mismatch between the convergence rate of the initial estimator and the marginal behavior of a cointegrated variable viewed in isolation. Consider a pair of inactive cointegrating variables, such as* (x_3^c, x_8^c) *in Table 1. The unknown cointegration relationship precludes transforming this pair into the cointegrating residual* v_3*. Without the rotation, OLS associated with the pair can only achieve* $\sqrt{n}$ *rate according to (18). The resulting penalty weight is insufficient to remove these variables that* individually *appear as I(1). In consequence, Alasso fails to eliminate both* x_3^c *and* x_8^c *wpa1. To the best of our knowledge, this is the first case of Alasso's variable selection inconsistency in an empirically relevant model.*

Remark 3.8 *Under condition (20) all variables in active cointegration groups have nonzero coefficients and Alasso selects them asymptotically according to (25). Despite potential variable over-selection in* $\mathcal{C}$*, (26) brings relief: in the limit the cointegration rank in Alasso's selected set,* $\mathrm{CoRk}(\widehat{M}^{Alasso})$*, must equal the active cointegration rank* $\mathrm{CoRk}(M^*)$*. Notice that under our agnostic perspective we do not need to know or use testing procedures to determine the value of* $\mathrm{CoRk}(M^*)$*.*

Let $C^{*c} = M^{*c} \cap \mathcal{C}$ *be the index set of inactive variables in* $\mathcal{C}$*. (25) implies that variables in* $C^{*c} \cap \widehat{M}^{Alasso}$*—Alasso's mistakenly selected inactive variables—cannot form cointegration groups. In mathematical expression,*

$$P\left(\mathrm{CoRk}\left(C^{*c} \cap \widehat{M}^{Alasso}\right) = 0\right) \to 1.$$

Let us again take (x_3^c, x_8^c) in Table 1 as an example. (26) indicates that Alasso is at least partially effective in that it prevents the inactive x_3^c and x_8^c from entering $\widehat{M}^{Alasso}$ simultaneously. It kills at least one variable in the pair to break the cointegration relationship. The intuition is as following. Suppose (x_3^c, x_8^c) are both selected. In the predictive regression this pair, due to coefficient estimation consistency, together behaves like the cointegrating residual v_3, which is an I(0). Alasso will not tolerate this inactive v_3 by paying a penalty on x_3 and x_8, because these coefficients are subject to a penalty weight $\widehat{\tau}_j = 1/O_p\left(n^{-1/2}\right)$, which is of the same order as the pure I(0) variables. Recall that Alasso removes all inactive pure I(0) variables when $n \to \infty$ under the same level of penalty weight.

Similar reasoning applies in another inactive cointegration group $\left(x_{5,}^c, x_7^c\right)$. In the regression x_5^c is inactive but x_7^c is active as it is involved in the active cointegration groups (x_1^c, x_6^c, x_7^c) and (x_2^c, x_7^c). While x_7^c is selected wpa1, suppose x_5^c is selected as well. Then x_5^c's contribution to the predictive regression would be equivalent to the I(0) cointegrating residual v_5. Its corresponding penalty is of the order $1/O_p\left(n^{-1/2}\right)$, which is sufficient to kick out the inactive v_5. Therefore, x_5 cannot survive Alasso's variable selection either.

The intuition gleaned in these examples can be generalized to cointegration relationship involving more than two variables and multiple cointegrated groups. The proof of Theorem 3.5 formalizes this argument by inspecting a linear combination of the corresponding Karush-Kuhn-Tucker condition for the selected variables.

In the literature Alasso always achieves oracle property by a single implementation so there is no point to run it twice. In our predictive regression with mixed roots, a single Alasso is likely to over-select inactive variables in the cointegration system. Given that the cointegrating ties are all shattered wpa1 in (26), it hints further action to fulfill the oracle property.

When the sample size is sufficiently large, with high probability those mistakenly selected inactive variables have no cointegration relationship, so they behave as pure unit root processes in the post-selection regression equation of y_i on $(W_{ij})_{j\in\widehat{M}^{Alasso}}$. This observation suggests running another Alasso—a post-selection Alasso. We obtain the post-Alasso initial OLS estimator

$$\widehat{\theta}^{postols} = \left(W'_{\widehat{M}^{Alasso}} W_{\widehat{M}^{Alasso}}\right)^{-1} W'_{\widehat{M}^{Alasso}} y.$$

The post-selection OLS estimator $\widehat{\theta}_j^{postols} = O_p\left(n^{-1}\right)$ for those over-selected inactive cointegrated variables $j \in C^{*c} \cap \widehat{M}^{Alasso}$, instead of $O_p\left(n^{-1/2}\right)$ as for the first-around initial $\widehat{\theta}_j^{ols}$. The resulting penalty level is heavy enough to wipe out these redundant variables in another round of post-selection Alasso, which we call TAlasso:

$$\widehat{\theta}^{TAlasso} = \arg\min_{\theta}\left\{\|y - \sum_{j\in\widehat{M}^{Alasso}} w_j\theta_j\|_2^2 + \lambda_n \sum_{j\in\widehat{M}^{Alasso}} \widehat{\tau}_j^{post}|\theta_j|\right\}, \tag{27}$$

where $\widehat{\tau}_j^{post} = |\widehat{\theta}_j^{postols}|^{-\gamma}$ is the new penalty weight. TAlasso asymptotic reclaims variable selection consistency for all types of variables.

Theorem 3.9 *Under the same assumptions and the same rate for λ_n as in Theorem 3.5, the TA-lasso estimator $\widehat{\theta}^{TAlasso}$ satisfies*

(a) Asymptotic distribution: $(R_n Q(\widehat{\theta}^{TAlasso} - \theta_n^*))_{M^*} \Longrightarrow (\Omega^+_{M^*})^{-1}\zeta^+_{M^*}$;

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

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 without potential testing error (Smeekes and Wijler, 2020). Our research provides a valuable guidance for practice. TAlasso is the first estimator that achieves the desirable oracle property without requiring prior knowledge on the persistence of multivariate regressors. Despite intensive study of LASSO estimation in recent years, Theorems 3.5 and 3.9 are eye openers. With the cointegration system in the predictors, the former shows that Alasso does not automatically adapt to the behavior of a *system* of regressors. Nevertheless, it at least breaks all redundant cointegration groups so its flaw can be easily rescued by another round of Alasso. This solution echoes the repeated implementation of a machine learning procedure as in Phillips and Shi (2019).

On the other hand, the weight $\widehat{\tau}_j$ in $\min_\theta \left\{\|y - W\theta\|_2^2 + \lambda_n \sum_{j=1}^p \widehat{\tau}_j |\theta_j|\right\}$ is a constant for Plasso, or it exploits merely the marginal variation of x_j for Slasso. As will be shown in the following subsection, such weighting mechanisms 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. The penalty of Alasso/TAlasso pays heed to the cointegration system thanks to the initial/post-selection OLS, but Plasso or Slasso does not.

3.4 Conventional LASSO with mixed roots

We now study the asymptotic theory of Plasso

$$\widehat{\theta}^{Plasso} = \arg\min_{\theta} \left\{\|y - W\theta\|_2^2 + \lambda_n \|\theta\|_1\right\}. \tag{28}$$

Corollary 3.10 *Suppose the linear model (12) satisfies Assumption 3.1.*

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

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

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

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

$$\frac{1}{\lambda_n} R_n Q(\widehat{\theta}^{Plasso} - \theta_n^*) \quad \Longrightarrow \quad \arg\min_v \left\{ v'\Omega^+ v + \sum_{j \in \mathcal{I}_0 \cup \mathcal{C}} D\left(1, v_j, \theta_j^{0*}\right) \right\}.$$

In Corollary 3.10 (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. The screening effect kicks in when the tuning parameter λ_n gets bigger. In view of the OLS rate of convergence in (18), we can call those θ_j associated with $\mathcal{I}_0 \cup \mathcal{C}$ the *slow coefficients* (at rate $\sqrt{n}$) and those associated with $\mathcal{I}_1$ the *fast coefficients* (at rate n). When λ_n is raised to the magnitude in (b), the term $D\left(1, v_j, \theta_j^{0*}\right)$ strikes variable screening among the slow coefficients but not the fast coefficients. If we further increase λ_n to the level of (c), then the convergence rate of the slow coefficients is dragged down by the large penalty but still there is no variable screening effect for the fast coefficients. In order to induce variable screening in $\widehat{\theta}_{\mathcal{I}_1}^{Plassso}$, the tuning parameter must be ballooned to $\lambda/n \to c_\lambda \in (0, \infty]$, but the consistency of the slow coefficients would collapse under such a disproportionately heavy λ_n. The result in Corollary 3.10 reveals a major drawback of Plasso in the mixed root model. Since it has one uniform penalty level for all variables, it is not adaptive to these various types of predictors.

Let us now turn to Slasso

$$\widehat{\theta}^{Slasso} = \arg\min_{\theta} \left\{ \|y - W\theta\|_2^2 + \lambda_n \sum_{j=1}^{p} \widehat{\sigma}_j \left|\theta_j\right| \right\}. \tag{29}$$

The I(0) regressors are accompanied by $\widehat{\sigma}_j = O_p(1)$, while for $j \in \mathcal{C} \cup \mathcal{I}_1$ the individually nonstationary regressors are coupled with $\widehat{\sigma}_j = O_p(\sqrt{n})$.

Corollary 3.11 *Suppose the linear model (12) satisfies Assumption 3.1.*

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

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

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

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

$$\frac{1}{\lambda_n} R_n Q(\widehat{\theta}^{Slasso} - \theta_n^*) \quad \Longrightarrow \quad \arg\min_{v} \left\{ v'\Omega^+ v + \sum_{j\in\mathcal{C}} D\left(d_j, v_j, \theta_j^{0*}\right) \right\}.$$

Remark 3.12 *The tuning parameter λ_n in Corollary 3.11(a) is $\sqrt{n}$-order smaller than that in Corollary 3.10(a) to produce the same asymptotic distribution as OLS. The distinction of Plasso and Slasso arises from the coefficients in the set $\mathcal{C}$. Their corresponding penalty terms have the multipliers $\widehat{\sigma}_j = O_p(\sqrt{n})$, instead of the desirable $O_p(1)$ that is suitable for their slow convergence rate under OLS. In other words, the penalty level is overly heavy for these parameters. The overwhelming penalty level incurs variable screening effect in (b) as soon as $\lambda_n \to c_\lambda \in (0, \infty)$. Moreover, (c) implies that for the consistency of $\widehat{\phi}^{Plasso}$ the tuning parameter λ_n must be small enough in the sense $\lambda_n/\sqrt{n} \to 0$; otherwise they will be inconsistent. In both (b) and (c) the penalty term $D\left(d_j, v_j, \theta_j^{0*}\right)$*

only screens those variables in $\mathcal{C}$. Although there is no variable screening effect for the component $\mathcal{I}$, its associated Slasso estimator differs in terms of asymptotic distribution from the OLS counterpart and can be expressed only in the argmin form due to the non-block-diagonal Ω_{zz}^{+} and Ω_{xx}^{+}. Similar to Plasso, simultaneously for all components Slasso cannot achieve consistent parameter estimation and variable screening.

To sum up this section, in the general model with various types of regressors, Alasso only partially maintains the oracle property under the standard choice of the tuning parameter but the oracle property can be fully restored by TAlasso. In contrast, Plasso using a single tuning parameter does not adapt to the different order of magnitudes of the slow and fast coefficients. Slasso suffers from overwhelming penalties for those coefficients associated with the cointegration groups.

4 Simulations

In this section, we examine via simulations the performance of the LASSO methods in forecasting as well as variable screening. We consider different sample sizes to demonstrate the approximation of the asymptotic theory in finite samples. Comparison is based on the one-period-ahead out-of-sample forecast.

4.1 Simulation Design

Following the settings in Sections 2 and 3, we consider three DGPs.

DGP 1 (Pure unit roots). This DGP corresponds to the pure unit root model in Section 2. Consider a linear model with eight unit root predictors, $x_i = (x_{ij})_{j=1}^{8}$ where each x_{ij} is drawn from independent random walk $x_{ij} = x_{i-1,j} + e_{ij}$, $e_{ij} \sim$ iid $N(0,1)$ across i and j. The dependent variable y_i is generated by $y_i = \gamma^* + x_i\beta_n^* + u_i$ where the intercept $\gamma^* = 0.25$, and $\beta_n^* = (1,1,1,1,0,0,0,0)'/\sqrt{n}$. The idiosyncratic error $u_i \sim$ iid $N(0,1)$, and so does those u_i's in DGPs 2 and 3.

DGP 2 (Mixed roots and cointegration). This DGP is designed for the mixed root model in Section 3. The dependent variable

$$y_i = \gamma^* + \sum_{l=1}^{2} z_{il}\alpha_l^* + \sum_{l=1}^{4} x_{il}^c\phi_{ln}^* + \sum_{l=1}^{2} x_{il}\beta_{ln}^* + u_i,$$

where $\gamma^* = 0.3$, $\alpha^* = (0.4, 0)$, $\phi^* = (0.3, -0.3, 0, 0)$, and $\beta_n^* = (1/\sqrt{n}, 0)$. The stationary regressors z_{i1} and z_{i2} follow two independent AR(1) processes with the same AR(1) coefficient 0.5; $z_{il} = 0.5z_{i-1,l} + e_{il}$, $e_{il} \sim$ iid $N(0,1)$. $x_{i\cdot}^c \in \mathbb{R}^4$ is an vector I(1) process with cointegration rank 2 based on the VECM, $\Delta x_{i\cdot}^c = \Gamma'\Lambda x_{i-1,\cdot}^c + e_i$, 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 $e_{i\cdot} = (e_{il})_{l=1}^{4}$, we set $e_{i2} = e_{i1} - \nu_{i1}$ and $e_{i4} = e_{i3} - \nu_{i2}$, where ν_{i1} and ν_{i2} 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.

DGP 3 (Stationary autoregression). The autoregressive distributed lag (ARDL) model is a classical specification for time series regressions. In addition to including lags of y_i, it is common to accommodate lags of predictors in predictive regressions, for example Medeiros and Mendes (2016). The stationary dependent variable in the following equation is generated from the ARDL model

$$y_i = \gamma^* + \rho^* y_{i-1} + \sum_{l=1}^{4} \phi_{ln}^* x_{il}^c + \beta_{1n}^* x_i + \beta_{2n}^* x_{i-1} + \sum_{l=1}^{3} (\alpha_{l1}^* z_{il} + \alpha_{l2}^* z_{i-1,l}) + u_i$$

where $\gamma^* = 0.3$, $\rho^* = 0.4$, $\phi^* = (0.75, -0.75, 0, 0)$, $\beta_n^* = (1.5/\sqrt{n}, 0)$, $\alpha_1^* = (0.6, 0.4)$, $\alpha_2^* = (0.8, 0)$, and $\alpha_3^* = (0, 0)$. $x_{i\cdot}^c \in \mathbb{R}^4$ is the same process as in DGP 2 with ν_{i1} and ν_{i2} are independent AR(1) processes with the AR(1) coefficient 0.4. The pure I(1) $x_{i\cdot}$ follows a random walk, and the pure I(0) 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.

As we develop our theory with regressors of fixed dimension, OLS is a natural benchmark. Another benchmark is the oracle OLS estimator under infeasible information. The sample sizes in our exercise range $n = 40$, 80, 120, 200, 400 and 800. We run 1000 replications for each sample size and each DGP.

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 (CV), 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 as specified 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 CV to obtain $c_\lambda^{(1)}, \ldots, c_\lambda^{(100)}$. Then we fix $c_\lambda = \text{median}(c_\lambda^{(1)}, \ldots, c_\lambda^{(100)})$ 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 calibrated 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 and TAlasso.

4.2 Performance Comparison

Table 2(a) reports the out-of-sample prediction accuracy in terms of the mean prediction squared error (MPSE), $E\left[(y_n - \widehat{y}_n)^2\right]$. By the simulation design, the unpredictable variation arises from the variance of the idiosyncratic error u_i is 1. Plasso and Slasso achieve variable screening and consistent estimation as the predictors are homogeneously unit root processes. They are slightly better in forecasting than Alasso/TAlasso when the sample size is small. As the sample size increases to $n = 800$, Alasso surpasses the conventional LASSO estimators, which suggests that variable selection is conducive for forecasting in a large sample. In DGP 2 and DGP 3 of mixed roots and cointegrated regressors, the settings are more complicated for the conventional LASSO to navigate.

TAlasso is the best performer, and it is followed by Alasso which also beats the conventional LASSO methods by a non-trivial edge.

Table 2(b) summarizes the variable screening performance. Recall that the set of relevant regressors is $M^* = \left\{j : \theta_j^* \neq 0\right\}$ and the estimated active set is $\widehat{M} = \left\{j : \widehat{\theta}_j \neq 0\right\}$. We define two *success rates* for variable screening:

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

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

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

These expectations SR, SR_1 and SR_2 are computed by the average in the 1000 simulation replications.

In terms of the overall selection measure SR, TAlasso is the most effective and Alasso takes the second place. As the sample size increases, TAlasso's success rates shoot to nearly 1 in all DGPs, which supports variable selection consistency. While the difference in SR_1 among these methods becomes negligible when the sample size is large, the gain of TAlasso and Alasso stems largely from SR_2. The asymptotic theory suggests $\lambda_n \asymp \sqrt{n}$ 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. In view of the results in Table 2, the advantage of TAlasso's variable screening capability helps with forecast accuracy in the context of predictive regressions where many included regressors actually exhibit no predictive power.[2]

Finally, we check in Table 3 variable screening of the inactive cointegration group (x_{i3}, x_{i4}) in both DGPs 2 and 3, where the true coefficients $\phi_3^{0*} = \phi_4^{0*} = 0$. Alasso is effective in preventing both redundant variables from remaining in the regression according to the third column, while in the second column when $n = 800$ it omits one variable about 1/3 of chance in DGP 2 and 1/8 of chance in DGP 3. In contrast, TAlasso successfully identifies both redundant variables wpa1, as shown in the first column. Plasso and Slasso break down in variable selection consistency as theory predicts.

5 Empirical Application

We apply the LASSO methods to Welch and Goyal (2008)'s dataset to predict stock returns. We focus on the improvement in terms of prediction error and variable screening.

[2] In Table 2(b) Slasso has the lowest SR_2. However, due to the presence of $\widehat{\tau}_j = \widehat{\sigma}_j = O_p(\sqrt{n})$ in the penalty term, in asymptotics it imposes heavier penalty on coefficients of I(1) regressors than Plasso does. The reason is that in our simulations we fix c_λ^{Plasso} and c_λ^{Slasso} by CV separately. CV selects the tuning parameter c_λ favoring lower MPSE and adjusts c_λ in finite sample. For example, in DGP 1, $c_\lambda^{Plasso} = 0.00563$ whereas $c_\lambda^{Slasso} = 0.00119$ which is much smaller than c_λ^{Plasso}. If we fix c_λ^{Plasso} by CV and let $c_\lambda^{Slasso} = c_\lambda^{Plasso}$, Slasso would have a much higher SR_2.

Table 2: MPSE and variable screening in simulations

(a) MPSE

	n	Oracle	OLS	Alas.	TAlas.	Plas.	Slas.
DGP 1	40	1.2812	1.5413	1.5015	1.4967	**1.4077**	1.4710
	80	1.0621	1.1849	1.1554	1.1555	**1.1203**	1.1332
	120	1.0535	1.1317	1.0989	1.1005	**1.0886**	1.0943
	200	0.9598	1.0222	0.9981	0.9998	0.9964	**0.9924**
	400	1.0685	1.0947	1.0934	1.0889	**1.0888**	1.0950
	800	0.9815	1.0071	1.0062	**1.0011**	1.0087	1.0326
DGP 2	40	1.1482	1.3804	1.3418	**1.3342**	1.3483	1.3658
	80	0.9910	1.0739	1.0555	**1.0514**	1.0621	1.0661
	120	1.0927	1.1592	1.1516	**1.1492**	1.1524	1.1526
	200	1.0454	1.0822	1.0631	**1.0599**	1.0718	1.0752
	400	1.0989	1.1260	1.1098	**1.1053**	1.1190	1.1241
	800	0.9930	1.0046	1.0043	**0.9999**	1.0152	1.0493
DGP 3	40	1.3975	1.7930	1.7152	**1.7053**	1.7340	1.7710
	80	1.1668	1.2667	1.2255	**1.2195**	1.2445	1.2527
	120	1.1025	1.1834	1.1311	**1.1297**	1.1590	1.1743
	200	1.0594	1.1009	1.0784	**1.0765**	1.0897	1.0938
	400	1.0760	1.0973	1.0951	**1.0924**	1.1058	1.1129
	800	0.9980	1.0048	1.0032	**1.0023**	1.0050	1.0344

Note: The bold number is for the best performance among all the feasible estimators.

(b) Variable screening

		SR				SR_1				SR_2			
	n	Alas.	TAlas.	Plas.	Slas.	Alas.	TAlas.	Plas.	Slas.	Alas.	TAlas.	Plas.	Slas.
DGP 1	40	0.557	**0.561**	0.542	0.520	0.878	0.860	0.906	**0.955**	0.235	**0.262**	0.178	0.085
	80	0.611	**0.623**	0.577	0.556	0.876	0.866	0.927	**0.955**	0.346	**0.381**	0.227	0.156
	120	0.679	**0.698**	0.618	0.602	0.908	0.901	0.949	**0.963**	0.450	**0.496**	0.286	0.242
	200	0.760	**0.773**	0.652	0.656	0.926	0.918	0.966	**0.972**	0.594	**0.629**	0.339	0.340
	400	0.889	**0.902**	0.712	0.756	0.973	0.969	**0.990**	0.985	0.806	**0.835**	0.435	0.527
	800	0.968	**0.973**	0.773	0.827	0.983	0.981	**0.998**	0.975	0.953	**0.965**	0.548	0.679
DGP 2	40	0.586	**0.603**	0.514	0.502	0.927	0.912	0.983	**0.989**	0.246	**0.294**	0.046	0.016
	80	0.667	**0.699**	0.538	0.520	0.965	0.958	0.995	**0.997**	0.369	**0.440**	0.082	0.044
	120	0.725	**0.764**	0.556	0.539	0.983	0.979	**0.997**	**0.997**	0.467	**0.549**	0.116	0.081
	200	0.800	**0.850**	0.585	0.570	0.991	0.990	0.999	**0.999**	0.609	**0.710**	0.171	0.141
	400	0.895	**0.950**	0.650	0.640	0.996	0.996	**1.000**	0.999	0.793	**0.904**	0.300	0.282
	800	0.954	**0.995**	0.725	0.692	0.998	0.998	**1.000**	0.986	0.910	**0.992**	0.450	0.397
DGP 3	40	0.680	**0.701**	0.565	0.546	0.963	0.957	0.993	**0.997**	0.350	**0.402**	0.066	0.019
	80	0.757	**0.785**	0.589	0.558	0.971	0.969	0.994	**0.997**	0.508	**0.571**	0.117	0.046
	120	0.816	**0.844**	0.615	0.573	0.972	0.969	0.993	**0.995**	0.635	**0.698**	0.175	0.080
	200	0.874	**0.902**	0.662	0.603	0.969	0.967	**0.992**	0.992	0.763	**0.827**	0.276	0.150
	400	0.939	**0.958**	0.747	0.661	0.972	0.971	0.992	**0.992**	0.901	**0.943**	0.462	0.274
	800	0.957	**0.966**	0.846	0.709	0.969	0.969	**0.992**	0.991	0.943	**0.963**	0.676	0.379

Note: The bold number is for the best performance in each category of measurement.

Table 3: Variable screening in DGP 2 and 3 for inactive cointegrated group $\phi_3^{0*} = \phi_4^{0*} = 0$

		Both $\widehat{\phi}_3, \widehat{\phi}_4 = 0$				One and only one of $\widehat{\phi}_3, \widehat{\phi}_4 = 0$				Neither $\widehat{\phi}_3, \widehat{\phi}_4 = 0$			
	n	Alas.	TAlas.	Plas.	Slas.	Alas.	TAlas.	Plas.	Slas.	Alas.	TAlas.	Plas.	Slas.
DGP 2	40	0.048	0.085	0.002	0.000	0.374	0.405	0.087	0.041	0.578	0.510	0.911	0.959
	80	0.118	0.225	0.005	0.001	0.475	0.431	0.181	0.133	0.407	0.344	0.814	0.866
	120	0.158	0.339	0.006	0.006	0.567	0.443	0.270	0.232	0.275	0.218	0.724	0.762
	200	0.265	0.520	0.012	0.017	0.573	0.372	0.394	0.413	0.162	0.108	0.594	0.570
	400	0.441	0.833	0.039	0.076	0.534	0.153	0.631	0.739	0.025	0.014	0.330	0.185
	800	0.658	0.986	0.085	0.216	0.342	0.014	0.836	0.741	0.000	0.000	0.079	0.043
DGP 3	40	0.104	0.175	0.002	0.000	0.433	0.426	0.115	0.048	0.463	0.399	0.883	0.952
	80	0.210	0.367	0.009	0.001	0.521	0.412	0.193	0.132	0.269	0.221	0.798	0.867
	120	0.328	0.551	0.012	0.007	0.534	0.343	0.289	0.226	0.138	0.106	0.699	0.767
	200	0.441	0.753	0.016	0.017	0.524	0.219	0.465	0.446	0.035	0.028	0.519	0.537
	400	0.714	0.955	0.053	0.086	0.284	0.043	0.687	0.713	0.002	0.002	0.260	0.201
	800	0.878	0.999	0.111	0.199	0.122	0.001	0.807	0.727	0.000	0.000	0.082	0.074

Note: Each cell is the fraction of occurrence in the 1000 replications of eliminating both variables (the first column, desirable outcome), eliminating one and only one variable (the second column) and eliminating neither variable (the third column).

5.1 Data

Welch and Goyal (2008)'s dataset is one of the most widely used in predictive regressions. Koo, Anderson, Seo, and Yao (2019) update this monthly data from January 1945 to December 2012, and we use the same time span. The dependent variable is *excess return* (ExReturn), defined as the difference between the continuously compounded return on the S&P 500 index and the three-month Treasury bill rate. The estimated AR(1) coefficient of the excess return is 0.149, indicating weak persistence. The 12 financial and macroeconomic predictors are introduced and depicted in Figure 1. Three variables, namely `ltr`, `infl` and `svar`, oscillate around the mean, whereas nine variables are highly persistent with AR(1) coefficients greater than 0.95. The two pairs (`tms`, `dfr`) and (`dp`, `dy`) are visibly moving at a synchronized pattern that suggests potential cointegration, while it is much more difficult to judge whether cointegration holds among (`dp`, `dy`, `ep`). `ep` fluctuates with `dy` before 2000 but the link dissolves afterward and the two series even diverge toward opposite directions during the Great Recession. The presence of stationary predictors and persistent ones fits the mixed roots environment studied in this paper, and our agnostic approach avoids decision errors from statistical testing.

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 forecast, we construct the long-horizon excess return as the sum of continuous compounded monthly excess return on the S&P 500 index

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

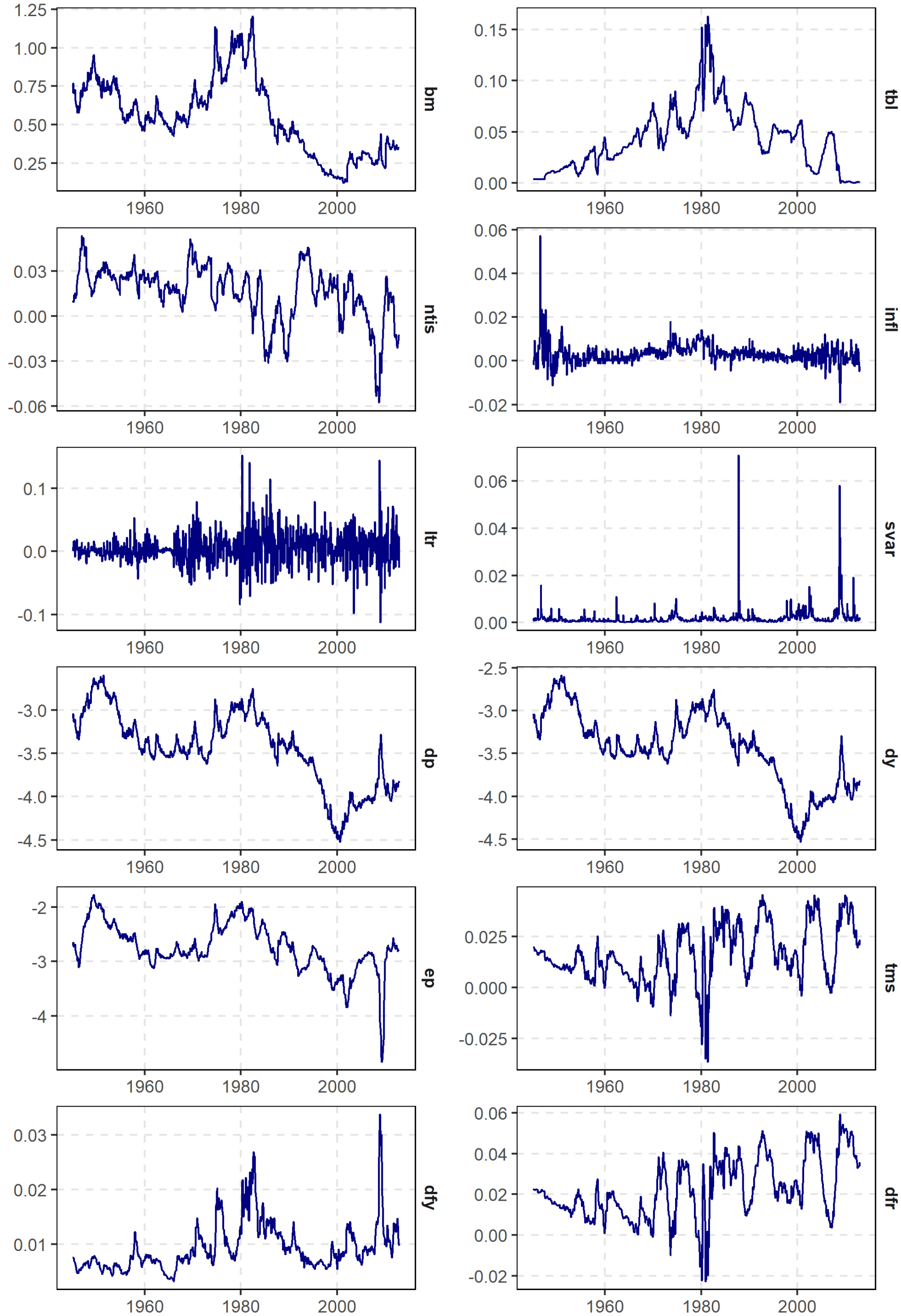


Figure 1: Time plot of 12 predictors of Welch and Goyal (2008)'s dataset. Except for the long-term return of government bonds (`ltr`), stock variance (`svar`) and inflation (`infl`), all the other nine variables' AR(1) coefficients are greater than 0.95. These persistent regressors include the dividend price ratio (`dp`), dividend yield (`dy`), earning price ratio (`ep`), term spread (`tms`), default yield spread (`dfy`), default return spread (`dfr`), book-to-market ratio (`bm`), and treasury bill rates (`tbl`).

where h is the length of the forecasting horizon. $h = 1$ stands for one year. We choose $h = 1/12$ (one month), $1/4$ (three months), $1/2$ (half a year), 1, 2, and 3 in our empirical exercises.

5.2 Performance

We apply the set of feasible forecasting methods as in Section 4 to forecast short-horizon and long-horizon stock returns recursively with either a 10-year or 15-year rolling window. All 12 variables are made available in the predictive regression, to which Welch and Goyal (2008) refer as the *kitchen sink model*. The tuning parameters for the shrinkage estimators are determined by 10-fold CV on MPSE with consecutive partitions in each estimation window. The CV method is a favorable choice for prediction purposes. For a robustness check, we also use the Bayesian information criterion (BIC) to decide the tuning parameters of Alasso and TAlasso. BIC is a popular choice geared to variable selection, but it is incompatible in our context with the conventional LASSO that cannot cope with variable screening and consistent estimation simultaneously.

The forecast returns of the LASSO methods are shown in Figure 2 along with the true realized return in gray color for $h = 1/12 = 0.083$ (short horizon), 1 (median horizon) and 3 (long horizon). When $h = 0.083$, the realized excess return resembles a white noise that is extremely difficult to forecast. When the horizon is extended to $h = 3$, the dynamics of the long-run aggregated return is evident. In most of the time Alasso and TAlasso track the realized return more closely.

Table 4 quantifies the forecast error in terms of the out-of-sample RMPSE (root MPSE) and mean predicted absolute error (MPAE) $E\left[|y_n - \widehat{y}_n|\right]$. In addition to OLS which involves all variables without any screening, we include random walk with drift (RWwD), i.e. the historical average of the excess returns, $\widehat{y}_{n+1} = \frac{1}{n}\sum_i^n y_i$, as another benchmark that utilizes no information from regressors at all. The results show that OLS loses in the short horizon whereas RWwD suffers in the long horizon, indicating ineffectiveness of either the all-in or all-out approach. Variable screening is essential for a balanced performance in this empirical example. Among the LASSO methods, in general Alasso and TAlasso forecast more precisely than the conventional Plass/Slasso. In particular, when the horizon is $h = 2$ or $h = 3$, TAlasso can achieve the smallest RMPSE and Alasso is also stronger than the Plasso and Slasso by a substantial margin. The results are robust when the tuning parameters are chosen by either CV or BIC.

There is an exceptional case of $h = 1$ with 15-year rolling window. Alasso fails to foresee the recovery trend after the financial crisis in 2008, and another round of Alasso further worsens TAlasso. As shown in Figure 2, when $h = 1$ Plasso's forecast happens to coincide with the movement of the realization during the recovery period after 2008 whereas Alasso/TAlasso swings to the opposite direction. While large deviation exacerbates RMPSE, under MPAE the gap in this case between Plasso and Alasso/TAlasso is narrowed or even reversed. Under the 10-year rolling window, all methods encounter difficulty around the financial crisis, and the difference between TAlasso and Slasso is negligible. Thus we view the unsatisfactory RMPSE of Alasso/TAlasso here as an adverse case under the specific rolling window.

In terms of prediction performance, it is known that eliminating irrelevant predictors is more im-

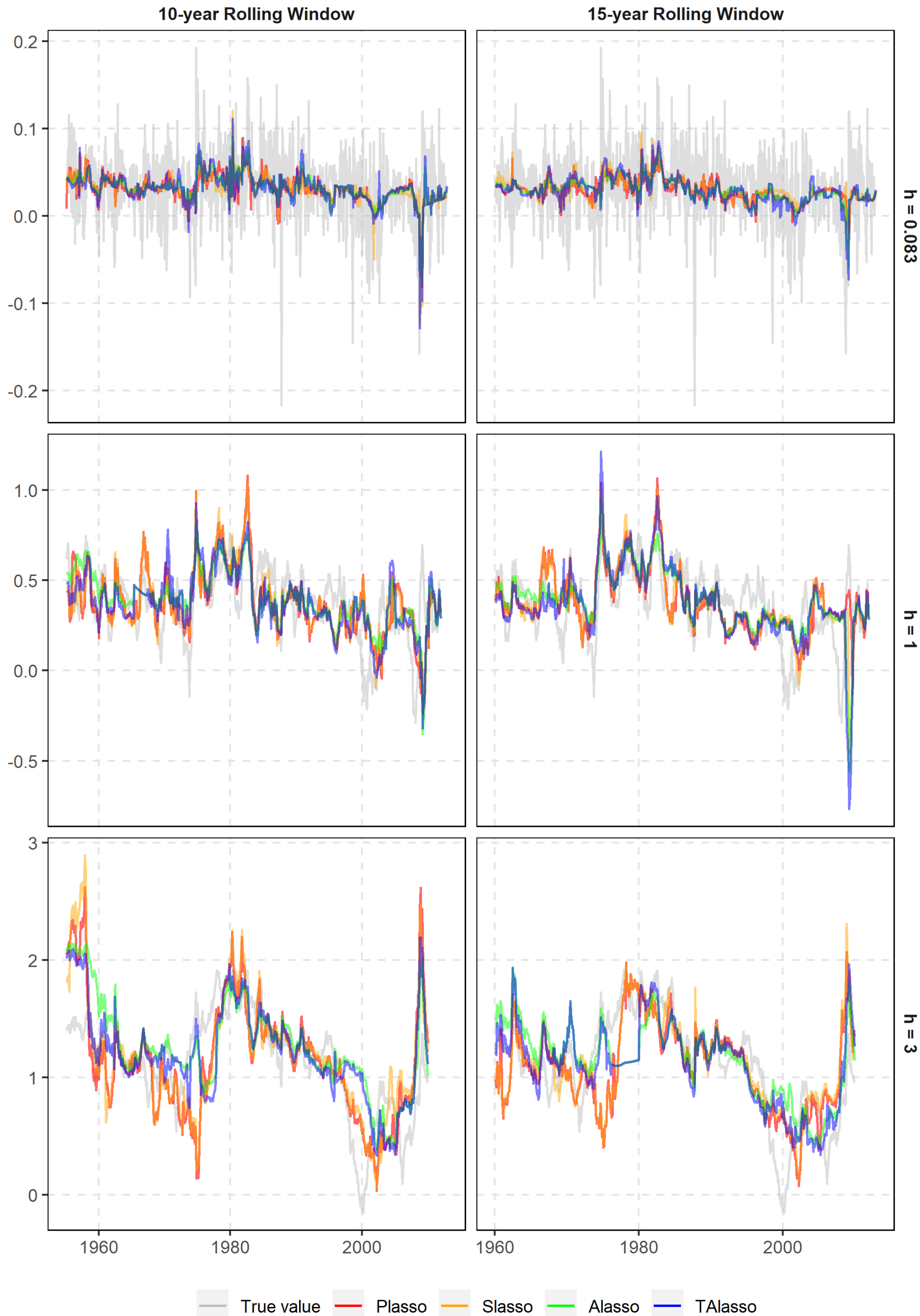


Figure 2: Realized return versus predicted returns. h is selected as 1/12, 1 and 3 to present the short horizon, medium horizon and long horizon, respectively.

Table 4: RMPSE and MPAE for predicting S&P 500 excess return

		OLS	RWwD	Plas.	Slas.	Alas.	TAlas.	Alas.	TAlas.
tuning para.		NA	NA	CV	CV	CV	CV	BIC	BIC
	h				RMPSE×100				
	1/12	*4.571*	4.344	4.314	4.328	**4.259**	4.350	4.339	4.343
10-year	1/4	*9.677*	8.144	9.149	8.707	**7.796**	7.947	8.081	8.116
rolling	1/2	*13.548*	12.821	12.688	12.388	**11.673**	11.958	12.444	12.407
window	1	18.449	*20.716*	17.579	17.219	17.671	17.474	17.404	**16.821**
	2	27.764	*36.011*	26.712	25.103	24.926	25.121	24.807	**24.007**
	3	44.795	*52.544*	39.970	42.135	38.234	36.095	38.282	**34.659**
	1/12	*4.499*	4.417	4.303	4.319	**4.284**	4.331	4.408	4.429
15-year	1/4	*9.091*	8.317	8.099	8.086	**7.919**	7.949	8.374	8.400
rolling	1/2	*14.175*	13.092	13.579	13.026	**12.505**	13.005	12.794	13.009
window	1	19.989	*21.092*	**17.165**	19.213	18.334	20.566	18.664	19.221
	2	24.386	*37.006*	22.932	23.219	20.970	**19.951**	21.097	20.006
	3	33.415	*54.035*	32.533	33.471	31.829	29.756	28.179	**27.518**
	h				MPAE×100				
	1/12	*3.422*	3.230	3.244	3.209	**3.174**	3.254	3.224	3.240
10-year	1/4	*6.538*	6.046	6.098	5.930	**5.637**	5.727	5.903	6.015
rolling	1/2	*10.038*	9.591	9.258	9.125	**8.538**	8.664	9.180	9.202
window	1	14.162	*16.149*	13.558	13.376	13.155	13.292	13.096	**12.682**
	2	21.669	*29.389*	19.560	19.190	17.856	18.069	18.014	**17.494**
	3	31.807	*44.231*	29.885	30.537	29.017	26.582	28.501	**25.901**
	1/12	*3.349*	3.277	3.251	**3.245**	3.248	3.295	3.277	3.313
15-year	1/4	*6.691*	6.238	6.055	6.062	**5.856**	5.917	6.193	6.192
rolling	1/2	*10.110*	9.940	9.727	9.522	**9.032**	9.439	9.130	9.378
window	1	15.003	*16.608*	13.812	14.553	**12.801**	13.991	13.186	13.166
	2	18.106	*30.876*	16.589	16.177	14.696	14.567	14.560	**14.077**
	3	24.126	*46.018*	24.362	24.491	25.081	23.297	21.644	**21.283**

Note: The bold number highlights the best performance in each row, while the italic number is for the worst.

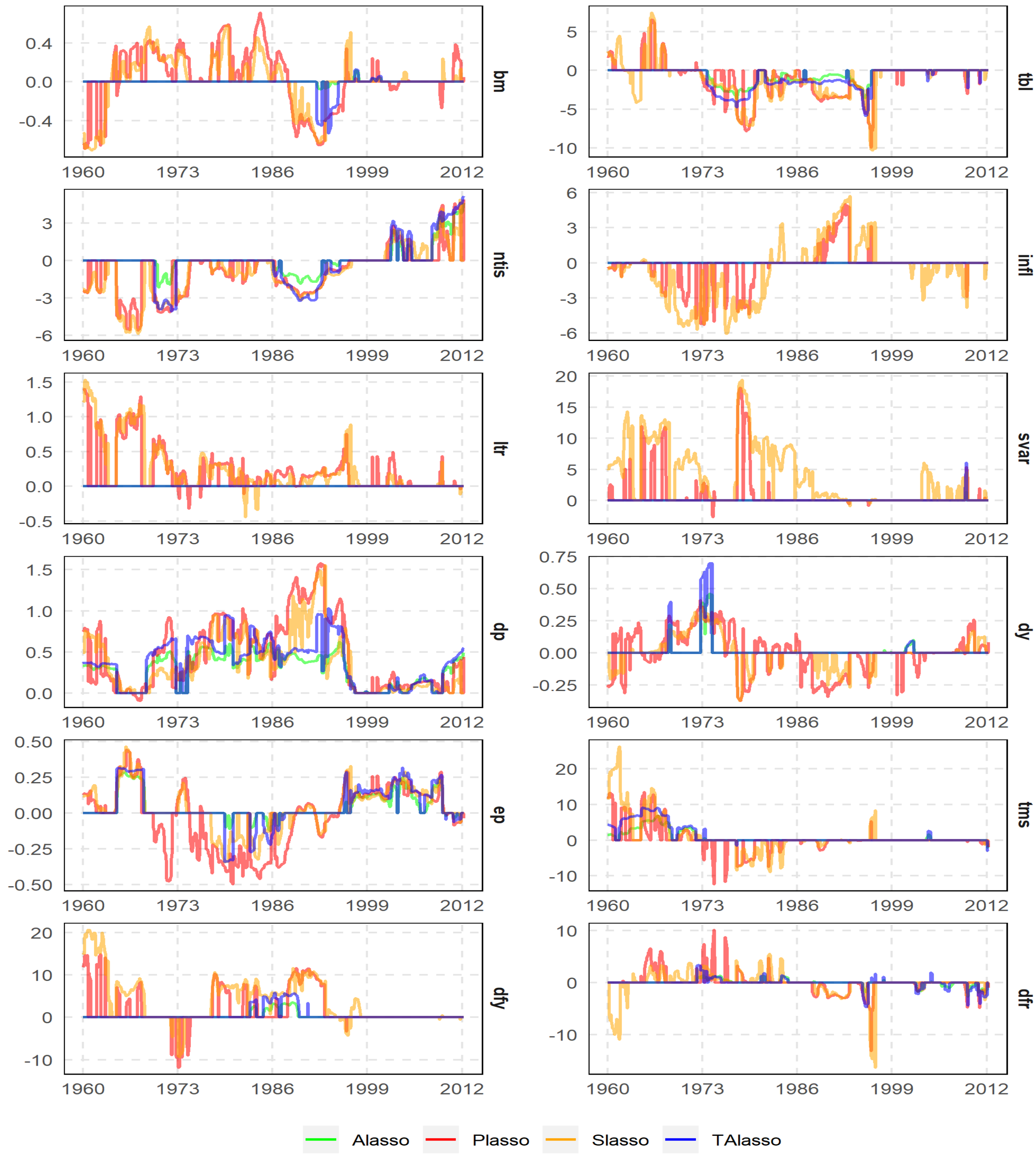
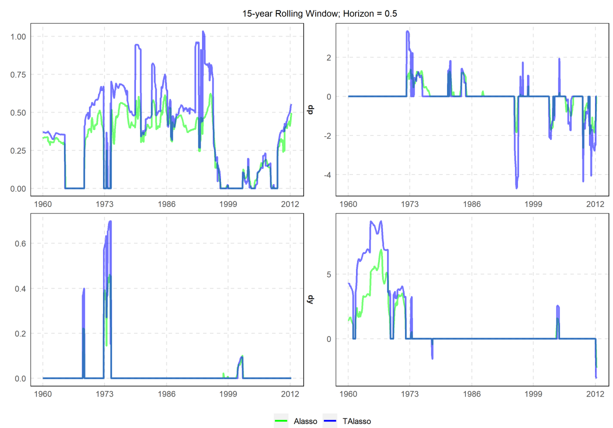


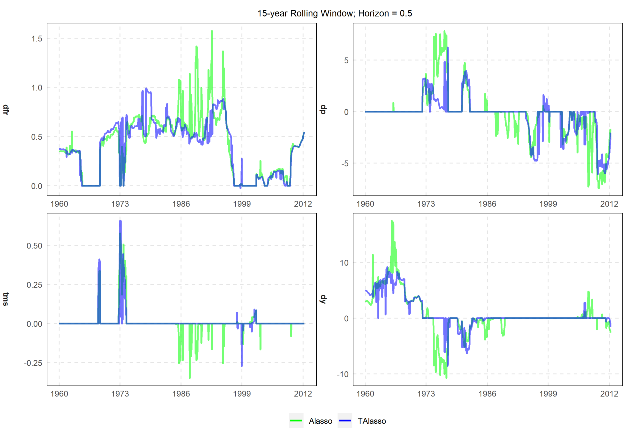
Figure 3: Estimated coefficients for all 12 variables (15-year rolling window, $h = 1/2$)

Fraction of active Alasso estimates eliminated by TAlasso

	`dp`	`dy`	`dfr`	`tms`
CV	0.018	0.250	0.263	0.014
BIC	0.016	0.460	0.236	0.308



(4.a) CV



(4.b) BIC

Figure 4: Estimated coefficient of potentially cointegrated variables

portant than including relevant ones, see Ploberger and Phillips (2003), for example. The inclusion of the irrelevant predictors could be detrimental in forecasting contexts, and including irrelevant I(1) predictors in predictive regression can be extremely harmful since stock returns are supposed to be stationary. In this sense, Alasso and TAlasso provide more conservative variable selection in predictive regressions. In Figure 3 the instance with $h = 1/2$ and 15-year rolling window is used to illustrate the estimated coefficients under CV. The shrinkage methods select different variables over the estimation windows, indicating the evolution of the predictive models across time. Alasso and TAlasso throw out more variables than Plasso or Slasso and hence deliver more parsimonious models. For example, they eliminate the variables `ltr` and `infl` completely. To highlight the potentially cointegrated variables, we zoom in Figure 4(a) the Alasso and TAlasso estimates of (`dp`, `dy`) and (`dfr`, `tms`), along with Figure 4(b) where BIC decides the tuning parameters and then produces the estimates. The table ahead of the subfigures lists the fraction, over the rolling windows, of the active Alasso estimates annihilated by TAlasso. BIC in general tends to freeze out more variables than CV, especially for `dy` and `tms`. Although the numbers vary, similar patterns are found in other combinations of the forecast horizon and the rolling window length.

6 Conclusion

We explore LASSO procedures in the presence of stationary, nonstationary and cointegrated predictors. While it no longer enjoys the well-known oracle property, Alasso breaks the link within inactive cointegration groups, and then its repeated implementation TAlasso recovers the oracle property thanks to the differentiated penalty on the zero and nonzero coefficients. TAlasso is adaptive to a system of multiple predictors with various degrees of persistence, unlike Plasso's uniform penalty or Slasso's penalty based only on the marginal variation of each predictor. Moreover, TAlasso 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 the predictors. The automatic penalty adjustment of TAlasso guarantees consistent model selection and the optimal rate of convergence. Such desirable properties may in practice improve the out-of-sample prediction under complex predictive environments with a mixture of regressors.

To focus on the mixed root setting, we adopt the simplest asymptotic framework with a fixed p and $n \to \infty$ to demonstrate the clear contrast between OLS, Alasso, TAlasso, Plasso and Slasso. This asymptotic framework is in line with the-state-of-art of the predictive regression studies in financial econometrics (Kostakis, Magdalinos, and Stamatogiannis, 2014; Phillips and Lee, 2016; Xu, 2018). On the other hand, a large number of potential regressors available in the era of big data are calling for theoretical extension to allow for an infinite number of regressors in the limit. As the restricted eigenvalue condition (Bickel, Ritov, and Tsybakov, 2009) is unsuitable in our context where the nonstationary part of the Gram matrix does not degenerate, in future research we are looking forward to 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 low-dimensional asymptotics to high-dimensional.

Another line of related literature concerns uniformly valid inference and forecasting after the LASSO model selection, see Belloni, Chernozhukov, and Kato (2015, 2018) or Hirano and Wright (2017), for example. These papers allow for model selection error by LASSO, and provide 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 make for interesting future research.

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.

Chinco, A. M., A. D. Clark-Joseph, and M. Ye (2019). Sparse signals in the cross-section of returns. *Journal of Finance forthcoming*.

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.

Giannone, D., M. Lenza, and G. E. Primiceri (2018). Economic predictions with big data: The illusion of sparsity. Technical report, Federal Reserve Bank of New York Staff Reports.

Gu, S., B. Kelly, and D. Xiu (2019). Empirical asset pricing via machine learning. *Review of Financial Studies forthcoming*.

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.

Kock, A. B. and H. Tang (2019). Inference in high-dimensional dynamic panel data models. *Econometric Theory 35*(2), 295–359.

Koo, B., H. M. Anderson, M. H. Seo, and W. Yao (2019). High-dimensional predictive regression in the presence of cointegration. *Journal of Econometrics forthcoming*.

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.

Ng, S. (2013). Variable selection in predictive regressions. In *Handbook of Economic Forecasting*, Volume 2, pp. 752–789. Elsevier.

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. (2015). Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics 13*(3), 521–555.

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. and Z. Shi (2019). Boosting: Why you can use the hp filter. *arXiv:1905.00175*.

Phillips, P. C. 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.

Shi, Z. (2016). Estimation of sparse structural parameters with many endogenous variables. *Econometric Reviews 35*(8-10), 1582–1608.

Smeekes, S. and E. Wijler (2018). Macroeconomic forecasting using penalized regression methods. *International Journal of Forecasting 34*(3), 408–430.

Smeekes, S. and E. Wijler (2020). Unit roots and cointegration. In *Macroeconomic Forecasting in the Era of Big Data*, pp. 541–584. Springer.

Su, L., Z. Shi, and P. C. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*(6), 2215–2264.

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

$$\begin{aligned} V_n(v) &= \Psi_n(v) - \Psi_n(0) \\ &= \|u - \frac{X'v}{n}\|^2 - \|u\|^2 + \lambda_n \left(\sum_{j=1}^p \widehat{\tau}_j |\beta_{jn}^* + \frac{v_j}{n}| - \sum_{j=1}^p \widehat{\tau}_j |\beta_{jn}^*| \right) \\ &= v'(\frac{X'X}{n^2})v - 2\frac{u'X}{n}v + \lambda_n \sum_{j=1}^p \widehat{\tau}_j (|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|). \end{aligned} \tag{30}$$

The first and the second terms in the right-hand side of (30) converge in distribution, as $\frac{X'X}{n^2} \Longrightarrow \Omega$ and $\frac{X'u}{n} = \frac{1}{n}\sum_{i=1}^n x_{i\cdot}' u_i \Longrightarrow \zeta$, 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

$$\widehat{\tau}_j = \left|\beta_{jn}^* + O_p\left(n^{-1}\right)\right|^{-\gamma} = |\beta_j^{0*}/\sqrt{n} + O_p\left(n^{-1}\right)|^{-\gamma}. \tag{31}$$

If $\beta_j^{0*} \neq 0$, then the β_{jn}^* dominates $n^{-1}v_j$ for a large n and

$$|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*| = n^{-1}v_j \mathrm{sgn}(\beta_{jn}^*) = n^{-1}v_j \mathrm{sgn}(\beta_j^{0*}). \tag{32}$$

Now (31) and (32) imply

$$\begin{aligned} \lambda_n \widehat{\tau}_j \cdot (|\beta_{jn}^* + \frac{v_j}{n}| - |\beta_{jn}^*|) &= \frac{\lambda_n}{n|\beta_j^{0*}/\sqrt{n} + O_p\left(n^{-1}\right)|^{\gamma}} v_j \mathrm{sgn}(\beta_j^{0*}) = \frac{\lambda_n n^{0.5\gamma-1}}{|\beta_j^{0*} + o_p\left(1\right)|^{\gamma}} v_j \mathrm{sgn}(\beta_j^{0*}) \\ &= O_p\left(\lambda_n n^{0.5\gamma-1}\right) = o_p\left(1\right) \end{aligned} \tag{33}$$

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 $v_j \neq 0$,

$$\lambda_n \widehat{\tau}_j \cdot (|\beta_j^* + \frac{v_j}{n}| - |\beta_j^*|) = \frac{\lambda_n}{n|\widehat{\beta}_j^{ols}|^{\gamma}} |v_j| = \frac{\lambda_n n^{\gamma-1}}{|n\widehat{\beta}_j^{ols}|^{\gamma}} |v_j| = \frac{\lambda_n n^{\gamma-1}}{O_p\left(1\right)} |v_j| \to \infty \tag{34}$$

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

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

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

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

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

$$\widehat{v}_{M^*}^{(n)} = n(\widehat{\beta}_{M^*}^{Alasso} - \beta_{M^*}^*) \Longrightarrow \Omega_{M^*}^{-1}\zeta_{M^*} \qquad \text{and} \qquad \widehat{v}_{M^{*c}}^{(n)} \Longrightarrow 0. \tag{35}$$

The first part of the above result establishes Theorem 2.1(b) about the asymptotic distribution for the coefficients in M^*.

Next we show variable selection consistency. As we only talk about Alasso in this proof, let $\widehat{M} = \widehat{M}^{Alasso}$. The result $P(M^* \subseteq \widehat{M}) \to 1$ immediately follows by the first part of (35), since $\widehat{v}_{M^*}^{(n)}$ converges in distribution to a non-degenerate continuous random variable. For those $j \in M^{*c}$, if the event $\{j \in \widehat{M}\}$ occurs, then the Karush-Kuhn-Tucker (KKT) condition entails

$$\frac{2}{n} x_j'(y - X\widehat{\beta}^{Alasso}) = \frac{\lambda_n \widehat{\tau}_j}{n}. \tag{36}$$

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

$$\frac{\lambda_n \widehat{\tau}_j}{n} = \frac{\lambda_n}{n|\widehat{\beta}_j^{ols}|^\gamma} = \frac{\lambda_n n^{\gamma-1}}{|n\widehat{\beta}_j^{ols}|^\gamma} = \frac{\lambda_n n^{\gamma-1}}{O_p(1)} \to \infty, \tag{37}$$

from the given rate of λ_n. However, using $y = X\beta_n^* + u$ and (35), the left-hand side of (36) is

$$\begin{aligned} \frac{2}{n} x_j'(y - X\widehat{\beta}^{Alasso}) &= \frac{2}{n} x_j'(X\beta_n^* - X\widehat{\beta}^{Alasso} + u) \\ &= 2\left(\frac{x_j'X}{n^2}\right) n(\beta_n^* - \widehat{\beta}^{Alasso}) + 2\frac{x_j'u}{n} \\ &= 2\left(\frac{x_j'X}{n^2}\right)\left(\widehat{v}_{M^*}^{(n)} + \widehat{v}_{M^{*c}}^{(n)}\right) + 2\frac{x_j'u}{n} \\ &\Longrightarrow 2\Omega_{j\cdot}(\Omega_{M^*}^{-1}\zeta_{M^*} + o_p(1)) + 2\zeta_j = O_p(1), \end{aligned} \tag{38}$$

where $\Omega_{j\cdot}$ is the j-th row of Ω. In other words, the left-hand side of (36) 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, $P(M^{*c} \subseteq \widehat{M}) \to 0$ or equivalently $P(\widehat{M} \subseteq M^*) \to 1$. We thus conclude the variable selection consistency in Theorem 2.1(a) $\blacksquare$

In the following proofs concerning Plasso and Slasso, we use a compact notation

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

for three generic vectors s, v, and β of the same dimension. It takes the the scalar-based symbol $D\left(\cdot, \cdot, \cdot\right)$ in the main text as a special case. Let the bold font $\mathbf{1}_p$ be a vector of p ones.

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 (30) 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 $V_n\left(v\right)$ and $V\left(v\right)$, the Convexity Lemma implies

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

In other words, the Plasso 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

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

$$\tilde{V}_n(v) = \tilde{\Psi}_n(v) - \tilde{\Psi}_n(0) = \frac{\lambda_n^2}{n^4} v'(X'X)v - \frac{\lambda_n}{n^2} 2u'Xv + \lambda_n \sum_{j=1}^{p} (|\beta_{jn}^* + \frac{\lambda_n}{n^2} v| - |\beta_{jn}^*|).$$

If $\beta_j^{0*} \neq 0$, in the limit $\frac{\lambda_n}{n^2} v$ is dominated by any $\beta_{jn}^* = \beta_j^{0*}/\sqrt{n}$ given the rate $\lambda_n / n^{3/2} \to 0$. For a sufficiently large n,

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

Notice that the scaled deviation $\widehat{v}^{(n)} = \lambda_n^{-1} n^2 (\widehat{\beta}^{Plasso} - \beta_n^*)$ can be expressed as $\widehat{v}^{(n)} = \arg\min_v \tilde{\Psi}_n(v)$. Since $\tilde{V}(v) = v'\Omega v + D\left(\mathbf{1}_p, v, \beta_0^*\right)$ is the limiting distribution of $\tilde{V}_n(v)$ and $\tilde{V}_n(v)$ is strictly convex, we invoke the Convexity Lemma to obtain $\frac{n^2}{\lambda_n}(\widehat{\beta}^{Plasso} - \beta_n^*) \Longrightarrow \arg\min_v \tilde{V}(v)$ as stated in the corollary. ■

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

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

where $\widehat{\sigma} = (\widehat{\sigma}_j)_{j=1}^p$ and $d = (d_j)_{j=1}^p$. Part (b) follows by the same argument as 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

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

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

Given the rate $\lambda_n/n^{3/2} \to 0$, 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

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

Since the limiting distribution of $\tilde{V}_n(v)$ is $\tilde{V}(v) = v'\Omega v + D\left(d, v, \beta^{0*}\right)$, the conclusion follows again by the Convexity Lemma. ■

A.2 Proofs in Section 3

Derivation of Eq.(16)

The columns in X^+ are 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 $u_i = \underset{1\times(p+1)}{\varepsilon_i'} \times \underset{(p+1)\times 1}{F_u(1)} - \triangle\tilde{\varepsilon}_{ui}$. Thus we have

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

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

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

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 (16).

Proof. [Proof of Theorem 3.2] We transform and scale-normalize the OLS estimator as

$$\begin{aligned} R_n Q\left(\widehat{\theta}^{ols}-\theta_n^*\right) &= R_n Q\left(W'W\right)^{-1} W'u \\ &= R_n Q\left(W'W\right)^{-1} Q'R_n\left(Q'R_n\right)^{-1} W'u \\ &= \left[R_n^{-1}Q'^{-1}W'WQ^{-1}R_n^{-1}\right]^{-1} R_n^{-1}Q'^{-1}W'u. \end{aligned} \tag{39}$$

The first factor

$$R_n^{-1}Q'^{-1}W'WQ^{-1}R_n^{-1} = \begin{pmatrix} \frac{Z^{+\prime}Z^+}{n} & \frac{Z^{+\prime}X^+}{n^{3/2}} \\ \frac{Z^{+\prime}X^+}{n^{3/2}} & \frac{X^{+\prime}X^+}{n^2} \end{pmatrix} \Longrightarrow \begin{pmatrix} \Omega_{zz}^+ & 0 \\ 0 & \Omega_{xx}^+ \end{pmatrix} = \Omega^+ \tag{40}$$

and the second factor

$$R_n^{-1}Q'^{-1}W'u = \begin{pmatrix} Z^{+\prime}u/\sqrt{n} \\ X^{+\prime}u/n \end{pmatrix} \Longrightarrow \zeta^+, \tag{41}$$

Thus the stated conclusion follows. ∎

To simplify notation, define R_{jn} as the j-th element of R_n, $\widehat{C} = \widehat{M}^{Alasso} \cap \mathcal{C}$, $C^* = M^* \cap \mathcal{C}$ and recall $C^{*c} = M^{*c} \cap \mathcal{C}$.

Proof. [Proof of Theorem 3.5] For a constant nonzero vector

$$\tilde{v} = \left(\tilde{v}_z', \tilde{v}_1', 0_{p_2}', \tilde{v}_x'\right)' \neq 0 \tag{42}$$

where the elements associated with $\mathcal{C}_2$ are suppressed as 0, we add a local perturbation

$$v_n = Q^{-1}R_n^{-1}\tilde{v} = \left(\frac{\tilde{v}_z'}{\sqrt{n}}, \frac{\tilde{v}_1'}{\sqrt{n}}, -\frac{\tilde{v}_1'A_1}{\sqrt{n}}, \frac{\tilde{v}_x'}{n}\right)$$

to θ_n^* so that the perturbed coefficient $\theta_n = \theta_n^* + v_n$. Let

$$\Psi_n(\tilde{v}) = \|Y - W\left(\theta_n^* + v_n\right)\|_2^2 + \lambda_n \sum_{j=1}^p \widehat{\tau}_j \left|\theta_{jn}^* + v_{jn}\right|,$$

and then define

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

We have shown in the proof of Theorem 3.2 that the first term

$$v_n' W' W v_n = \tilde{v}' R_n^{-1} Q'^{-1} W' W Q^{-1} R_n^{-1} \tilde{v} \Longrightarrow \tilde{v}' \Omega^{+} \tilde{v} \tag{43}$$

by (40) and the second term

$$2 v_n' W' u = 2 \tilde{v}' R_n^{-1} Q'^{-1} W' u \Longrightarrow 2 \tilde{v}' \zeta^{+} \tag{44}$$

by (41).

We focus on the third term. Theorem 3.2 has shown that the OLS estimator $\widehat{\theta}_j^{ols} - \theta_{jn}^* = O_p\left(R_{jn}^{-1}\right)$ for each $j \in \mathcal{M}_Q$. Given any fixed $\tilde{v}_j \neq 0$ and a sufficiently large n:

- For $j \in \mathcal{I}_0 \cup \mathcal{C}_1$, by the definition of v each element $v_{jn} = \tilde{v}_j / \sqrt{n}$. If $\theta_j^{0*} \neq 0$, we have $(|\theta_j^* + \tilde{v}_j/\sqrt{n}| - |\theta_j^*|) = n^{-1/2} \tilde{v}_j \text{sgn}(\theta_j^{0*})$, and thus $\lambda_n \widehat{\tau}_j \cdot (|\theta_j^{0*} + \frac{\tilde{v}_j}{\sqrt{n}}| - |\theta_j^{0*}|) = O_p\left(\lambda_n n^{-1/2}\right) = o_p\left(1\right)$. If $\theta_j^{0*} = 0$, we have $\lambda_n \widehat{\tau}_j \cdot (|\theta_j^{0*} + \frac{\tilde{v}_j}{\sqrt{n}}| - |\theta_j^{0*}|) = \frac{\lambda_n n^{\frac{\gamma-1}{2}}}{O_p(1)} \left|\tilde{v}_j\right| = O_p\left(\lambda_n n^{(\gamma-1)/2}\right) \to \infty$ when $\tilde{v}_j \neq 0$.
- For $j \in \mathcal{I}_1$, the true coefficient $\theta_{jn}^* = \theta_j^{0*}/\sqrt{n}$ depending on n while $v_{jn} = \tilde{v}_j/n$. If $\theta_j^{0*} \neq 0$, then θ_{jn}^* dominates $\tilde{v}_j/n$ in the limit and $(|\theta_{jn}^* + \tilde{v}_j/n| - |\theta_{jn}^*|) = n^{-1} \tilde{v}_j \text{sgn}(\theta_j^{0*})$. By the same derivation in (33), $\lambda_n \widehat{\tau}_j \cdot (|\theta_{jn}^* + \frac{\tilde{v}_j}{n}| - |\theta_{jn}^*|) = O_p\left(\lambda_n n^{(0.5\gamma - 1)}\right) = o_p\left(1\right)$ given the condition (21). If $\theta_j^{0*} = 0$, according to the derivation in (34) $\lambda_n \widehat{\tau}_j \cdot (|\theta_{jn}^* + n^{-1} \tilde{v}_j| - |\theta_{jn}^*|) = \frac{\lambda_n n^{\gamma-1}}{O_p(1)} \left|\tilde{v}_j\right| = O_p\left(\lambda_n n^{\gamma-1}\right) \to \infty$ when $\tilde{v}_j \neq 0$.

The above analysis indicates $V_n(\tilde{v}) \Longrightarrow V(\tilde{v})$ for every fixed $\tilde{v}$ in (42), where

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

Let $\widehat{v}^{(n)} = \widehat{\theta}^{Alasso} - \theta_n^*$. The same argument about the strict convexity of $V_n\left(\tilde{v}\right)$ and $V\left(\tilde{v}\right)$ implies

$$\left(R_n Q \widehat{v}^{(n)}\right)_{M_Q^*} \Longrightarrow \left(\Omega_{M_Q^*}^{+}\right)^{-1} \zeta_{M_Q^*}^{+} \tag{45}$$

$$\left(R_n Q \widehat{v}^{(n)}\right)_{M_Q^{*c}} \Longrightarrow 0 \tag{46}$$

We have established Theorem 3.5(a).

Next, we move on to discuss the effect of variable selection. For any $j \in \widehat{M}$ where $\widehat{M} = \widehat{M}^{Alasso}$. The KKT condition with respect to θ_j entails

$$2 W_j' (y - W \widehat{\theta}^{Alasso}) = \lambda_n \widehat{\tau}_j. \tag{47}$$

We will invoke similar argument as in (37) and (38) to show the disparity of the two sides of the KKT condition. The left-hand side of (47) is the j-th element of the $p \times 1$ vector $2W'(y - W\widehat{\theta}^{Alasso})$.

Pre-multiply the diagonal matrix $\frac{1}{2}R_n^{-1}Q'^{-1}$ to the vector:

$$\begin{aligned}
& R_n^{-1}Q'^{-1}W'(y - W\widehat{\theta}^{Alasso}) \\
= \; & R_n^{-1}Q'^{-1}W'\left(W\left(\theta_n^* - \widehat{\theta}^{Alasso}\right) + u\right) \\
= \; & \left(R_n^{-1}Q'^{-1}W'WQ^{-1}R_n^{-1}\right)R_nQ\left(\theta_n^* - \widehat{\theta}^{Alasso}\right) + R_n^{-1}Q'^{-1}W'u \\
= \; & R_n^{-1}Q'^{-1}W'WQ^{-1}R_n^{-1}O_p(1) + R_n^{-1}Q'^{-1}W'u \\
= \; & \left(\Omega^+ + o_p(1)\right)O_p(1) + \left(\zeta^+ + o_p(1)\right) = O_p(1)
\end{aligned} \tag{48}$$

where the third equality follows by (45) and (46), and the fourth equality by (40) and (41).

Suppose $j \in M^{*c}$. For $j \in \mathcal{I}$, the rotation Q does not change these variables so the order of the left-hand side of (47) is the same as (48). If $j \in \mathcal{I}_0$, multiply $0.5n^{-1/2}$ to the right-hand side of (47):

$$\frac{1}{\sqrt{n}}\lambda_n\widehat{\tau}_j = \frac{\lambda_n}{\sqrt{n}|\widehat{\theta}_j^{ols}|^\gamma} = \frac{\lambda_n n^{0.5(\gamma-1)}}{|\sqrt{n}\widehat{\theta}_j^{ols}|^\gamma} = O_p\left(\lambda_n n^{0.5(\gamma-1)}\right) \to \infty$$

as $\gamma \geq 1$ and $\lambda_n \to \infty$. Similarly, if $j \in \mathcal{I}_1$ multiply $0.5n^{-1}$ to the right-hand side of (47)

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

We have verified that for $j \in \mathcal{I}$ the right-hand side of (47) is of bigger order than its left-hand side. It immediately follows that given the specified rate for λ_n, for any $j \in M^{*c} \cap \mathcal{I}$ we have (24) since $P(j \in \widehat{M} \cap M^{*c}) \to P\left(O_p(1) = \infty\right) = 0$. We have established (24).

Estimation consistency (22) immediately implies (25). One the other hand, it is possible that variables in $\mathcal{C}$ are wrongly selected into $\widehat{M}$. If the event $\{j \in \widehat{M}\}$ occurs for some $j \in \mathcal{C}$, the KKT condition may still hold in the limit. To see this, pre-multiply $\frac{1}{2n}$ to the the left-hand side of (47) and it is of order $O_p(1)$ according to (48). However, the right-hand side becomes $\frac{1}{2n}\lambda_n\widehat{\tau}_j = \frac{\lambda_n n^{(0.5\gamma-1)}}{2|\sqrt{n}\widehat{\theta}_j^{ols}|^\gamma}$. Although the numerator shrinks to zero asymptotically according to (21), the denominator $\sqrt{n}\widehat{\theta}_j^{ols} = O_p(1)$ by Theorem 3.2 so that we cannot rule out the possibility that the two sides of (47) being equal when $\widehat{\theta}_j^{ols}$ is close to zero.

Finally, to show (26) we argue by contraposition. For those $j \in \widehat{C}$, the counterpart of (47) is the following KKT condition

$$2x_j^{c\prime}(y - W\widehat{\theta}^{Alasso}) = \lambda_n\widehat{\tau}_j, \quad \forall\, j \in \widehat{C}. \tag{49}$$

(25) already rules out $\text{CoRk}(M^*) > \text{CoRk}(\widehat{M})$ in the limit. Now suppose $\text{CoRk}(M^*) < \text{CoRk}(\widehat{M})$. If so, for any $\widehat{M}$ there exists a constant cointegrating vector $\psi \in \mathbb{R}^{p_c}$ such that $\psi'\psi = 1$ (scale normalization), $\psi_{C^{*c}\cap\widehat{C}} \neq 0$ (must involve wrongly selected inactive variables) and the linear combination $x_{i\cdot}^c\psi$ is an I(0). While ψ is not unique in general, we use ψ to represent any one of them.

Such a ψ can linearly combine the multiple equations in (49) to generate a single equation

$$\frac{2}{\sqrt{n}} \sum_{j \in \widehat{C}} \psi_j x_j^{c\prime} (y - W\widehat{\theta}^{Alasso}) = \frac{\lambda_n}{\sqrt{n}} \sum_{j \in \widehat{C}} \psi_j \widehat{\tau}_j. \tag{50}$$

The left-hand side of (50) is $O_p(1)$ since $\sum_{j \in \widehat{C}} \psi_j x_{ji}^{c\prime} = O_p(1)$ due to cointegration. The right hand side of (50) can be decomposed into two terms

$$\frac{\lambda_n}{\sqrt{n}} \sum_{j \in \widehat{C}} \psi_j \widehat{\tau}_j = \frac{\lambda_n}{\sqrt{n}} \sum_{j \in \widehat{C} \cap C^*} \psi_j \widehat{\tau}_j + \frac{\lambda_n}{\sqrt{n}} \sum_{j \in \widehat{C} \cap C^{*c}} \psi_j \widehat{\tau}_j =: I + II.$$

The first term

$$I \leq \frac{\lambda_n}{\sqrt{n}} \sum_{j \in \widehat{C} \cap C^*} |\psi_j| \, |\widehat{\tau}_j| \leq \frac{\lambda_n}{\sqrt{n}} \sum_{j \in \widehat{C} \cap C^*} |\widehat{\tau}_j| \leq \frac{\lambda_n}{\sqrt{n}} \sum_{j \in C^*} |\widehat{\tau}_j| = \frac{\lambda_n}{\sqrt{n}} O_p(1) = o_p(1)$$

as $\widehat{\tau}_j = O_p(1)$ for $j \in \mathcal{C}^*$, whereas the second term

$$II = \lambda_n n^{0.5(\gamma-1)} \sum_{j \in \widehat{C} \cap C^{*c}} \frac{\psi_j}{|\sqrt{n}\widehat{\theta}_j^{ols}|^\gamma} \to \infty \text{ or } -\infty$$

as $\sqrt{n}\widehat{\theta}_j^{ols} = O_p(1)$ for $j \in C^{*c}$, ψ_j is a constant with nonzero elements, and $\lambda_n n^{0.5(\gamma-1)} \to \infty$ given the specified rate for λ_n. Whether the right-hand side diverges to $+\infty$ or $-\infty$ depends on the configuration of ψ. The equation (50) holds with probability approaching 0 when n is sufficiently large. That is, no inactive cointegrating residuals can be formed within the selected variables wpa1, because a redundant cointegration group would induce such a cointegration vector ψ that sends the right-hand side of (50) to either positive infinity or negative infinite in the limit. ■

Proof. [Proof of Theorem 3.9] In Eq.(12) we have separated the cointegrated variables into the active ones and the inactive ones

$$\begin{aligned} y_i &= z_{i\cdot}\alpha + + x_{i\cdot}\beta + u_i \\ &= z_{i\cdot}\alpha + \sum_{l \in C^*} x_{il}^c \phi_l + \sum_{l \in C^{*c}} x_{il}^c \phi_l + x_{i\cdot}\beta + u_i. \end{aligned} \tag{51}$$

We proceed our argument conditioning on the three events

$$\begin{aligned} S_1 &= \left\{ M^* \cap \mathcal{I} = \widehat{M} \cap \mathcal{I} \right\} \\ S_2 &= \left\{ C^* \subseteq \widehat{C} \right\} \\ S_3 &= \left\{ \text{CoRk}(M^*) = \text{CoRk}(\widehat{M}) \right\}. \end{aligned}$$

According to Theorem 3.5, these events occur wpa1, given sufficiently large sample size.

Under these three events, the regression equation (51) is reduced to

$$y_i \;=\; z_{iM^*}\alpha_{M^*} + \sum_{l\in C^*} x^c_{il}\phi_l + \sum_{l\in C^{*c}\cap\widehat{C}} x^c_{il}\phi_l + x_{iM^*}\beta_{M^*} + u_i, \tag{52}$$

where on the right-hand side of the above equation the first and the fourth terms are present by the event S_1, the second and the third terms by S_2. Due to event S_3, there is no cointegration group in the third term, so we can re-write (52) as

$$y_i = z_{iM^*}\alpha_{M^*} + \sum_{l\in\mathcal{C}} x^c_{il}\phi_l + \tilde{x}^+_{i\cdot}\tilde{\beta}^+ + u_i, \tag{53}$$

where $\tilde{x}^+_{i\cdot} = \left((x^c_{il})_{l\in C^{*c}\cap\widehat{C}}, x_{M^*}\right)$ is the collection of *augmented* pure I(1) processes in the post-selection regression equation (52) and $\tilde{\beta}^+ = \left((\phi_j)_{l\in C^{*c}\cap\widehat{C}}, \beta_{M^*}\right)$ is the corresponding coefficient.

For $l \in C^{*c}\cap\widehat{C}$, the true coefficients are 0. Now these variables appear as pure I(1) in (53), for which Theorem 3.5 gives variable selection consistency. We implement the post-selection Alasso in (53), or equivalently TAlasso as in (27). Among the variables in $\tilde{x}_{i\cdot}$ TAlasso will eliminate $(x^c_{il})_{l\in C^{*c}\cap\widehat{C}}$ wpa1 while the true active variables x_{M^*} will be maintained. In the mean time, all variables in the first and second terms in (53) are active and Theorem 3.5's (25) guarantees their maintenance wpa1. The asymptotic distribution naturally follows by applying Theorem 3.5(a) to (52). ■

Proof. [Proof of Corollary 3.10] For Part (a) and (b), we start with the local perturbation $\check{v} = (\check{v}'_z, \check{v}'_1, \check{v}'_2, \check{v}'_x)'$, $v_n = Q^{-1}R_n^{-1}\check{v}$ and $\theta_n = \theta^*_n + v_n$. Notice that $\check{v}$ is different from $\tilde{v}$ in Theorem 3.5 as we do not impose $\check{v}_2 = 0$. Define

$$V_n(\check{v}) = v_n'W'Wv_n - 2v_n'W'u + \lambda_n\sum_{j=1}^{p}(|\theta^*_{jn} + v_n| - |\theta^*_{jn}|). \tag{54}$$

In view of (43) and (44),

$$V_n(\check{v}) \Longrightarrow V(\check{v}) = \check{v}'\Omega^+\check{v} - 2\check{v}'\zeta^+ + \lim_{n\to\infty}\lambda_n\sum_{j=1}^{p} R^{-1}_{jn}D(1,\tilde{v}_j,\theta^{0*}_j).$$

We invoke the Convexity Lemma. For Part (a), $\lambda_n/R_{jn}\to 0$ for all j so the penalty vanishes in the limit and the asymptotic distribution is equivalent to that of OLS. For Part (b), the tuning parameter's rate is $\lambda_n/\sqrt{n}\to c_\lambda\in(0,\infty)$ so that only the penalty term associated with $\mathcal{I}_1$ vanishes in the limit.

Part (c) needs more subtle elaboration. Let $v_{\lambda_n} = \frac{\lambda_n}{\sqrt{n}}Q^{-1}R_n^{-1}\check{v} = \frac{\lambda_n}{\sqrt{n}}v_n$ and $\theta_{\lambda_n} = \theta^*_n + v_{\lambda_n}$. Define

$$V_{\lambda_n}(\check{v}) = v'_{\lambda_n}W'Wv'_{\lambda_n} - 2v'_{\lambda_n}W'u + \lambda_n\sum_{j=1}^{p}(|\theta^*_{jn} + v_{\lambda_n,j}| - |\theta^*_{jn}|),$$

and multiply n/λ_n^2 on both sides,

$$\left(\frac{n}{\lambda_n^2}\right) V_{\lambda_n}(\check{v}) = v_n' W' W v_n - \frac{\sqrt{n}}{\lambda_n} 2 v_n' W' u + \frac{n}{\lambda_n} \sum_{j=1}^{p} (|\theta_{jn}^* + v_{\lambda_n,j}| - |\theta_{jn}^*|).$$

By the rate condition of λ_n, the second term $\frac{\sqrt{n}}{\lambda_n} 2 v_n' W' u = o_p(1)$. Given $\tilde{v}_j \neq 0$ and n large enough:

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

$$\frac{n}{\lambda_n}(|\theta_j^{0*} + \frac{\lambda_n}{\sqrt{n}} \frac{\check{v}_j}{\sqrt{n}}| - |\theta_j^{0*}|) = \frac{n}{\lambda_n} D\left(1, \frac{\lambda_n}{n} \check{v}_j, \theta_j^{0*}\right) = D\left(1, \check{v}_j, \theta_j^{0*}\right). \tag{55}$$

- If $j \in \mathcal{I}_1$, the coefficient $\theta_{jn}^* = \theta_j^{0*}/\sqrt{n}$ shrinks faster than $\frac{\lambda_n}{n^{3/2}}$. The reverse triangular inequality $||a+b| - |a|| \leq |b|$ for any $a, b \in \mathbb{R}$ guarantees

$$\left| \frac{n}{\lambda_n} \left(|\theta_{jn}^* + \frac{\lambda_n}{n^{3/2}} \check{v}_j| - |\theta_{jn}^*| \right) \right| \leq \frac{n}{\lambda_n} \left| \frac{\lambda_n}{n^{3/2}} \check{v}_j \right| = O\left(n^{-1/2}\right), \tag{56}$$

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

Thus we conclude

$$\left(\frac{n}{\lambda_n^2}\right) V_{\lambda_n}(\check{v}) \Longrightarrow \tilde{v}' \Omega^+ \tilde{v} + \sum_{j \in \mathcal{I}_0 \cup \mathcal{C}} D\left(1, \tilde{v}_j, \theta_j^{0*}\right)$$

as stated in the Corollary. ■

Proof. [Proof of Corollary 3.11] We start with the same local perturbation $\check{v} = (\check{v}_z', \check{v}_1', \check{v}_2', \check{v}_x')'$, $v_n = Q^{-1} R_n^{-1} \check{v}$ and $\theta_n = \theta_n^* + v_n$ as in the proof of Corollary 3.10. We focus on the counterpart of the terms in (54).

- if $j \in \mathcal{I}_0$, we have $\widehat{\sigma}_j = O_p(1)$ and the coefficient θ_j^{0*} is independent of n so that

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

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

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

  since these indices are associated with a unit root process x_j^c and therefore $\frac{\widehat{\sigma}_j}{\sqrt{n}} \Longrightarrow d_j$ is degenerate;

- if $j \in \mathcal{I}_1$, for these unit root processes in x_j similarly we have

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

The above analysis implies

$$V_n(\check{v}) \quad\Longrightarrow\quad V\left(\check{v}\right)=\check{v}'\Omega^{+}\check{v}-2\check{v}'\zeta^{+}+c_\lambda\sum_{j\in\mathcal{C}}D\left(d_j,v_j,\theta^{0*}_j\right), \tag{57}$$

and Part (a) and (b) follow.

For Part (c), let $\tilde{R}_n=R_n/\lambda_n$ and $\theta_n=\theta^*_n+\tilde{R}_n^{-1}\tilde{v}$. Define

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

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

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

by the rate condition of λ_n. Again we study the last term. For $\check{v}_j\neq 0$ and a sufficiently large n:

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

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

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

$$\frac{1}{\lambda_n}\widehat{\sigma}_j\left(|\theta^{0*}_j+\frac{\lambda_n}{\sqrt{n}}\check{v}_j|-|\theta^{0*}_j|\right)=\frac{1}{\lambda_n}D\left(\widehat{\sigma}_j,\frac{\lambda_n}{\sqrt{n}}\check{v}_j,\theta^{0*}_j\right)=D\left(\frac{\widehat{\sigma}_j}{\sqrt{n}},\check{v}_j,\theta^{0*}_j\right)=D\left(d_j,v_j,\theta^{0*}_j\right)=O_p\left(1\right);$$

- for $j\in\mathcal{I}_1$, the rate condition $\lambda_n/n^{0.5}\to 0$ makes sure that $\theta^*_{jn}=\theta^{0*}_j/\sqrt{n}$ dominates $\frac{\lambda_n}{n}$ in the limit, so that

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

We obtain $\frac{\tilde{V}_n(\check{v})}{\lambda_n^2}\Longrightarrow\check{v}'\Omega^{+}\check{v}+\sum_{j\in\mathcal{C}}D\left(d_j,v_j,\theta^{0*}_j\right)$ and the conclusion follows. ■