

LASSO-based predictive regression for stock returns

Master's thesis
for the Master's degree programme Economics in the Faculty of
Business, Economics and Social Sciences at the
Christian-Albrechts-Universität zu Kiel

Submitted by

Hieu, Hoang

First assessor: Prof. Dr. Matei Demetrescu
Second assessor: Prof. Dr. Kai Carstensen
Kiel, January, 2020

Table of Contents

Contents

1	Introduction	2
2	Literature Review	3
3	The Adaptive Lasso and Its Advantages	5
4	Simulation study	11
5	Application: Goyal's data set	15
6	Conclusion	20

1 Introduction

Prediction of stock returns has always been an important subject in finance since an accurate prediction can help investors decide the portions of safe and risky assets in their portfolios, generating optimal wealth for their clients and themselves. As a result, a large volume of literature has been dedicated to developing asset pricing theory and predictive statistical models. One of the most popular and long-standing bases in the realm of asset pricing is the efficient market hypothesis (EMH), summarized and popularized in Fama and Malkiel (1970). According to EMH, equity premium is constant and reflects all available information. Hence, market anomalies and historical average of premia are analyzed to forecast asset returns.

On the other hand, financial econometricians focus on including relevant lagged financial and macroeconomic variables as predictors for equity premium as a way to exploit market inefficiency. Fama and French (1988), Schiller and Campbell (1998), among others explored the power of using valuation ratios, e.g. dividend-price ratio, dividend yield, earnings-price, book-to-market ratio to forecast long-term returns of stock, while Fama (1990), Schwert (1990), and related papers showed correlation between bonds (treasury and corporate) and stock returns. As a result, a healthy number of literature demonstrated and cemented this idea, see, among Hodrick (1992), Kothari and Shanken (1997), Lamont (1998), Pontiff and Schall (1998).

As of the late 1990s and early 2000s, the common consensus within the field is that excess stock returns can be predicted (Welch and Goyal, 2008). However, along with new tools come new challenges. One of which is that the results of aforementioned findings may very well be spurious. (Stambaugh, 1999) showed that when the innovation of a predictor is correlated with excess return, which is the case for many valuation ratios, the resulting estimated coefficient is biased and exhibits sharply different finite-sample properties from the standard case. One apparent example is the dividend yield which has the same return component with equity premium. A second source for spurious result stems from the high persistence of predictors. It has been well-known that a regression model with integrated or near-integrated predictors may sometimes produce "non-sense correlation" where highly significant betas and high R^2 values are obtained while no "real" and meaningful correlation exists, except for the case of co-integrated series; see among Yule (1926), Granger and Newbold (1974), Phillips (1986), Granger and Newbold (2001), and Engle and Granger (1987). Last but not least, the problem of near-collinear predictors is also apparent for financial econometricians. As stated above, many highly persistent time series variables exhibit meaningless high cross-correlation scenario, but high correlation nonetheless. Furthermore, some ratios are by construction derived from other ratios or macroeconomic variables, hence the possible long-term co-movement between said ratios. This in turn causes the design matrix to become nearly singular (asymptotically singular), which pro-

duce estimation inconsistencies and failures in central limit theory in least square regression even in the case of stationary predictors and strong regression signal (Phillips, 2016), much less for our case of high-persistence.

For the reasons stated above, OLS may not be the best way to go. The late 1990s and early 2000s also witnessed a relatively new method for estimating linear models: the lasso (Tibshirani, 1996). Instead of just minimizing the residual sum of squares, the lasso further applies a constraint to the sum of absolute values of estimated coefficients. This penalty term helps shrink coefficient estimates and at the same time encourages some variables to take on zero as coefficient, effectively eliminate them from the model. Thus, the lasso has the advantages of both (continuous) model selection and variance reduction by trading it for Specifically, the stochastic some amount of bias, which may increase predictive performance. Belonging to the penalized least squared family of regression method, lasso also benefits from stability when there exists high collinearity between predictors by preventing coefficient inflation as in the case of its brother shrinkage variation, the ridge regression. Still, lasso suffers from a number of problems, mainly related to inference. Additionally, we still have the problem of near-singular design matrix and near-integrated and integrated series, potentially cointegrated in our hand.

This paper will be organized as followed. This section gave an introduction to the paper. Next, I will review some literature regarding the problem of asymptotically degenerate design matrix, and how it makes OLS estimations invalid. An overview of lasso (and its variants) and how lasso can be used to combat our prevailing problems will also be discussed in this section. The third section will be about the technical details of Adaptive Lasso (henceforth alasso), with focus on the inference of coefficient estimates. The fourth section compares predictive performance of alasso with regular lasso, autoregressive of order 1, and OLS in simulation settings. Lastly, I will apply alasso to Goyal’s data set used in Welch and Goyal (2008), with updates until 2018, to assess its real-world predictive performance. Conclusion and extension will be given as closing thoughts.

2 Literature Review

In this section, I will first review a number of the literature that discusses our three big problems in the context of OLS regression: correlation between innovation of lagged predictor and regression disturbance; mixed roots, high persistence predictors; and near-singular design matrix. Next, we will take a look at how lasso-type regression, specifically alasso, can help alleviate parts of our problems.

OLS, or ordinary least squares, is a long-standing powerhouse in the scene of linear regression. The objective of OLS is to find a set of coefficients that minimizes the squared differences between observed dependent variables and its predicted value. Thank to its readily available analytical solution, fast compu-

tation, and well-studied inference, it is widely used in both cross-sectional and time series data alike. However, there are a set of assumptions required to make the OLS estimates valid. Specifically, innovation of predictors are generally not allowed to be correlated with dependent variable. Such assumption does not hold very well for predictive regression due to overlapping in the construction of valuation ratios, macroeconomic variables, and equity premium. Next, the stochastic processes involved must be stationary and ergodic. However, evidences for stationarity of valuation ratios are mixed and shaky. Roll (2002) argues that under rational expectation, asset price is non-stationary due to its dependence on expectation of future quantities. Yet, metrics that are constructed as functions of price, e.g market-to-book, earning-price ratios, dividend yields,... may exhibit different root characteristics (Phillips, 2015). At the same time, most remaining series show high yet imprecisely determined degree of persistence, leading to the problem of mixed roots, possibly cointegrated, regression; and nonstationarity leads to the endogeneity effect in the limit, leading to non-standard limit theory (Phillips, 2015). The same paper also discusses "misbalancing" issue, where predicted variable and predictors have different memory types. The solution out is not straight-forward. Elliott and Stock (1994) discussed two common simple solutions: ignore the problem altogether, or determine the post-regression inference by pretesting predictors for unit roots. Both lead to the substantial over-rejection of the null of no significance. The solution proposed in the same paper involves Bayesian statistic, which may not be appealing to some. In another approach, a local-to-unity autoregressive specification in the form of $\rho = 1 + \frac{c}{n}$ is used to conduct asymptotic theory. However, the introduction of the unknown parameter brings more issues. Since localizing coefficient c is not consistently estimable, asymptotic bias cannot be corrected, leading to nonstandard limit theory. Phillips and Lee (2013) discussed c and suggested possible solutions.

Another violation of OLS assumptions comes in the form of high to perfect correlation between predictors. In the case of perfect correlation, removing inappropriate regressor(s) is a common remedy. When the correlation is not perfect, determinant(s) of design matrix gets into the vicinity of zero, causing computational difficulty in matrix inversion and inflated coefficient estimates. However, removal of regressors is not always an preferable option since each regressor may contain some additional information that can improve the model fit or prediction. In the case of predicting excess returns, some predictors may contain information about market inefficiency despite high correlation due to a common variable in their construction. This kind of construction also leads to possible co-movement of predictors, causing singularity in the limit (near-singular design matrix). As variable frequency increases such as in the case of financial data, singularity can come very quickly.

In this paper, I would like to introduce a relatively new method of estimation for linear model that has the ability to hopefully overcome some of the aforementioned issues. Proposed and discussed by Tibshirani (1996), the lasso (least absolute shrinkage and selection operator) exhibits some more preferable proper-

ties than the well-known OLS.

First, while shrinking coefficients introduces some amount of bias, it helps to prevent estimate inflation in presence of high-collinearity. In fact, for the case of near-singular design, the lasso estimates are consistent, and with an appropriate choice for shrinkage parameter λ , limiting distribution is normal (Knight et al., 2000; Knight, 2008). This result is especially handy for the case in this paper.

Second, lasso can set some coefficients to zero, effectively performs continuous model selection. Via this mechanism, variance is reduced and hence accuracy may increase in the case of predictive regression. Continuous model selection has some advantages over discrete model selection, for example subset selection. Small changes in data can lead to substantially different selection outcome for subset selection, or the selection event can be trapped in a local optimum (Breiman, 1995). Furthermore, as the number of predictors increases, discrete selection is computationally hard. On the other hand, continuous selection process is more stable, interpretable, and can scale easily with a large number of predictors (Tibshirani, 1996). However, plain-vanilla lasso is not always consistent in identifying the right subset of predictors, and does not always exhibit "oracle properties" (consistency in model selection and asymptotic normality) (Meinshausen and Bühlmann, 2004; Zou, 2006). Hence, the Adaptive Lasso (alasso) is proposed by Zou (2006) as an alternative. Alasso assigns different weights to each coefficients, and if the weights are cleverly chosen and data-driven, alasso can enjoy oracle properties.

Last but not least, alasso works well in the case of mixed degree of persistence in predictive regression. It can even adapt to system of predictors that exhibits cointegration by assigning appropriate penalty level inside the system without knowing the identity of these predictors. Lee et al. (2018) establish and demonstrate a simple condition on λ that leads alasso to oracle properties without knowledge of persistence level in advance.

With all the favorable theory at hand, we will embark on testing the performance of alasso in both simulation and real data settings. But first, I will re-establish important results mentioned above in a more concrete manner.

3 The Adaptive Lasso and Its Advantages

In this section, we will discuss the advantage of alasso. The theoretical framework will closely follow Tibshirani (1996); Knight (2008); Zou (2006); Lee et al. (2018). The proof for theorems and theoretical results are not given here, and can be found in referenced papers/articles.

THE MODEL The linear model is assumed to be as following, adapted from

Lee et al. (2018):

$$\begin{aligned}
y_i &= \sum_{l=1}^{p_z} z_{il} \alpha_l^* + \sum_{l=1}^{p_x} x_{il} \beta_l^* + \sum_{l=1}^{p_c} x_{il}^c \phi_l^* + \varepsilon_i \\
&= \mathbf{z}_i' \boldsymbol{\alpha}^* + \mathbf{x}_i' \boldsymbol{\beta}^* + \mathbf{x}_i^c \boldsymbol{\phi}^* + \varepsilon_i \\
&= \mathbf{w}_i' \boldsymbol{\theta} + \varepsilon_i, \\
\mathbf{y} &= \mathbf{w}' \boldsymbol{\theta} + \boldsymbol{\varepsilon}.
\end{aligned} \tag{1}$$

for $i = 1, \dots, n$, where $\mathbf{z}_i = (z_{i1}, \dots, z_{ip_z})'$, $\mathbf{x}_i^c = (x_{i1}^c, \dots, x_{ip_z}^c)'$, and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip_x})'$ represent the stationary, cointegrated, and unit root regressors, respectively, and $p = p_z + p_c + p_x$ is the total number of regressors in the model. \mathbf{w}_i is $p \times 1$ vector of all predictors; $\boldsymbol{\theta}$ is $p \times 1$ vector of all associated coefficients; and \mathbf{w} , \mathbf{y} , $\boldsymbol{\varepsilon}$ are the observation-stacked vectors/matrices. The presence of heterogeneous degrees of persistence and cointegration in our predictors nicely follows practical situation in predicting excess return with multiple valuation ratios and macroeconomic variables.

One deviation from the model laid out in Lee et al. (2018) is that predictors are allowed to have increasingly strong correlation between one another. Specifically, as sample size n increases, the degree of correlation between said predictors also increases and approaches unity. This setting is included to emulate near-singular design matrix phenomenon frequently encountered in stock return predictive regression. Later, we will see that the sequence of tuning parameter λ proposed by Lee et al. (2018) intended for mixed root in predictors is also helpful in combating near-singular design. Formally, define matrix \mathbf{C}_n as in Knight (2008):

$$\mathbf{C}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{w}_i \mathbf{w}_i' \tag{2}$$

that is nonsingular for each n except for when $n \rightarrow \infty$, then

$$\mathbf{C}_n \rightarrow \mathbf{C}, \tag{3}$$

where \mathbf{C} is singular. In practice, near-singular design can be detected by checking whether the smallest eigenvalue of \mathbf{C}_n is small in comparison to its trace. For some sequence $\{a_n\}$ tending to infinity, assume that

$$a_n(\mathbf{C}_n - \mathbf{C}) \rightarrow \mathbf{D}_0 \tag{4}$$

where \mathbf{D}_0 is positive definite on the null space of \mathbf{C} , that is $\mathbf{v}' \mathbf{D}_0 \mathbf{v} > 0$ for non-zero \mathbf{v} with $\mathbf{C} \mathbf{v} = 0$. Another assumption is that near-singularity affect all predictors in our model.

To ensure stationarity of y_i , the effect of non-stationary predictors (including unit root components in cointegrated ones) are kept small using local-to-zero coefficients. This type of coefficient design is also useful to model the weak signal-to-noise ratio in predictive regression (Phillips and Lee, 2013; Lee et al., 2018).

The $p \times 1$ true coefficient $\theta_n^* = (\theta_{jn}^* = \theta_j^{0*}/n^{\delta_j})_{j=1}^p$ where $\theta_j^{0*} \in \mathbb{R}$ is a fixed constant independent of sample size, and $\delta_j \in [0, 1)$. In cases where $\theta_j^{0*} = 0$, δ_j is also set to zero. θ_n^* thus gets smaller with increasing sample size for $\theta_j^{0*} \neq 0$ and $\delta_j \in (0, 1)$, and approach zero as n tends to infinity.

On the front of identification, assumptions about error terms must also be made. While correlation between regression error and the innovation of non-stationary predictors \mathbf{x}_i is allowed, correlation between regression errors and innovation of stationary and cointegrated predictors is excluded (see **Assumption 3.1** and **Remark 3.1**, Lee et al. (2018)).

THE LASSO Lasso is a technique for estimation of linear models that utilize regularization in order to shrink coefficients and perform variable selection at the same time (Tibshirani, 1996). Lasso objective function for model (1) are defined to be

$$\sum_{i=1}^n (y_i - \mathbf{w}_i' \boldsymbol{\theta})^2 + \lambda_n \sum_{j=1}^p |\theta_j|, \quad (5)$$

which is essentially least square with an additional ℓ_1 penalty that helps force estimates of "small" parameter towards zero. Despite introducing potential bias into the estimates, lasso may reduce estimation variance. In the limit, truly "small" parameters are zero with probability tends to 1 while all others are discernibly not zero. Therefore, in cases where true parameters are zero, no biases are produced and variance is reduced, a win-win situation (superefficiency, as termed in Knight (2008)). On the other hand, such regularization causes bias in estimates for non-zero true coefficients while typically does not improve estimation variance considerably. Naturally, we want estimators that achieve superefficiency when $\theta_j^* = 0$ and produce no asymptotic bias otherwise. Studying how estimators behave in the limit regarding the choice of tuning parameter λ is hence a great way to come up with a desirable one. Knight et al. (2000) find that if λ is treated as a sequence dependent on sample size, and design matrix is non-singular, $O(n)$ growth rate of $\{\lambda_n\}$ is sufficient to obtain \sqrt{n} -consistency and non-degenerate limiting distribution. However, in this paper we are more interested in cases where the design matrix is near-singular.

NEAR-SINGULAR DESIGN MATRIX Assume model (1) with \mathbf{C}_n satisfies (2), (3), and (4), that \mathbf{C} is singular and D_0 is positive definite in the null space of \mathbf{C} . Define $b_n = (n/a_n)^{1/2}$ for $\{a_n\}$ satisfies (4), and define Z_n to be

$$Z_n(\mathbf{u}) = \sum_{i=1}^n [(\varepsilon_i - \mathbf{u}' \mathbf{w}_i / b_n)^2 - \varepsilon_i^2] + \lambda_n \sum_{j=1}^p (|\theta_j + u_j / b_n| - |\theta_j|). \quad (6)$$

This equation is a rescaled version of the objective function in (5) with constants subtracted so that convergence is ensured. If $\hat{\boldsymbol{\theta}}$ minimizes (5) then $b_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ minimizes (6). The following theorem is adapted from Knight et al. (2000):

Theorem 1 (adapted from Theorem 1, Knight (2008)) Define $\boldsymbol{\Omega}$ to be a

zero mean multivariate normal random vector such that $\text{Var}(\mathbf{u}'\boldsymbol{\Omega}) = \sigma^2 \mathbf{u}'\mathbf{D}_0\mathbf{u}$ positive for each nonzero \mathbf{u} that satisfies $\mathbf{C}\mathbf{u} = \mathbf{0}$. Let $\hat{\boldsymbol{\theta}}_n$ minimizes (5) for $\lambda_n \geq 0$. If $\lambda_n/b_n \rightarrow \lambda_0 \geq 0$ then

$$b_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \arg \min\{Z(\mathbf{u}) : \mathbf{C}\mathbf{u} = \mathbf{0}\},$$

where

$$Z(\mathbf{u}) = -2\mathbf{u}'\mathbf{W} + \mathbf{u}'\mathbf{D}_0\mathbf{u} + \lambda_0 \sum_{j=1}^p \{u_j \text{sgn}(\theta_j) + |u_j|I(\theta_j = 0)\}.$$

Theorem 1 reveals many interesting insights. With a proper choice of $\{\lambda_n\}$, that is the sequence converges to a finite non-negative value λ_0 , lasso estimate has normal limiting distribution. In the case of $\lambda_0 = 0$, lasso, somewhat trivially, is consistent and has the same limiting distribution as the OLS does since the penalty term vanishes. Less trivial is the case where λ_0 takes on positive value. Suppose all true parameters $\theta_1, \dots, \theta_p$ are non-zero, lasso estimate is biased, and the amount of bias depends on the vector \mathbf{u} and the signs of each element in the vector of true parameter $\boldsymbol{\theta}$. By assumption in Theorem 1, the null space of \mathbf{C} is the space of vectors \mathbf{u} with $u_1 + \dots + u_p = 0$, therefore $\text{Bias}(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}$ if true parameters all have the same sign. In another example case where $\theta_1 \neq 0$ and $\theta_2 = \dots = \theta_p = 0$, the joint limiting distribution of $b_n(\hat{\boldsymbol{\theta}}_{nj} - \theta_j)$ for $j = 2, \dots, p$ will have positive probability mass at $\mathbf{0}$. Since the limiting distribution lies in the null space of \mathbf{C} as in (4), $b_n(\hat{\theta}_{n1} - \theta_1)$ is implied to have positive probability mass at 0 (Knight, 2008). This indicates possible asymptotic bias in the estimates of non-zero parameters. Another downside of singularity in the limit is the slower convergence rate of estimates to their limiting distributions due to $a_n \rightarrow \infty$ and $b_n = (n/a_n)^{1/2}$ hence b_n is only of $o(\sqrt{n})$. The exact margin however, still depends strongly on the growth rate of a_n .

MIXED ROOT Aside from near-singular design matrix, unknown degree of persistence also poses problems for not only OLS, but also for lasso. In the presence of possible unit roots (exact roots are unknown) in regressors, OLS estimates is bias in the limit due to serial dependence in the innovations. With lasso, the variable screening effect is very sensitive to the choice of tuning parameter, and each set of predictors is affected differently (**Corollary 3.7** and **Remark 3.8**, Lee et al. (2018)). As expected, when $\lambda_0 = 0$, lasso's limiting distribution collides with that of OLS since no selection takes place anymore. More interesting are cases where $\lambda_0 \neq 0$. For $\lambda_0 \in (0, \infty)$, only the set of stationary predictors receives selection. In this setting, λ_0 is still too small to have an effect on non-stationary part. Choosing the sequence to growth even faster, where $\lambda_n/\sqrt{n} \rightarrow \infty$ and $\lambda_n/n \rightarrow 0$, drags down convergence rate of estimates for the stationary set while screening still does not hit non-stationary set. Raising growth rate further however starts to introduce inconsistency (**Lemma 3**, Zou (2006)). As a result, it is impossible for lasso to have at the same time both variable screening and consistent estimation due to it having a single penalty for all $I(0)$ and $I(1)$ predictors.

ADAPTIVE LASSO Introduced by Zou (2006), alasso is designed to overcome possible inconsistencies in variable selection of lasso. Alasso involves applying individual weight to the tuning parameter of each regressor. The objective function of alasso is defined as:

$$\sum_{i=1}^n (y_i - \mathbf{w}_i' \boldsymbol{\theta})^2 + \lambda_n \sum_{j=1}^p \hat{\tau}_j |\theta_j|, \quad (7)$$

where $\hat{\tau}_j = |\hat{\beta}_j|^{-\gamma}$ is the individual weight for the corresponding regressor j , with $\hat{\beta}_j$ denoting estimate from another regression method. Popular choices for $\hat{\beta}_j$ include estimates from OLS, ridge regression, or even lasso. γ is a hyperparameter that is, in practice, chosen such that $\gamma \geq 1$ to prevent non-convex optimization problem (Lee et al., 2018). One can also use cross-validation to optimize for γ . The solution for objective function (7) is a vector of alasso estimators $\hat{\boldsymbol{\theta}}^{al}$.

The choice of the vector of weights $\hat{\boldsymbol{\tau}}$ is an important aspect of alasso implementation. Usually, a data driven method such as estimates from another regression scheme is used to obtain $\hat{\boldsymbol{\beta}}$. In our case, we will consider using estimates from OLS and ridge regression due to some favorable properties. First is the OLS, $\hat{\tau}_j = |\hat{\beta}_j^{ols}|^{-\gamma}$. In the case of mixed roots, there is an asymptotic bias in the limiting distribution of OLS estimates due to serial dependence in the innovations since we allow for correlation between regression error and the innovation of non-stationary regressors. Still, OLS estimates converge at the same rate as in the case of stationary regression. For true-zero coefficients, $\hat{\beta}_j^{ols}$ is small so $\hat{\tau}_j$ is large. This put a heavier penalty on such estimates. On the other hand, the weights for predictors with true non-zero coefficients converge to small values, putting less weights on the estimates. In singular settings, however, calculating weights using OLS fails due to the elimination or false inflation of estimates for highly correlated variables. Weight using estimates from ridge regression can be used as a replacement. By setting lambda value close to zero, ridge estimate is argued to be a reasonable approximation for OLS estimate (Knight et al., 2000).

Under the aforementioned choice of weight vector, Theorem 1 can be extended to accommodate for alasso (Knight et al., 2000). Furthermore, for $\lambda_n \rightarrow \infty$, alasso estimator achieve (asymptotical) "oracle properties" for $\lambda_n/\sqrt{n} \rightarrow 0$ and $\lambda_n n^{(\gamma-1)/2}$ (Zou, 2006), which means the alasso estimator

- identifies the right subset model: $P(\hat{M}_n = M^*) \rightarrow 1$
- has the optimal estimation rate, $\sqrt{n}(\hat{\boldsymbol{\theta}}^{al} - \boldsymbol{\theta}^*) \rightarrow N(\mathbf{0}, \boldsymbol{\Sigma}^*)$ where $\boldsymbol{\Sigma}^*$ is the covariance matrix knowing the true subset model.

Here, $M^* = \{j \in \{1, \dots, p\} : \theta_j^* \neq 0\}$ is the set of regressors with true-large coefficients, $\hat{M}_n = \{j \in \{1, \dots, p\} : \hat{\theta}_j \neq 0\}$ is the set of regressors with non-zero estimates (selected regressors). As straightforward as it sounds, the case in our

hand here is a bit more complicated. In the presence of mixed roots, the oracle properties for lasso estimate still exist, albeit with different conditions on λ_n . Lee et al. (2018) deal with this in Theorem 3.4. The optimal λ_n is such that $\lambda_n \rightarrow \infty$ and

$$\frac{\lambda_n}{n^{(1/2) \wedge (1-\gamma.\bar{\delta})}} + \frac{1}{\lambda_n n^{(\gamma-1)/2}} \rightarrow 0, \quad (8)$$

where the wedge denotes the minimum operator, and $\bar{\delta} = \max_{j \leq p} \delta_j$. In practice, $\gamma \geq 1$ is usually chosen to ensure convex optimization for lasso implementation. With this optimal rate, consistent model selection is achieved, and rate of convergence is \sqrt{n} which is optimal for estimates of stationary set, and n for that of non-stationary set. Set $\gamma = 1$, $\hat{\delta} = 1/2$, and use the usual formulation $\lambda_n = c_\lambda b_n n^{1/2}$, the restriction (8) turns out to be

$$c_\lambda b_n + \frac{1}{c_\lambda b_n n^{1/2}} \rightarrow 0, \quad (9)$$

where c_λ is a constant. Any slowly shrinking sequence such as $b_n = (\log \log n)^{-1}$ fulfills this restriction. However, there exists asymptotic bias in the limiting distribution of true-large non-stationary parameters while estimates for true-zero parameters and true-large parameters for stationary regressors are consistent. On the other hand, the problem of near-singularity in the design matrix still needs to be addressed. The asymptotic theory developed by Knight (2008) and explored above is also applicable to lasso, as it allows for different lambda sequence for each estimate. Specifically, to deal with condition (2), (3), (4), more stringent conditions must be put to slow the growth rate of λ_n , similar to the explored lasso case. That is, estimates for stationary regressors' true-large parameters are no longer \sqrt{n} -consistent, and slower than n for the non-stationary case all due to the sequence $b_n = o(\sqrt{n})$. Additionally, singular \mathbf{C} causes further possible bias in the limit of the estimates for all true-large parameters while keeping estimates for true-zero parameters unbiased. Here, I propose a more slowly growing sequence of lambda. Keeping the same construction of λ_n as above, set $b_n = 1/\log(n)^\kappa$, with κ is any positive real number. The reason for a slower growth rate of λ is to not too far exceed the convergence rate of OLS estimators used as weights. In the context of mixed roots, and near-singular design, OLS estimates are generally biased except for estimates for true-large stationary regressors' parameters, and the convergence rate is slower than non-singular design case. Therefore, λ_n for lasso needs not to grow too fast, or else the sequence will dominate convergence of the weight sequence, forcing all estimates toward zero.

EXACT POST-SELECTION INFERENCE On the front of post-selection inference for regularization least square, the prospect is not bright. Recently, Lee et al. (2016) derive a way to compute exact confidence interval for lasso estimators, and also make a package for application in R. However, the results are not applicable to the case at hand here due to the now-familiar problems of mixed persistence and singularity in the limit of design matrix.

In the next section, I will present simulation results to test if the ideas presented in Section 3 has any merit.

4 Simulation study

In this section, the performance of lasso and alasso will be assessed by means of one-step-ahead mean squared forecasting error and variable screening success rate. The scheme will followed closely **Section 4** of Lee et al. (2018), albeit with different specifications for data generating process.

DATA GENERATING PROCESS To emulate both problems we have discussed in previous Section, we set up the DGP as followed. The dependent variable y_i is generated by the process

$$y_i = \gamma^* + \sum_{j=1}^2 z_{ij} \alpha_j^* + \sum_{j=1}^3 x_{ij} \beta_j^* + \sum_{j=1}^4 x_{ij}^c \phi_{ij}^* + u_j,$$

where $\gamma^* = 0.3$ and $\theta^* = (\alpha^*, \beta_n^*, \phi_n^*) = (0, 0.3, 0, -0.4, \frac{1}{\sqrt{n}}, 0.2, -0.2, 0, 0)$. z_{i1} and z_{i2} follow two stationary AR(1) processes with the same autoregressive coefficients $\rho_{z1} = \rho_{z2} = 0.6$. Their error terms exhibit increasing dependency, and are generated from a bivariate normal distribution $MN(\mathbf{0}, \Sigma_n)$ where $\Sigma_n = \begin{pmatrix} 1 & \rho_n \\ \rho_n & 1 \end{pmatrix}$, and $\rho_n = \frac{n}{n+1} \frac{1-\rho_{z1}\rho_{z2}}{\sqrt{(1-\rho_{z1}^2)(1-\rho_{z2}^2)}}$. This construction of Σ_n allows for increasingly strong correlation between two processes z_{i1} and z_{i2} , and the correlation approaches unity as $n \rightarrow \infty$. x_{i1}, x_{i2} , and x_{i3} follow three independent AR(1) processes, with the respective coefficients of $(-0.98, 0.4, 1)$. This is to emulate mixed degrees of persistence in our later real-world example. Lastly, $\mathbf{X}_i^c \in \mathbb{R}^4$ is an I(1) process with cointegration rank 2 based on the VECM, $\Delta \mathbf{X}_i^c = \Gamma' \Lambda \mathbf{X}_{i-1}^c + \mathbf{e}_i$, where the cointegrating matrix $\Lambda = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}$ and the loading matrix $\Gamma = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. In the error term $\mathbf{e}_i = (e_{i1}, e_{i2}, e_{i3}, e_{i4})'$, we set $e_{i2} = e_{i1} + \nu_{1i}$ and $e_{i4} = e_{i3} + \nu_{2i}$ where ν_{1i} and ν_{2i} are independent AR(1) processes with the AR(1) coefficients $(0.2, 0.4)$.

BENCHMARK To assess the performance of alasso estimator, I will pitch it against OLS and lasso estimators. I also add the so-called Oracle OLS estimator, where I only include the predictors that have non-zero coefficient to use in OLS estimators. The sample size settings are chosen to be $n = 40, 80, 160, 250, 500, 1000$. For each n , 1000 replications will be generated and studied to eliminate any non-systematic random error. To choose the initial tuning parameter c_λ for alasso and lasso estimators, another exploratory sample with size 200 will be generated, and 10-fold cross validation will be employed to find the value that minimizes CV-MSE. This exploratory study is to be replicated 100 times, and in the full scale study, we set $c_\lambda = \text{median} \left(c_\lambda^{(1)}, c_\lambda^{(2)}, \dots, c_\lambda^{(100)} \right)$. In the case of alasso, in each settings for n , we construct $\lambda_n = c_\lambda b_n \sqrt{n}$ as discussed in the previous section, where b_n takes on 4 different sequences $\left(\frac{1}{\log(\log(n))}, \frac{1}{\log(n)}, \frac{1}{\log(n)^2}, \frac{1}{\log(n)^3} \right)$. For lasso, λ_n is obtained by multiplying c_λ with $(\sqrt[3]{n}, \sqrt{n}, n)$. Set $g_n^{(al)} = \lambda_n^{(al)} + \frac{1}{\lambda_n^{(al)} n^{1/2}}$.

Table 1: Mean Prediction Square Error

n	Oracle	OLS	Alasso				Lasso		
			$\lambda_n^{(al,1)}$	$\lambda_n^{(al,2)}$	$\lambda_n^{(al,3)}$	$\lambda_n^{(al,4)}$	$\lambda_n^{(pl,1)}$	$\lambda_n^{(pl,2)}$	$\lambda_n^{(pl,3)}$
40	1.3839	1.6221	1.5201	1.4618	1.5088	1.5651	1.7574	1.6333	1.5842
80	1.1545	1.2621	1.5223	1.3125	1.2233	1.2376	1.8643	1.7277	1.6190
160	1.0391	1.0568	1.4095	1.1578	1.0609	1.0501	1.6813	1.5335	1.4523
280	1.0290	1.0524	1.4000	1.1515	1.0587	1.0477	1.6889	1.5824	1.4817
500	1.0644	1.0781	1.4416	1.1750	1.0823	1.0730	1.7494	1.5937	1.4826
1000	0.9641	0.9708	1.2592	1.0260	0.9715	0.9683	1.5879	1.4268	1.3132

Note: Best MPSE for each sample size setting is noted in bold.

Performance of candidate estimators will be assessed using out-of-sample mean prediction square error (MPSE), $E[(y_{T+1} - \hat{y}_{T+1})^2]$ and success rate for variable screening (SR). Let the set of relevant predictors be $M^* = \{j \in \{1, \dots, p\} : \theta_j^* \neq 0\}$ and the estimated active set (contains predictors with non-zero estimates) be $\hat{M} = \{j \in \{1, \dots, p\} : \hat{\theta}_j \neq 0\}$. Success rates for variable screening are defined as:

$$\begin{aligned}
SR &= \frac{1}{p} E \left[\left| \left\{ j : j \in \{1, \dots, p\} : I(\theta_j^* = 0) = I(\hat{\theta}_j = 0) \right\} \right| \right], \\
SR_1 &= \frac{1}{|M^*|} E \left[\left| \left\{ j : j \in \hat{M}, j \in M^* \right\} \right| \right], \\
SR_2 &= \frac{1}{|M^{*c}|} E \left[\left| \left\{ j : j \in M^{*c}, j \in \hat{M}^c \right\} \right| \right],
\end{aligned}$$

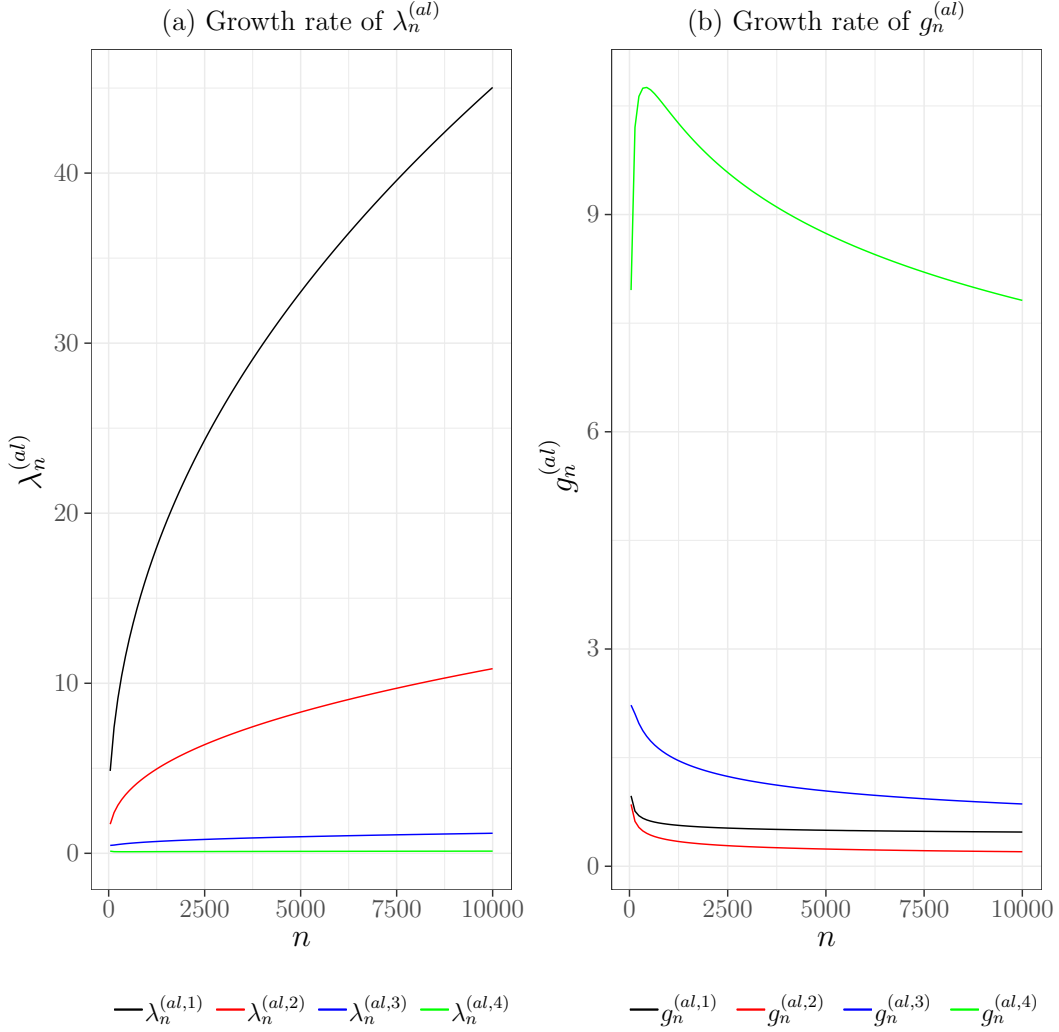
where SR denotes overall success rate of classification into zero coefficients and non-zero coefficients, SR_1 the percentage of the correct selection in the active set, and SR_2 the percentage of correct elimination of the zero coefficients. Expectation is calculated by taking the average of all 1000 replications in each sample size setting.

RESULTS Let $(\lambda_n^{(al,1)}, \lambda_n^{(al,2)}, \lambda_n^{(al,3)}, \lambda_n^{(al,4)}) = c_\lambda^{(al)} \left(\frac{\sqrt{n}}{\log(\log(n))}, \frac{\sqrt{n}}{\log(n)}, \frac{\sqrt{n}}{\log(n)^2}, \frac{\sqrt{n}}{\log(n)^3} \right)$ and $(\lambda_n^{(pl,1)}, \lambda_n^{(pl,2)}, \lambda_n^{(pl,3)}) = c_\lambda^{(pl)}(n, \sqrt{n}, \sqrt[3]{n})$ be the feasible set of lambda sequences where c_λ^{al} and c_λ^{pl} are the initial lambda values for alasso and lasso, respectively. We will first examine the growth rate of each sequence for alasso.

As shown in Figure 1, $\lambda_n^{(al,1)}$ is the fastest-growing sequence that satisfies (9), followed by $\lambda_n^{(al,2)}$, and $\lambda_n^{(al,3)}$, where as $\lambda_n^{(al,4)}$ is the slowest-growing sequence, that it in fact barely converges at all. Additionally, $\lambda_n^{(al,4)}$ also takes too long to converge to zero, and only converges after n reaching around 400. We will later see the effect of different convergence rates on estimators' performance.

Table 1 reports the one-step-ahead MPSE for each of the feasible estimators. In the case of alasso, 4 different lambda sequences are used. According to Table 1, the OLS is a very strong estimator, frequently outperforming both shrinkage estimators, sometimes by non-trivial margin. OLS is especially strong at higher

Figure 1: Growth rate of lambda sequences for alasso



sample size due to its optimal convergence rate. Lasso, on the other hand loses out completely for all candidate λ sequence. At small sample sizes $n = 40$, alasso estimator wins by nontrivial margin for any chosen sequence of tuning parameter. In samples with large size, alasso estimator with tuning parameter $\lambda_n^{(al,3)}$ trails behind OLS, but not too far, while $\lambda_n^{(al,4)}$ gives alasso best MPSE performance overall and even comes close to (unfeasible) Oracle OLS estimator at $n = 1000$. However, this outstanding performance comes at a cost, which is apparent in the next table. Another, expected, drawback of alasso and lasso is the slower convergence rate of estimates as sample size increases, characterized by smaller MPSE differences between increasing sample size settings. Interestingly, slowest-diverging λ sequence reduce MPSE fastest for both of our shrinkage estimators.

Table 2: Variable Screening Performance, Adaptive Lasso

n	SR				SR1				SR2			
	$\lambda_n^{(al,1)}$	$\lambda_n^{(al,2)}$	$\lambda_n^{(al,3)}$	$\lambda_n^{(al,4)}$	$\lambda_n^{(al,1)}$	$\lambda_n^{(al,2)}$	$\lambda_n^{(al,3)}$	$\lambda_n^{(al,4)}$	$\lambda_n^{(al,1)}$	$\lambda_n^{(al,2)}$	$\lambda_n^{(al,3)}$	$\lambda_n^{(al,4)}$
40	0.5689	0.6229	0.6264	0.5957	0.3406	0.5462	0.7404	0.8570	0.8543	0.7188	0.4840	0.5957
80	0.6021	0.6726	0.6832	0.6301	0.3800	0.5914	0.7908	0.8984	0.8798	0.7740	0.5488	0.6301
160	0.6259	0.7167	0.7553	0.6697	0.4100	0.6400	0.8516	0.9320	0.8958	0.8125	0.6350	0.6697
280	0.6321	0.7461	0.7912	0.6919	0.4120	0.6658	0.8820	0.9396	0.9073	0.8465	0.6778	0.6919
500	0.6450	0.7543	0.8168	0.7096	0.4384	0.6808	0.8932	0.9368	0.9033	0.8463	0.7213	0.7096
1000	0.6766	0.7882	0.8518	0.7432	0.4878	0.7202	0.9212	0.9394	0.9125	0.8733	0.7650	0.7432

Note: Best Variable Screening performance for each sample size setting is noted in bold.

Table 3: Variable Screening Performance, Plain-vanilla Lasso

n	SR			SR1			SR2		
	$\lambda_n^{(pl,1)}$	$\lambda_n^{(pl,2)}$	$\lambda_n^{(pl,3)}$	$\lambda_n^{(pl,1)}$	$\lambda_n^{(pl,2)}$	$\lambda_n^{(pl,3)}$	$\lambda_n^{(pl,1)}$	$\lambda_n^{(pl,2)}$	$\lambda_n^{(pl,3)}$
40	0.4458	0.4609	0.4924	0.0034	0.1006	0.2508	0.9988	0.9113	0.7945
80	0.4464	0.4726	0.4948	0.0042	0.0786	0.1976	0.9993	0.9650	0.8663
160	0.4462	0.4814	0.5154	0.0046	0.0784	0.1768	0.9983	0.9853	0.9388
280	0.4444	0.4781	0.5209	0.0002	0.0716	0.1770	0.9998	0.9863	0.9508
500	0.4444	0.4817	0.5303	0.0000	0.0770	0.1846	1.0000	0.9875	0.9625
1000	0.4444	0.4849	0.5380	0.0000	0.0788	0.1932	1.0000	0.9925	0.9690

Note: Best Variable Screening performance for each sample size setting is noted in bold.

Regarding variable selection, reported in Table 3, alasso outperforms lasso in both overall classification (SR) and correct selection in the active set (SR1) for all $\lambda_n^{(al)}$, where lasso leads in term of correct elimination of zero coefficients (SR2) for all $\lambda_n^{(pl)}$. However, this result for SR2 is misleading, because due to fast-diverging lambda sequence, lasso eliminates almost ALL predictors, hence correctly eliminates all predictors with true-zero coefficients as well. This means lasso behaves almost like fitting a constant to the dependent variable. Decreasing the growth rate of $\lambda_n^{(pl)}$ improves screening performance, but only marginally. We can also observe that SR1 ties closely with MSPE performance, as seen in the case of $\lambda_n^{(al,4)}$, which is the slowest growing sequence. It shows that choosing the correct predictors are, to a certain extent, more important than eliminating the irrelevant ones; and the estimates for included-but-irrelevant predictors are small anyway due to small OLS estimates hence large penalty. High SR1 comes at a cost of significantly lower SR2, however, and $\lambda_n^{(al,3)}$ is the best choice for variable screening performance while staying closely behind OLS and $\lambda_n^{(al,4)}$ in term of MPSE. Therefore, one should choose the λ sequence based on one's priority: if the set of predictors exhibits mixed roots and increasing degree of correlation with low signal-to-noise ratio, a slower-growing lambda sequence is preferred for MPSE performance, whereas a relatively faster-growing sequence is better for variable screening effect. Note that with any λ sequence that tends to infinity and satisfies (8), the selection event of alasso estimator is consistent. The difference here is the speed of convergence, which relates to practical usage. Overall, $\lambda_n^{(al,3)}$ is the preferred choice for both prediction and variable screening performance.

5 Application: Goyal's data set

In this section, alasso and lasso estimators are used to predict equity premium using macroeconomics and financial variables. The (updated to 2018) dataset comes from Welch and Goyal (2008).

DATA Here, we use the updated monthly dataset as in (Welch and Goyal, 2008) from the period 01-1927 to 12-2018. The 14 included independent variables as predictors are as followed: *Dividend-price Ratio* (**d/p**), the difference between log of 12-month moving sums of dividends paid on the S&P 500 index and the log of prices of the index itself; *Dividend Yield* (**d/y**), the difference between log of 12-month moving sums of dividends and the log of lagged prices of the S&P 500 index; *Earning-price Ratio* (**e/p**), the difference between log of 12-month moving sums of earnings on the index and the log of prices of the index; *Dividend Payout Ratio* (**d/e**), the difference between the log of dividends and the log of earnings; *Stock Variance* (**svar**), sum of squared daily returns on the S&P 500; *Book-to-market Ratio* (**b/m**), ratio of book value to market value for the Dow Jones Industrial Average; *Corporate Issuing Activity* which is characterized by *Net Equity Expansion* (**ntis**), the ratio of 12-month moving sums of net issues

Table 4: Estimated AR(1) coefficients

Variables	d/p	d/y	e/p	d/e	svar	b/m	ntis
AR(1) Coef	0.9921	0.9921	0.9867	0.9910	0.6322	0.9853	0.9798

Variables	tbl	lty	ltr	tms	dfy	dfr	infl
AR(1) Coef	0.9930	0.9958	0.0441	0.9602	0.9751	- 0.1181	0.4800

Bold number denotes high persistence

by NYSE listed stocks divided by the total end-of-year market capitalization of NYSE stocks; *Treasury Bill Rates* (**tbl**), the 3-month Treasury Bill rates; *Long Term Yield* (**lty**), long-term government bond yield; *Long Term Rate of Returns* (**ltr**), the rate of returns of long-term government bonds; *Term Spread* (**tms**), the difference between the long term yield on government bonds and the Treasury-bill; *Default Yield Spread* (**dfy**), the difference between BAA and AAA-rated corporate bond yields; *Default Return Spread* (**dfr**), the difference between long-term corporate bond and long-term government bond returns; and finally *Inflation Rate* (**infl**). The dependent variable is *excess return*, that is the difference between the continuously compounded return on the S&P 500 index and three-month Treasury bill rate.

In addition to the usual one-month-ahead short-horizon prediction, the long-horizon prediction is also viable since the signal of persistent predictors may amplify over time (Cochrane, 2009). Following Lee et al. (2018), I construct the long-horizon excess return as the sum of continuous compounded monthly excess return on the S&P 500 index. Let h be the length of the forecasting horizon, $h = \frac{1}{12}, \frac{1}{4}, \frac{1}{2}, 1, 2, 3$,

$$LongReturn_i = \sum_{k=i}^{i+12 \times h-1} ExReturn_k.$$

Here, $h = \frac{1}{12}$ stands for one-month-ahead short-horizon prediction, and $h = 1$ stands for one-year-ahead prediction. In each horizon scheme, the corresponding $LongReturn^{(h)}$ becomes the dependent variable for estimation.

We will now look at the results of a few simple test to diagnose problems with our set of predictors. Table 4 shows estimated AR(1) coefficients for all predictors. The numbers show mixed persistence, with some near-unity coefficients. ADF test fails to reject the null hypothesis of unit root in predictors **tbl**, **lty** at 5% level, and **d/p**, **d/y**, **e/p** at 1% level. Further inspection of the residuals by regressing **tbl** on **lty** shows evidence for stationarity. As a result, the two series may exhibit cointegrating relationship. The dependent variable **premium** is stationary with estimated AR(1) coefficient of around 0.0914. We also check if the design matrix is near-singular. First, \mathbf{C}_n is computed as in (2). Smallest absolute eigenvalue of \mathbf{C}_n is many orders of magnitude smaller than its trace, which practically indicates near-singularity. In fact, a preliminary fit using OLS and all available predictors

Table 5: Mean Prediction Square Error

h	MSPE				Percentage relative to OLS			
	OLS	RWwD	Alasso	Lasso	OLS	RWwD	Alasso	Lasso
10-year rolling window								
$\frac{1}{12}$	0.0024	0.0020	0.0022	0.0022	1.0000	0.8574	0.9261	0.9099
$\frac{1}{4}$	0.0039	0.0061	0.0051	0.0052	1.0000	1.5498	1.2905	1.3219
$\frac{1}{2}$	0.0132	0.0130	0.0125	0.0123	1.0000	0.9857	0.9473	0.9321
1	0.0201	0.0254	0.0202	0.0219	1.0000	1.2673	1.0063	1.0921
2	0.0289	0.0444	0.0313	0.0316	1.0000	1.5381	1.0834	1.0950
3	0.0227	0.0585	0.0296	0.0289	1.0000	2.5772	1.3060	1.2711
15-year rolling window								
$\frac{1}{12}$	0.0033	0.0017	0.0018	0.0018	1.0000	0.7484	0.7709	0.7662
$\frac{1}{4}$	0.0090	0.0054	0.0045	0.0048	1.0000	1.6247	1.3412	1.4256
$\frac{1}{2}$	0.0187	0.0120	0.0114	0.0116	1.0000	1.3327	1.2641	1.2956
1	0.0296	0.0260	0.0188	0.0208	1.0000	1.3883	1.0060	1.1118
2	0.0317	0.0517	0.0313	0.0313	1.0000	1.7471	1.0591	1.0587
3	0.0024	0.0738	0.0352	0.0352	1.0000	2.3252	1.1095	1.1098
20-year rolling window								
$\frac{1}{12}$	0.0100	0.0017	0.0017	0.0017	1.0000	0.7266	0.7352	0.7376
$\frac{1}{4}$	0.0212	0.0054	0.0043	0.0046	1.0000	1.4923	1.1942	1.2706
$\frac{1}{2}$	0.0340	0.0120	0.0105	0.0110	1.0000	1.1969	1.0522	1.0962
1	0.0407	0.0260	0.0210	0.0217	1.0000	1.2248	0.9894	1.0225
2	0.0340	0.0517	0.0363	0.0354	1.0000	1.5186	1.0664	1.0404
3	0.0407	0.0738	0.0423	0.0403	1.0000	1.8117	1.0389	0.9893

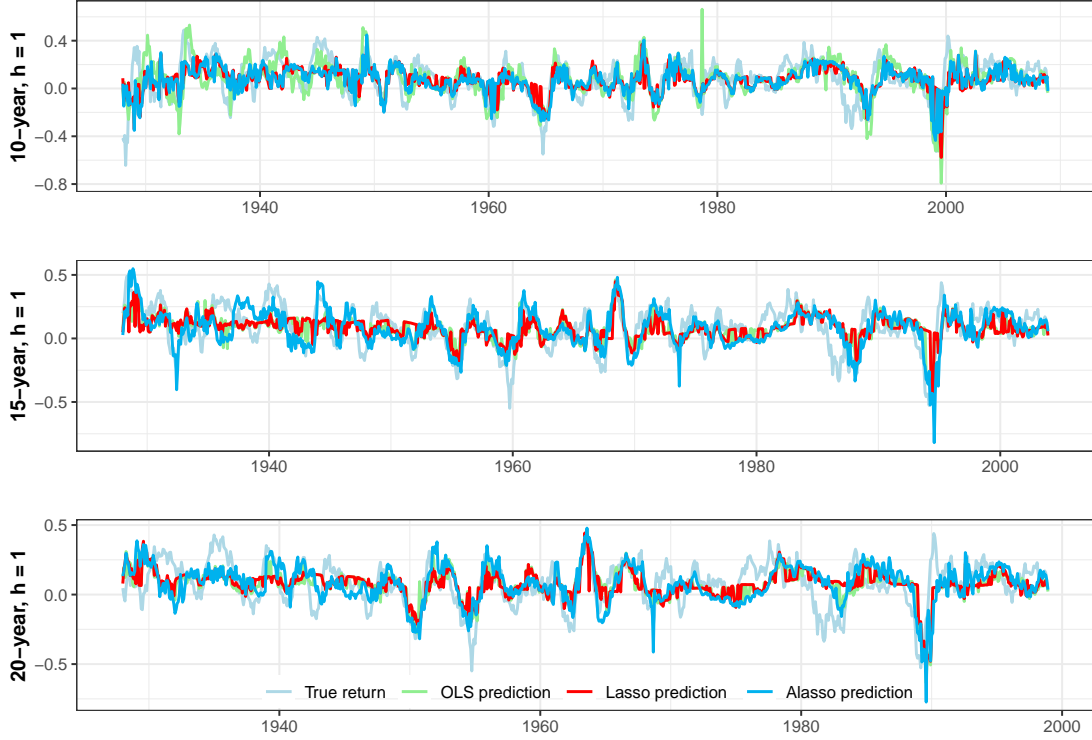
Best forecasting performance is noted in bold

(*kitchen sink* model, Welch and Goyal (2008)) and full sample sees the removal of predictor d/e due to multicollinearity. As a result, in the following implementation of alasso, estimates from ridge regression with near-zero λ will be used instead of OLS estimates to compute the weight vector $\hat{\tau}$.

PERFORMANCE COMPARISON To make a comparison, a *kitchen sink* model with OLS estimates is included. We also include the prediction from historical average of the excess returns $\hat{y}_{n+1} = \frac{1}{n} \sum_{i=1}^n y_i$, denoted Random Walk with Drift (RWwD). The forecasting performance is based on the *one-step-ahead out-of-sample MPSE* and the *percentage* defined as the ratio of one particular method to that of OLS. In total, 3 estimation window settings are in used: 10-year, 15-year, and 20-year rolling window. The lambda value will be determined automatically by 10-fold cross-validation in each estimation window, and one-step-ahead out-of-sample prediction is then made for each window. The estimation window is then move forward by one month and estimation procedure is repeated until the end of the sample.

Table 5 displays the MPSE performance of 4 feasible estimators. OLS is the overall, sometimes only marginal, winner. Alasso and lasso estimators perform well in the one-month-ahead short-horizon prediction, trailing behind RWwD.

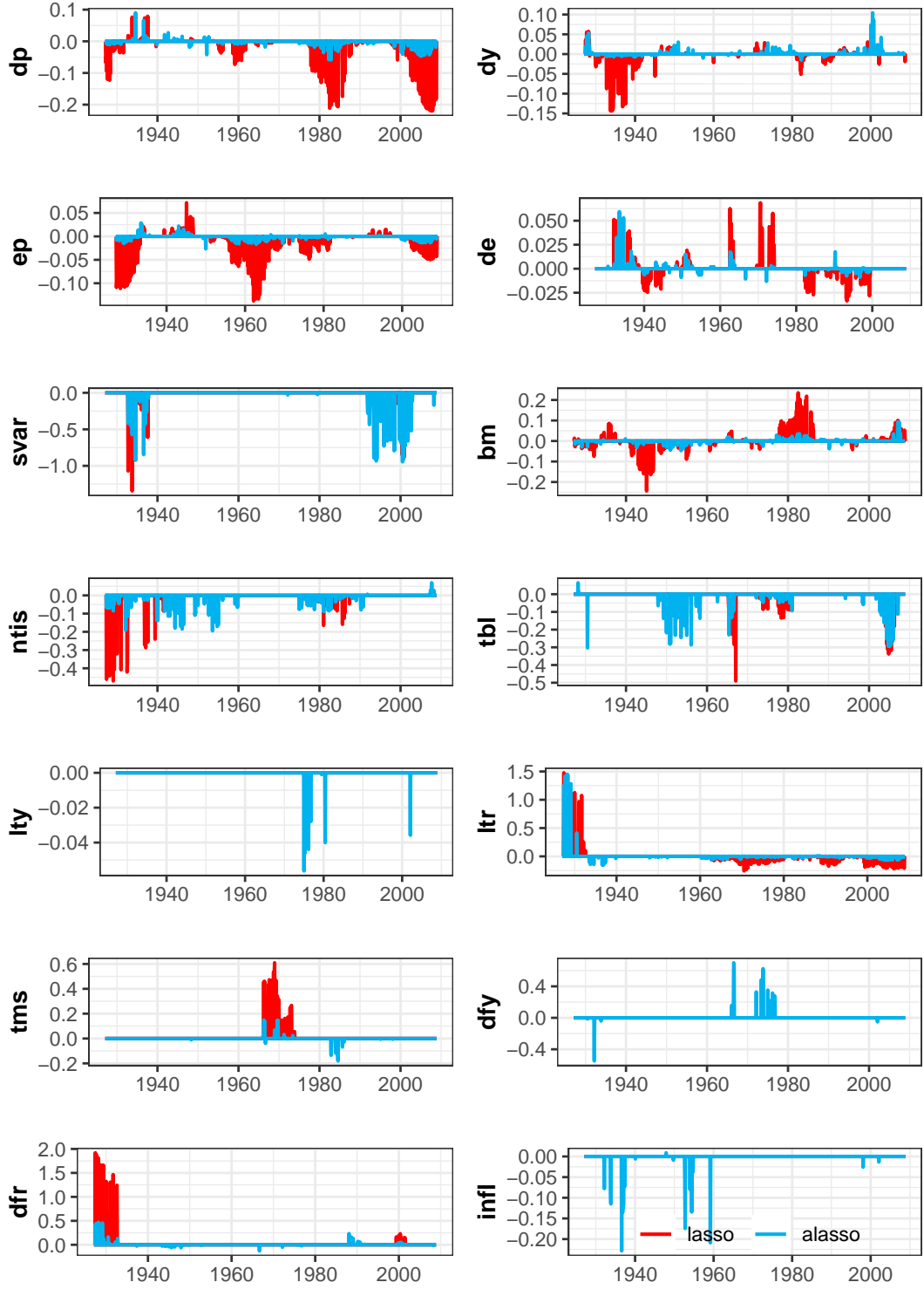
Figure 2: True Returns vs Predicted Return ($h = 1$)



The reason for outstanding performance of RWwD estimator when $h = \frac{1}{12}$ is the stationary nature of monthly excess return. Such nature is detected by alasso and lasso by means of shrinking most coefficients to 0 in most estimation windows (Figure 3), whereas OLS does not possess the mechanics to select predictors. As the prediction horizon h increases, long excess returns increasingly deviate from historical means while signals from predictors accumulate and their contribution can be picked up by other estimators. As sample size increases, performance of alasso and lasso estimators catch up to OLS, especially in longer horizon ($h = 1, 2, 3$). Alasso wins at 20-year rolling window and one-year-ahead prediction performance, whereas lasso is best at 20-year rolling window and three-year-ahead prediction. If estimation window further expands, I expect more improvement in MPSE for the shrinkage estimators.

Figure 2 shows how well predictions from OLS, alasso, and lasso estimators track true excess return. In general, from around 1950 to date, alasso and lasso estimators track true return quite well, and get better the larger the sample size, though lagging behind true changes at some points. Alasso estimator adapts better to spikes while the path for lasso estimates is relatively 'smoother'. OLS estimator, on the other hand, adapts better to spikes for 10-year window, and gives more conservative estimates for 15- and 20-year window. All estimators are able to track the financial crisis, but do not follow the subsequent recovery very well except for alasso in 15-year estimation window.

Figure 3: Estimated Coefficients (10-year rolling window ($h = \frac{1}{12}$))



Regarding screening, both shrinkage estimators select different predictors as estimation windows roll over. Some predictors are selected more frequently than others, and different rolling-windows call for different sets of predictors. The most frequently included predictors by alasso are **d/p**, **d/y**, **ntis**, **tbl**, and **ltr**. Some predictors even change sign as estimation windows roll over. These phenomena suggest the contribution of valuation ratios and macroeconomics predictors to excess return is dynamic. For example, inflation rate's estimated coefficients peak in between the 30s and 60s, coinciding with the period when the U.S experienced great fluctuation in inflation rate. Such high temporary variance may be reflected in the prices of stock market, giving the inflation rate higher contribution to excess return than in the periods of stability. In general, as alasso eliminates more predictors and estimates smaller coefficients, it thus leans toward smaller models that are more parsimonious (Lee et al., 2018) with similar or in few cases better predictive performance than that of lasso and OLS. Illustrations for the case of 10-year rolling window and $h = 1$ are shown in Figure 4.

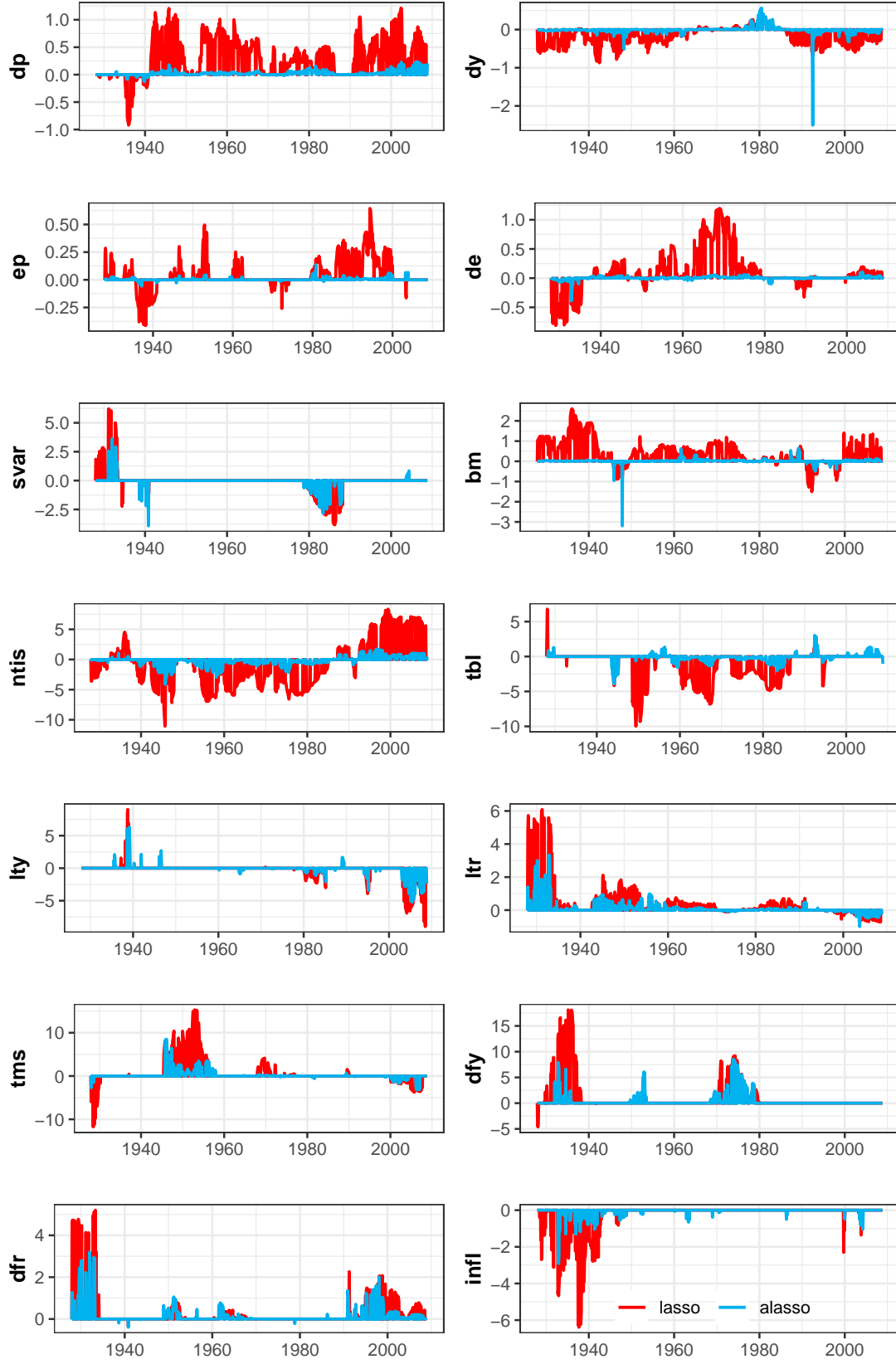
6 Conclusion

This paper discusses asymptotic theories and tests the performance of alasso and lasso estimator in context of two practical problems: mixed roots and near-singular design matrix. In order to achieve that goal, each problems are discussed at length, and an attempt to reconcile theories that address them are made. The theories are promising, especially for alasso: it has the ability to differentiate the true subset of variables even when the model contains a mix of $I(0)$, $I(1)$, and cointegrated processes thanks to an appropriate choices for the weights and tuning parameter λ while conventional lasso can only impose the restriction on $I(0)$ processes if consistency is to be kept (Lee et al., 2018). Moreover, the tuning parameter for alasso can be further restricted to accommodate for near-singular design matrix, giving the estimator valid asymptotic distribution (Knight, 2008). Although the estimates are biased and convergence rate is less than optimal, it can correctly identify the true subset of variables as sample size grows.

Regarding the choice of tuning parameter, its growth rate plays a tremendous role in the performance of our shrinkage estimators. Generally, a slower-growing λ sequence emphasizes picking out variables with true-large coefficients while sacrificing the success rate of identifying irrelevant ones; the opposite is true for a (relatively) faster-growing sequence. As all success rates slowly converge to 1, the "best" choice of λ sequence lies in the hand of the user, who ultimately decides whether including or dropping variables yields more desirable outcome¹. For starter, in this kind of complex predictive environment discussed here, I suggest

¹Keep in mind that sample size for convergence may lie in the thousands. That is what mixed roots and near-singular design bring. However, in some field, this may not troublesome for much longer, as the amount of data available to users is growing very rapidly

Figure 4: Estimated Coefficients (10-year rolling window ($h = 1$))



$\lambda_n = c_\lambda \frac{\sqrt{n}}{\log(n)^2}$ for alasso, and some λ_n that grow slower than \sqrt{n} for lasso.

Results for the empirical study are less than stellar. Our shrinkage estimators only occasionally perform better than OLS in terms of out-of-sample prediction error. Yet, looking past MPSE, alasso and lasso bring more to the table. The variable selection mechanism gives insight to the contribution of each predictors through time, as demonstrated in the inflation rate example above. Besides, more predictors can be included in the model, which may improve predictive power.

For the complex nature of predictors presented in this paper, no exact post-selection inference for the discussed shrinkage estimators exists². This is a truly serious drawback. Due to the incorporated variable selection mechanism, the reasonable thing to do is to pass ever more predictors into the model so that shrinkage estimators can pick out the better ones. As p increases, complications and complex interactions are bound to happen. Further research needs to be made so that a theory that can accommodate such scenario, increasing the practicality of shrinkage estimators in predictive regression.

References

- Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* 37(4), 373–384.
- Cochrane, J. H. (2009). *Asset pricing: Revised edition*. Princeton university press.
- Elliott, G. and J. H. Stock (1994). Inference in time series regression when the order of integration of a regressor is unknown. *Econometric theory* 10(3-4), 672–700.
- Engle, R. F. and C. W. Granger (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, 251–276.
- Fama, E. F. (1990). Stock returns, expected returns, and real activity. *The journal of finance* 45(4), 1089–1108.
- Fama, E. F. and K. R. French (1988). Dividend yields and expected stock returns. *Journal of financial economics* 22(1), 3–25.
- Fama, E. F. and B. G. Malkiel (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance* 25(2), 383–417.
- Granger, C. and P. Newbold (1974). Spurious regressions in econometrics. *Journal of Econometrics* 2(2), 111–120.

²or known to the author

- Granger, C. and P. Newbold (2001). Spurious regression in econometrics. *ECONOMETRIC SOCIETY MONOGRAPHS* 33, 109–118.
- Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *The Review of Financial Studies* 5(3), 357–386.
- Knight, K. (2008). Shrinkage estimation for nearly singular designs. *Econometric Theory* 24(2), 323–337.
- Knight, K., W. Fu, et al. (2000). Asymptotics for lasso-type estimators. *The Annals of statistics* 28(5), 1356–1378.
- Kothari, S. P. and J. Shanken (1997). Book-to-market, dividend yield, and expected market returns: A time-series analysis. *Journal of Financial Economics* 44(2), 169–203.
- Lamont, O. (1998). Earnings and expected returns. *The journal of Finance* 53(5), 1563–1587.
- Lee, J. D., D. L. Sun, Y. Sun, J. E. Taylor, et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44(3), 907–927.
- Lee, J. H., Z. Shi, and Z. Gao (2018). On lasso for predictive regression.
- Meinshausen, N. and P. Bühlmann (2004). Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich.
- Phillips, P. C. (1986). Understanding spurious regressions in econometrics. *Journal of econometrics* 33(3), 311–340.
- Phillips, P. C. (2015). Halbert white jr. memorial jfec lecture: Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics* 13(3), 521–555.
- Phillips, P. C. (2016). Inference in near-singular regression. In *Essays in Honor of Aman Ullah*, pp. 461–486. Emerald Group Publishing Limited.
- 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.
- Pontiff, J. and L. D. Schall (1998). Book-to-market ratios as predictors of market returns. *Journal of Financial Economics* 49(2), 141–160.
- Roll, R. (2002). Rational infinitely lived asset prices must be non-stationary. *Journal of banking & finance* 26(6), 1093–1097.
- Schiller, J. and J. Campbell (1998). Valuation ratios and the long-run stock market outlook. *Journal of*.

- Schwert, G. W. (1990). Stock returns and real activity: A century of evidence. *The Journal of Finance* 45(4), 1237–1257.
- Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics* 54(3), 375–421.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1), 267–288.
- Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21(4), 1455–1508.
- Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?—a study in sampling and the nature of time-series. *Journal of the royal statistical society* 89(1), 1–63.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476), 1418–1429.

Affirmation

I hereby declare that I have composed my Master's thesis "LASSO-based predictive regression for stock returns" independently using only those resources mentioned, and that I have as such identified all passages which I have taken from publications verbatim or in substance. I agree that the work will be reviewed using plagiarism testing software.

Neither this thesis, nor any extract of it, has been previously submitted to an examining authority, in this or a similar form

I have ensured that the written version of this thesis is identical to the version saved on the enclosed storage medium.

A handwritten signature in black ink, consisting of a large, stylized 'U' followed by a cursive 'el', underlined with a horizontal line.

06.01.2019