High-dimensional predictive regression in the presence of cointegration*

Bonsoo Koo† Monash University

Heather Anderson Monash University

Myung Hwan Seo Seoul National University

Wenying Yao Deakin University

Abstract

While a great number of predictive variables for stock returns have been suggested, their prediction power is unstable. We propose a Least Absolute Shrinkage and Selection Operator (LASSO) estimator of a predictive regression in which stock returns are conditioned on a large set of lagged covariates, some of which are highly persistent and potentially cointegrated. We establish the asymptotic properties of the proposed LASSO estimator and validate our theoretical findings using simulation studies. The application of this proposed LASSO approach to forecasting stock returns suggests that the cointegrating relationship among the persistent predictors leads to a significant improvement in the prediction of stock returns over various competing models in the mean squared error sense.

Keywords: return predictability, predictive regression, cointegration, LASSO

JEL: C13, C22, G12, G17

*The authors acknowledge financial support from the Australian Research Council Discovery Grant DP150104292.

†Corresponding author. Author contact details are bonsoo.koo@monash.edu, heather.anderson@monash.edu, myung.h.seo@gmail.com, wenying.yao@deakin.edu.au

1 Introduction

This paper focuses on the predictive regression for stock returns. We propose a Least Absolute Shrinkage and Selection Operator (LASSO) approach to the predictive regression in the presence of cointegration among highly persistent predictors. Specifically, in the predictive regression for stock returns, the information set we consider consists of a mixture of many stationary predictors as well as a few highly persistent predictors that are integrated of order 1 ($I(1)$ series). We find that there exists at least one cointegrating relationship among the persistent predictors commonly used in the finance literature for predicting stock returns. Our empirical study shows significant improvement in return predictability using the proposed LASSO approach compared with the traditional ordinary least squares (OLS), historical average, and AR(1) models in most cases. This could be due in large part to the presence of cointegration. Our empirical study finds that the proposed LASSO approach improves the forecasting of stock returns significantly compared to other methods, including the historical mean, the OLS, and more recent forecasting combinations.

Return predictability has been of perennial interest to financial practitioners and scholars within the economics and finance professions. To this end, empirical studies employ linear predictive regression models, (for instance, refer to Stambaugh (1999), Binsbergen et al. (2010) and the references therein). The predictive regression allows forecasters to focus on the prediction of asset returns conditioned on a set of one-period lagged predictors. In a typical example of predictive regression, the dependent variable is the rate of return on stocks, bonds or the spot rate of exchange, among many assets. The predictors are usually an array of financial ratios and macroeconomic variables.

A considerable amount of economic and financial literature has investigated the predictability of stock returns. Consequently, a number of plausible predictive variables for stock returns are proposed and well-documented (See Koijen and Nieuwerburgh, 2011, for the most recent survey on the literature). Nevertheless, there is no clear-cut evidence as to whether those variables are indeed successful predictors. Rather, there have been heated debates on whether stock returns are predictable, and the jury is still out. Recently, Welch and Goyal (2008) have argued that stock return prediction based on the predictors suggested in the literature is poor in terms of both in-sample fit and out-of-sample forecasting.

Arguably, while predictive regression is the main tool for forecasting stock returns, the predictive regression model has been largely empirically driven and involves several econometric issues. See Phillips (2015) for a summary and recent developments on those issues. Among those, we restrict our attention to two specific problems, which we believe could be tackled in a more systematic way. Firstly, martingale difference features of excess returns are not readily compatible with the highly persistent predictors suggested by the empirical literature. It is well-documented that time series of financial ratios are often highly persistent

and could be (nearly) unit root processes. This is a typical problem of unbalanced regression. Unbalanced regression is difficult to reconcile with the assumptions required for the conventional regression estimation and could render the usual statistical inference invalid. See Campbell and Yogo (2006) and Phillips (2014) for more details.

Secondly, there is no clear guidance that explains why a certain set of predictors is included or excluded in the predictive regression, and under what circumstances some predictors have predictive power and others do not. Given a large pool of covariates available for return predictability, selecting the appropriate covariates for a given period certainly helps to predict the stock return more precisely. However, choosing the right predictors is extremely difficult in practice. As a result, precision is sacrificed for parsimony in empirical studies, and quite commonly a single predictor, or only a handful of predictors at most, are used. This issue is exacerbated by the unstable nature of return predictability over time. The explanatory power of predictors appears to be sensitive to the specific sample period. In this sense, stock return predictability epitomizes two contemporary forecasting challenges: (i) sporadic predictability of predictors; and (ii) incongruence between in-sample predictive content and out-of-sample forecasting ability.[1] To date, few econometric methods and finance theories have been devised to successfully address the aforementioned challenges within a unified framework.

Against this backdrop, this paper employs the highly celebrated LASSO estimation for the predictive regression in the presence of cointegration among predictors. We are particularly inspired by (i) LASSO's capability in model selection, and (ii) cointegration among persistent predictors that could circumvent the unbalanced regression. As will become clear in our empirical study, there exists at least one cointegrating relationship among persistent predictors for stock returns, which lends strong support to our approach based on cointegration. Allowing for cointegration is also compatible with many empirical findings that stock returns might not be predictable in the short run, but predictable in the long run.[2]

Although predictive regression enables us to analyze the explanatory power of each individual predictor, our primary objective is on improving the overall prediction of stock returns given a set of predictors. Therefore, we focus on the out-of-sample forecasting performance by predictors selected via the LASSO estimation. In addition, this paper investigates the large sample properties of the LASSO estimation of the linear predictive regression. We show the consistency of the LASSO estimator of coefficients on both stationary and nonstationary predictors within the linear predictive regression framework. Furthermore, in the presence of cointegration, we derive the limiting distribution of the cointegrating vector under certain regularity conditions, in contrast with the usual LASSO estimation, in which case the limit-

[1] See Stock and Watson (1996, 2003) and Rossi (2013).

[2] Although the unbalanced regression could be tackled by incorporating an accommodating error structure such as (fractionally) integrated noise, it is not easily compatible with the predictive regression framework and its related assumptions.

ing distribution of the estimator is not readily available. To the best of our knowledge, this paper is the first work that establishes the limiting distribution for the LASSO estimator of a cointegrating vector.

The remainder of this paper is organized as follows. We briefly discuss related literature in Section 2. Section 3 introduces predictive regression models, and Section 4 discusses the LASSO selection of relevant predictors within the model framework. In particular, Section 4.1 demonstrates the consistency of the proposed LASSO estimators, and Section 4.2 derives the limiting distribution of the LASSO estimator of a cointegrating vector. Simulation results are discussed in section 5. Also, a real application of our methodology is given in Section 6. Section 7 concludes. All proofs of propositions, lemmas and related regularity conditions are found in Appendix A.

2 Relevant literature

The interest in forecasting stock returns has been unattenuated in the fields of economics and finance since the early work of Dow (1920). However, recent survey papers suggest a shift from finding the appropriate predictors to amalgamating a number of plausible predictors, in order to yield a better forecast for stock returns (Koijen and Nieuwerburgh, 2011; Rapach and Zhou, 2013).

Previously, the majority of the literature focused on whether a certain variable has prediction power for stock returns. Based on the present-value relationship between stock prices and their cash flows, an array of financial variables including financial ratios have been proposed as predictors. Among many others, documented examples are the earnings-price ratio (Campbell and Shiller, 1988), the yield spread (Fama and Schwert, 1977; Campbell, 1987), the book-to-market ratio (Fama and French, 1993; Pontiff and Schall, 1998), the short interest (Rapach et al., 2016) and the dividend-payout ratio (Lamont, 1998). In addition, other types of predictors have followed suit. The assorted macroeconomic variables include the consumption-wealth ratio (Lettau and Ludvigson, 2001) and the investment rate (Cochrane, 1991). Ang and Bekaert (2007) claimed that the return predictability is visible for a short horizon although it is weak for a long horizon. Welch and Goyal (2008) revisited these predictors to verify whether they are able to produce a better forecast than the historical mean of stock returns, but the results do not lend any credence to stock return predictability. Rather, they conclude that most models are unstable at best.

In many empirical models for analyzing stock return predictability, the return is conditioned on only one predictor. This practice leads to potential model mis-specification and misleading statistical inference. Most recently, however, along with statistical and econometric development involving new forecasting techniques, the literature has delved into how to incorporate the large pool of predictors suggested in the literature within a unified frame-

work. This research trend comes with an acknowledgment that these predictors are potentially useful in forecasting stock returns in one way or another, but their prediction power can be very unstable. The high dimension of the predictors poses a significant challenge due to the degrees-of-freedom problem, as noted in Ludvigson and Ng (2007). In addition, temporal instability of their predictive power renders forecasting extremely difficult (Rossi, 2013). In this regard, a number of different approaches have been investigated; for instance, dynamic factor approach (Ludvigson and Ng, 2007), Bayesian averaging (Cremers, 2002), a system of equations involving vector autoregressive models (Pástor and Stambaugh, 2009) and technical indicators (Neely et al., 2014).

Our LASSO approach in the presence of cointegration is an addition to this strand of literature. Since the seminal paper of Tibshirani (1996), the LASSO estimation has gained popularity in various fields of study. See Fan and Lv (2010) and Tibshirani (2011) for the most recent development. The wide variety of predictors available for forecasting stock returns necessitate a wise selection of predictors that capture the quintessential dynamics of future returns. Moreover, the unstable nature of their prediction power at different time periods renders covariate selection even more crucial. In this regard, the LASSO approach in this paper works on both cross-sectional and time-series dimensions.

A few studies have investigated the use of LASSO in the presence of non-stationary variables. Caner and Knight (2013) proposed a unit root test involving Bridge estimators in order to differentiate unit root from stationary alternatives. Liao and Phillips (2015) applied the adaptive LASSO approach to cointegrated system in order to estimate a vector error correction model (VECM) with an unknown cointegrating rank and unknown transient lag order. Quite recently, Kock (2016) further discussed the oracle efficiency of the adaptive LASSO in stationary and non-stationary autoregressions. These papers differ from ours in many aspects. First of all, instead of using autoregressive models, our paper focuses on the predictive regression where the stationary and non-stationary predictors co-exist. Secondly, we allow for the presence of cointegration and have a different objective from the other studies in the literature, which leads to a different approach.

Although LASSO has been widely employed in various fields of studies, not much financial literature utilises this powerful model selection and estimation tool to address stock return predictability. Only two recent papers have emerged. Buncic and Tischhauser (2016) and Li and Tsiakas (2016) suggest that LASSO contributes to better forecasts of stock returns because it enables researchers to select the most relevant factors among a broad set of predictors. Both papers are closely related to the works of Ludvigson and Ng (2007) and Neely et al. (2014). However, both of them fall short in explaining the theoretical grounds behind their empirical results, and they do not take into account the impact of persistent regressors on the linear predictive regression. Moreover, in order to achieve their objectives, both Buncic and Tischhauser (2016) and Li and Tsiakas (2016) imposed coefficient constraints. In contrast,

we do not require any constraint on the sign of coefficients, and the persistence of predictors are explicitly considered in the model. In fact, we provide theoretical credence to the LASSO approach to the linear predictive regression in the presence of persistent predictors.

Our study is also in line with the strand of literature on forecasting under instability based on the predictive regression (see Rossi, 2013, and references therein). As Timmermann (2006) duly noted, there are largely two schools of thought about how to improve forecasting when the underlying model is uncertain. The first suggests employing an approach using various forecasting combinations, which has been actively investigated (see Elliott et al., 2013; Koo and Seo, 2015, among many others). The other one considers the entire set of possible predictors in a unified framework, allowing the data to select the best combination of predictors in an attempt to search for the best forecasting model. Most literature based on ℓ_1-regularization, including the LASSO, falls under this approach.[3] This paper addresses the unstable nature of stock return predictability by utilising the automated selection of the LASSO at different points in time by rolling the estimation window within the sample period.

3 Econometric framework: Predictive regression

Our econometric framework is a linear predictive regression in which a large set of conditioning variables is available, but the prediction power of each variable could be unstable.[4] Some of the conditioning variables could be non-stationary, *i.e.* $I(1)$ variables and might be cointegrated. Our model is framed in the context of stock return predictability because forecasting stock returns epitomizes the issues we are attempting to address: (i) possibly large set of conditioning variables; (ii) some of them are quite persistent; and (iii) the prediction power of each conditioning variable is unstable. Let us consider the following linear predictive regression model:

$$y_t = x_{t-1}'\boldsymbol{\beta} + z_{t-1}'\boldsymbol{\alpha} + u_t, \tag{1}$$

where $\beta \in \mathbb{R}^k$ for a fixed k with $x_t \sim I(1)$, and $\alpha \in \mathbb{R}^p$ for $p \to \infty$ with $z_t \sim I(0)$. $I(q)$ denotes an integrated process of order q. The high persistence of x_t is compatible with the property of many commonly used stock return predictors, such as the dividend-price ratio and the earning-price ratio.[5] We use z_t in model (1) to represent the other group of less persistent return predictors, for example the long term bond rate and inflation. See Neely et al. (2014) and Ludvigson and Ng (2007) for a pool of plausible financial and macroeconomic predictors.

[3] The former of the two is pooling forecasts whereas the latter is pooling information. See Timmermann (2006) for pros and cons of two different approaches.

[4] Our model is intrinsically an array structure, where the true coefficient values and the sparsity structure are allowed to be varying along sample size n. However, we suppress this dependence on n in our notation to ease notational complexity. Furthermore, one can view the changing values of the cointegrating vector β as the weightings changing among multiple cointegrating relationship in x_t.

[5] Welch and Goyal (2008) showed that many return predictors including the treasury bill rate, the long term bond yield and the term spread are also highly persistent, with a first order autocorrelation above 0.99.

Model (1) is in line with the most recent literature on stock return predictability. It is a multi-factor model employed in the empirical asset pricing literature. The primary objective is to estimate the conditional mean of y_t, given a mixture of possibly high-dimensional stationary and fixed number of non-stationary conditioning variables available at time $t-1$. In the context of the present paper, the prediction objective y_t is the rate of return on a broad market index over the risk free rate, commonly referred to as the equity premium. It is commonly accepted that the equity premium displays martingale difference features, whereas many return predictors are quite persistent. The stationarity feature of the martingale difference sequence on the left-hand side of model (1) hardly matches with the highly persistent predictors on the right-hand side. Furthermore, a close examination of the historical data of various financial ratios reveals a strong co-movement among these persistent variables, which lends strong credence to the existence of at least one cointegrating relationship among them.

Without the loss of generality in the forecasting context, we assume that there is at least one cointegrating relationship within x_t. Specifically, the cointegrating vector $\boldsymbol{\delta}$ and the corresponding constant c are such that $x_t'\boldsymbol{\beta} = x_t'\boldsymbol{\delta}c$. For identification when it comes to estimating $\boldsymbol{\delta}^0$ and c^0, we standardize $\boldsymbol{\delta}$ such that $\hat{\boldsymbol{\delta}} = \frac{\hat{\boldsymbol{\beta}}}{\|\hat{\boldsymbol{\beta}}\|}$ and $\hat{c} = \|\hat{\boldsymbol{\beta}}\|$ for some generic norm $\|\cdot\|$.

4 LASSO selection of predictors

This section examines an automated selection of predictors among many covariates whose cardinality can be even larger than the sample size under the sparsity assumption. Particularly, we propose a LASSO estimation in which a ℓ_1-type penalty is applied for this purpose. Note that in a matrix form, model (1) can be written as

$$Y = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \mathbf{u}. \tag{2}$$

The LASSO estimator for $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ can be obtained by

$$(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\alpha}}')' = \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\arg\min} \|Y - X\boldsymbol{\beta} - Z\boldsymbol{\alpha}\|_2^2/n + \lambda(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\alpha}\|_1), \tag{3}$$

where $\|\cdot\|_m$ denotes a ℓ_m norm and $\lambda \sim \ln p/\sqrt{n}$.

Equation (3) describes the usual LASSO objective function regardless of the presence of a cointegrating relationship. We investigate how the presence of a cointegrating relationship

affects the LASSO estimation approach by reparameterisation. With $c = \|\boldsymbol{\beta}\|_1$ and $\boldsymbol{\delta} = \boldsymbol{\beta}/c$,

$$\begin{aligned} Y - X\boldsymbol{\beta} - Z\boldsymbol{\alpha} &= \mathbf{u} - X(\boldsymbol{\beta} - \boldsymbol{\beta}^0) - Z(\boldsymbol{\alpha} - \boldsymbol{\alpha}^0) \\ &= \mathbf{u} - X(\boldsymbol{\delta} - \boldsymbol{\delta}^0)c - X\boldsymbol{\delta}^0(c - c^0) - Z(\boldsymbol{\alpha} - \boldsymbol{\alpha}^0) \\ &= \mathbf{u} - \frac{\tilde{X}}{\sqrt{n}}\sqrt{n}(\boldsymbol{\delta} - \boldsymbol{\delta}^0) - W(c - c^0) - Z(\boldsymbol{\alpha} - \boldsymbol{\alpha}^0) \\ &= \mathbf{u} - D(\boldsymbol{\theta} - \boldsymbol{\theta}^0), \end{aligned}$$

where $\boldsymbol{\theta} = (\sqrt{n}\boldsymbol{\delta}', \boldsymbol{\gamma}')'$ with $\boldsymbol{\gamma} = (c, \boldsymbol{\alpha}')'$ and $D = (\tilde{X}/\sqrt{n}, G)$ with $\tilde{X} = Xc$ and $G = (W, Z)$. In addition, the ℓ_1 penalty terms can be rewritten as

$$\lambda(\|\boldsymbol{\beta}\|_1 + \|\boldsymbol{\alpha}\|_1) = \lambda(\|c\|_1 + \|\boldsymbol{\alpha}\|_1) = \lambda\|\boldsymbol{\gamma}\|_1,$$

with a given constraint $\|\boldsymbol{\delta}\|_1 = 1$ by construction.

Consequently, in the presence of a cointegrating relationship, the LASSO estimator for $\boldsymbol{\theta}$ can be represented by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}, \|\boldsymbol{\delta}\|_1 = 1}{\arg\min} \|\mathbf{u} - D(\boldsymbol{\theta} - \boldsymbol{\theta}^0)\|_2^2/n + \lambda\|\boldsymbol{\gamma}\|_1. \tag{4}$$

Note that (4) is a constrained minimisation problem given that the estimator for $\boldsymbol{\theta}$ is obtained under the constraint, $\|\boldsymbol{\delta}\|_1 = 1$, i.e. the constraint on the parameter space to which $\boldsymbol{\delta}$ belongs is imposed. Also, note that $\lambda\|\boldsymbol{\gamma}\|_1$ can be replaced with $\lambda\|\boldsymbol{\theta}\|_1$ without changing the minimisation problem in this set-up. Notably, our approach reframes the stock return predictability in a variable selection context in such a way that we can use the standard LASSO algorithm and its variations readily available without any modification.

In what follows, we establish the large sample properties associated with the proposed LASSO estimator, $\hat{\boldsymbol{\theta}} = (\sqrt{n}\hat{\boldsymbol{\delta}}', \hat{\boldsymbol{\gamma}}')'$. We start with the consistency of the proposed estimator and its convergence rate and proceed to the large-sample distribution of the LASSO estimator for the cointegrating vector $\boldsymbol{\delta}$. All proofs for our theoretical results are provided in Appendix A with corresponding regularity conditions. It is worth noting that the conditions are quite general, accounting for time series features of the data and the model in Section 3.

4.1 Consistency and rate of convergence

Proposition 1 *Under regularity conditions **A1**-**A3** specified in Appendix A,*

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_1 = O_p\left(\frac{s_0 \ln p}{\sqrt{n}}\right), \tag{5}$$

where s_0 is the number of non-zero elements in $\boldsymbol{\theta}^0$ and p is allowed to tend to infinity.

Proof. See Appendix A. ■

Proposition 1 ensures that LASSO can serve as a predictor-selection tool in the linear predictive regression for forecasting stock returns even when there are two different types of covariates, either stationary (z_t) or nonstationary (x_t), and the dimension for z_t is potentially very large. This property is essential for stock return predictability because the number of plausible return predictors is very large and could be nonstionary (financial ratios) or stationary (macroeconomic variables). LASSO is an automated method that selects relevant predictors under a sparsity assumption so as to achieve parsimony without sacrificing precision.[6] This is more important when the number of observations is relatively small compared to the number of predictors available.

Proposition 1 shows that the convergence rate for $\hat{\boldsymbol{\theta}}$ is slower than $O_p(\sqrt{\ln p/n})$, the usual convergence rate of the LASSO estimator for the linear models, due to the dependence structure of the data.

4.2 Large sample distribution of an estimator for the cointegrating vector

This section discusses the large sample distribution of the proposed estimator for the cointegrating vector, $\boldsymbol{\delta}$. From (3), when $\hat{c}$ and $\hat{\boldsymbol{\alpha}}$ are fixed, we can focus on the estimator for $\boldsymbol{\delta}$, the coefficient associated with $\tilde{X}\hat{c}$. The objective function (3) becomes

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}, \|\boldsymbol{\delta}\|=1}{\arg\min} \|\tilde{Y} - \tilde{X}\boldsymbol{\delta}\|_2^2/n, \tag{6}$$

where $\tilde{Y} = Y - Z\hat{\boldsymbol{\alpha}}$ and $\tilde{X} = X\hat{c}$. It is worth noting that there is no shrinkage on the cointegrating vector due to the restriction on the parameter space, $\|\boldsymbol{\delta}\| = 1$. This is quite intuitive because the coefficients constituting the cointegrating vector are not penalised by the usual LASSO estimation in the presence of a cointegrating relationship. This is more compelling for the return predictive regression, in which the excess return (prediction objective) is stationary while some of the predictors are persistent. As seen from (6), we can derive the large sample distribution of $\hat{\boldsymbol{\delta}}$ based on the standard constrained least squares regression models.

Proposition 2 *The large sample distribution of appropriately scaled* $\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0$ *can be obtained by the minimizer of the limit objective function* $D(b)$ *such that*

$$D(b) = -2c_0 b' \left[\int_0^1 \mathscr{W}(r) d\mathscr{U}(r) + \Lambda \right] + c_0^2 b' \int_0^1 \mathscr{W}(r)\mathscr{W}(r)' dr b,$$

with

$$\Lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n-1} \sum_{j=1}^{i-1} \mathrm{E}(\varepsilon_{i-j} u_{i+1}),$$

[6] Here, sparsity means that the number of truly relevant predictors is small among many candidate variables.

where $\varepsilon_t = x_t - x_{t-1}$ *and* $(\mathscr{W}(\cdot), \mathscr{U}(\cdot))'$ *is the weak limit of the partial sum process of* $(\varepsilon_t, u_t)'$.

Under regularity conditions ***A1-A3*** *specified in Appendix A,*

$$n(\hat{\delta} - \delta_0) \xrightarrow{d} \underset{b:sgn(\delta_0)'b=-\sum_j |b_j| \mathbb{1}_{\{\delta_{0j}=0\}}}{\arg\min} D(b). \tag{7}$$

Proof. See Appendix A. ■

Proposition 2 allows us to make an inference for the cointegration vector δ. Proposition 2 is nonstandard because there exists a restriction on the parameter space. If it were not for the restriction on the parameter space, we could obtain the closed-form expression by deriving the minimizer of the limit objective function $D(b)$ in Proposition 2. Nevertheless, we can easily simulate the limit distribution for the inference purpose.[7] Lastly, the conditions required for Proposition 2 are quite general. For instance, we do not require that the error process is the martingale difference sequence. It suffices that $\{\varepsilon_t\}$ and $\{u_t\}$ are strong-mixing.

5 Simulation study

In this section, we use a simulation study to investigate the forecasting performance of LASSO when there exists a mixture of stationary and cointegrated non-stationary predictors.

5.1 Simulation design

For our simulation study, we consider the standard linear predictive regression model outlined in (1):

$$y_t = \mathbf{x}_{t-1}'\boldsymbol{\beta} + \begin{pmatrix} 1 & \mathbf{z}_{t-1}' \end{pmatrix} \begin{pmatrix} 0.15 \\ \boldsymbol{\alpha} \end{pmatrix} + u_t, \tag{8}$$

where the dimension of regressors $(\mathbf{x}_t', \mathbf{z}_t')'$ is fixed, $\mathbf{x}_t \sim I(1)$ and $\mathbf{z}_t \sim I(0)$. We simulate $\mathbf{x}_t$ and $\mathbf{z}_t$ from the following processes:

$$\Delta \mathbf{x}_t = \begin{pmatrix} 0.2 & 0.5 & 0 & 0 \end{pmatrix}' \begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix} \mathbf{x}_{t-1} + \varepsilon_{x,t}, \tag{9}$$

$$\mathbf{z}_t = 0.6\, \mathbf{I}_6\, \mathbf{z}_{t-1} + \varepsilon_{z,t}. \tag{10}$$

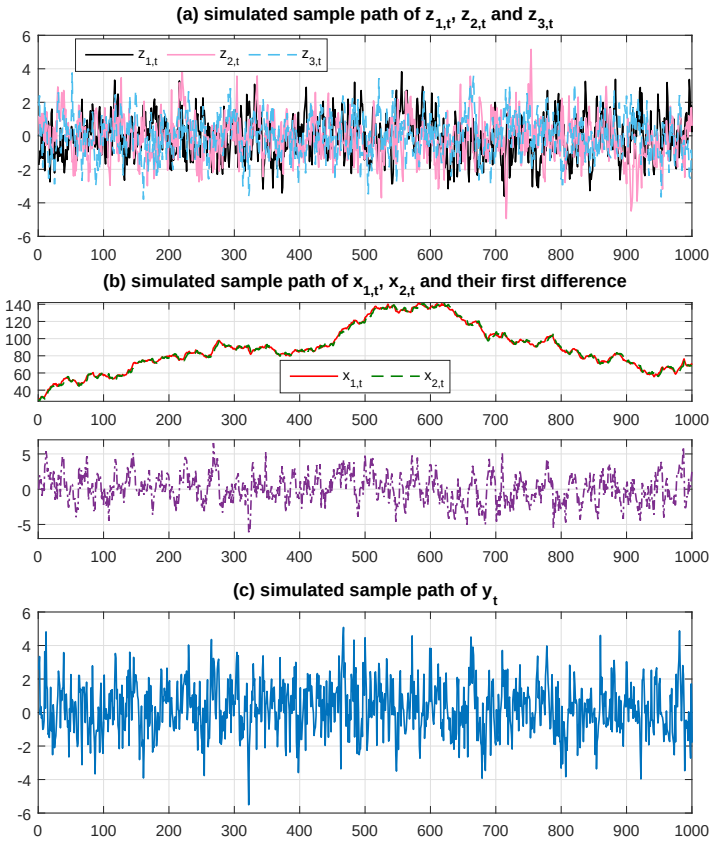
Note that (9) is the usual vector error correction representation in which the first part represents a $k \times \ell$ loading matrix (ℓ: cointegrating rank) while the second part denotes the cointegrating vector. Hence, $\mathbf{x}_t$ is a 4-dimensional $I(1)$ process with cointegrating rank 1 and cointegrating vector $\delta = (1, -1, 0, 0)'$. $\mathbf{z}_t$ is a 6-dimensional stationary AR(1) pro-

[7] The limit distribution cannot be tabulated as it is not asymptotically pivotal but can be simulated. In particular, the restriction on b can be imposed by replacing δ_{0j} with $\tilde{\delta}_j = \hat{\delta}_j \mathbb{1}_{\{|\hat{\delta}_j| \geq c_n\}}$ for some $c_n \to 0$ and $n \times c_n \to \infty$. Due to the super-consistent convergence rate of $\hat{\delta}$, this event is equivalent to the restriction on the parameter space with a probability approaching one.

cess. Coefficients in the predictive regression model (1) are set to be $\boldsymbol{\beta} = c\boldsymbol{\delta} = 0.4\boldsymbol{\delta}$, and $\boldsymbol{\alpha} = (-0.1, 0.02, 0.7, 0, 0, 0)'$. All of the error terms $(u_t, \varepsilon_{x,t}, \varepsilon_{z,t})$ are i.i.d $\mathcal{N}(0,1)$ and uncorrelated contemporaneously.

We start with the simple data generating process (DGP) outlined in equations (8)-(10) by maintaining a modest level of autocorrelation in the stationary variables $\boldsymbol{z}_t$. The strength of persistence in the non-stationary variables $\boldsymbol{x}_t$ is also relatively weak. Parameters in this DGP are chosen such that we preserve a weak dependence structure. Figure 1 shows one simulated sample path of $\{y_t, x_{1,t}, x_{2,t}, z_{1,t}, z_{2,t}, z_{3,t}\}$ with $T = 1000$ observations. Panel (a) shows that $\{z_{1,t}, z_{2,t}, z_{3,t}\}$ display the usual stationary time series, whereas $I(1)$ variables, $x_{1,t}$ and $x_{2,t}$ are cointegrated with the cointegrating vector $(1, -1)'$ in panel (b). Panel (c) of Figure 1 confirms that the $I(1)$ variables $\mathbf{x}_t$ affect y_t through the cointegrating relationship, and hence y_t is a stationary process.

Figure 1: One of the simulated sample paths, $T = 1000$



5.2 Estimation

We consider five different estimation methods. We omit the constant term henceforth for ease of notation, although the constant is included in our simulation study. Firstly, we use the OLS to estimate $(\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ and evaluate the one-step-ahead forecasting error. Secondly, we predict y_t using its historical average $\hat{y}_{T+1} = \frac{1}{T}\sum_{i=1}^{T} y_i$. The third method is to estimate y_t using an AR(1) model:

$$y_t = \rho\, y_{t-1} + u_t. \tag{11}$$

For the fourth method, we employ a forecast combination à la Timmermann (2006). More specifically, we compute the combined forecast, $\hat{y}_{T+1}$, as an equally weighted average of all forecasts that are separately obtained from distinctive linear predictive regression models, each conditioning on only one predictor:

$$\hat{y}_{T+1} = \frac{1}{m}\sum_{m=1}^{M} \hat{y}^{m}_{T+1}, \tag{12}$$

where $M = k + p$ is the total number of predictors, and $\hat{y}^{m}_{T+1}$ is a forecast based on a bivariate linear predictive regression that conditions on m-th predictor only.

Finally, we apply a standard LASSO method with the following objective function:

$$\arg\min_{\boldsymbol{\beta},\boldsymbol{\alpha}} \left\{ \frac{1}{n}\sum_{t=1}^{n} \left(y_t - \boldsymbol{x}'_{t-1}\boldsymbol{\beta} - \boldsymbol{z}'_{t-1}\boldsymbol{\alpha}\right)^2 + \lambda\|\boldsymbol{\beta}\|_1 + \lambda\|\boldsymbol{\alpha}\|_1 \right\}, \tag{13}$$

where $\|\cdot\|_1$ is the L_1-norm. The penalty term λ is estimated using the cross validation algorithm given by Qian et al. (2013). We present results for two estimated λ's: the one that minimizes the cross-validated error loss (LASSO_{min}) as well as the largest λ that produces an error that is smaller than the minimized error loss plus its one standard deviation (LASSO_{1se}). The latter choice is meant to mitigate in-sample overfitting for LASSO by selecting a bigger penalty term.

We consider different sample sizes to evaluate the forecasting performance of each method. For $T = 100$, 200, 400, and 1000, we compare the one-step-ahead forecasting errors of the four estimation methods.

To evaluate the comparative forecasting performance between different methods, we consider two measures: forecasting mean squared error (FMSE) and out-of-sample R^2. The former is defined as the expected squared difference between the actual realisation and the one-step-ahead forecast: $\text{FMSE} := E[y_{T+1} - \hat{y}_{T+1}]^2$. The latter is defined as $R^2{:=}1 - FMSE_i/FMSE_{OLS}$ for each model i. We set the OLS method as our benchmark when calculating the out-of-sample R^2 because we want our simulation study to reflect the predictive regression on return predictability of the financial ratios studied in our empirical study.

The choice of estimation methods considered in this paper merits further explanation. Not

only are those methods employed in many empirical studies but also they can shed light on the trade-off between estimation bias and statistical uncertainty around estimation embedded in forecasts. We note that the OLS and LASSO estimation methods are based on the correct specification while the historical average (the constant model henceforth) and AR model are misspecified. We intend to investigate the impact of the bias induced by the LASSO estimation in comparison with the constant and AR models and the reduction in the estimation variance LASSO could achieve in comparison with the OLS. We choose the AR model instead of the random walk model because the dependent variable $\{y_t\}$ is stationary, and hence the random walk is not compatible with the simulated data. Furthermore, the forecasting performance of the random walk model is quite poor. Finally, we include forecast combination as one of the competing model strategies because it has been widely recognized to improve forecast performance compared to the individual model (Timmermann, 2006; Rapach et al., 2010). Similar to LASSO, the combining forecast model approach alleviate the impact of model uncertainty and parameter instability.

5.3 Discussion on the simulation results

Table 1 presents FMSE and out-of-sample R^2 of the forecasting strategies based on 1000 repetitions. Overall, both LASSO_{min} and LASSO_{1se} estimations always produce smaller FMSEs than the benchmark OLS forecasts across four different sample sizes, and hence positive values of the out-of-sample R^2. On the other hand, the historical average, AR(1) model and forecast combination always lead to inferior forecasting performance compared to the OLS benchmark. In particular, the forecast combination does not appear to be successful in this case. One possible reason is that the data generating process (DGP) has a stable dependence structure. We can infer that the bias induced by adopting the LASSO would be small whereas the reduction in the estimation variance would be substantial.

We conjecture that the LASSO approach renders the presence of a cointegrating relationship conspicuous due to the shrinkage imposed on the small or negligible but nonzero coefficients, which could lead to better forecasting performance. Therefore, as the number of observations increases, the forecasting performance based on the OLS gets closer to that of the LASSO. This is confirmed from Table 1. When we have a small sample size $T = 100$, the improvement in the FMSE of LASSO_{1se} over the OLS benchmark, as measured by the out-of-sample R^2, is 5%. However, as the sample size increases to $T = 1000$, the out-of-sample R^2 of LASSO_{1se} reduces to 3.56%.

Finally, we note that LASSO_{1se} performs better than LASSO_{min}, which corroborates that the LASSO in some cases suffers from the well-known overfitting issue. Overfitting is a data-dependent problem. One solution recommended by Qian et al. (2013) is to use a larger penalty term λ than the one that minimizes the cross-validated error, which is what we employed in the LASSO_{1se} case. In what follows, we use an alternative DGP to conduct the

Table 1: Forecasting performance of different model strategies

T	OLS	Average	AR(1)	F-comb	LASSO_{min}	LASSO_{1se}
FMSE						
100	2.1118	2.5081	2.1963	2.3719	2.0544	**2.0055**
200	2.0851	2.4927	2.1535	2.3862	2.0579	**1.9936**
400	2.1003	2.4787	2.2089	2.3930	2.0869	**2.0277**
1000	1.9272	2.3709	2.0333	2.2741	1.9181	**1.8586**
Out-of-sample R^2						
100		-0.1877	-0.0400	-0.1232	0.0272	**0.0503**
200		-0.1955	-0.0328	-0.1444	0.0130	**0.0439**
400		-0.1801	-0.0517	-0.1394	0.0064	**0.0346**
1000		-0.2303	-0.0551	-0.1800	0.0047	**0.0356**

[a] The out-of-sample R^2 is defined as $R^2 = 1 - FMSE_i / FMSE_{OLS}$ for each model i.

same simulation exercise, in which case overfitting does not appear to have much influence on the forecasting performance of LASSO.

5.4 Robustness: Alternative DGPs

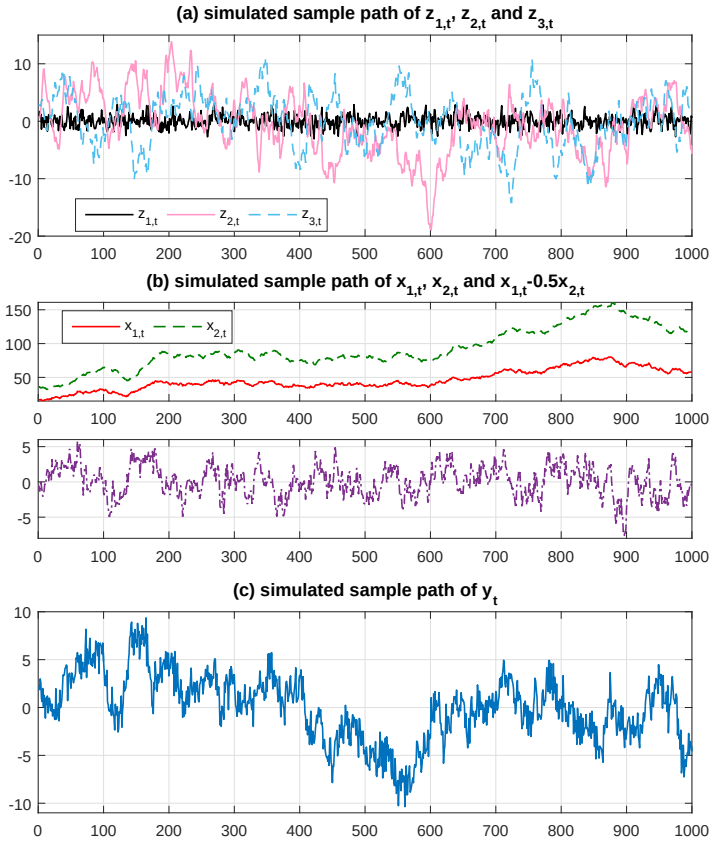
We consider an alternative DGP with richer dynamics to examine the robustness of our simulation study. This might be more compatible with the financial or macro data observed in reality. The specification of the alternative DGP is as follows.

$$\Delta \mathbf{x}_t = \begin{pmatrix} 0 & 0.3 & 0 & 0.2 \end{pmatrix}' \begin{pmatrix} 1 & -0.5 & 0 & 0 \end{pmatrix} \mathbf{x}_{t-1} + \varepsilon_{x,t}, \tag{14a}$$

$$\begin{aligned} \mathbf{z}_t = & \begin{pmatrix} 0.3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.31 & 0.29 & 0.12 & 0.04 & 0 \\ 0 & -0.21 & 1.25 & -0.24 & 0.04 & 0 \\ 0 & 0.07 & 0.03 & 1.16 & 0.01 & 0 \\ 0 & 0.08 & 0.27 & -0.07 & 1.25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.3 \end{pmatrix} \mathbf{z}_{t-1} \\ & + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.35 & -0.28 & -0.07 & -0.02 & 0 \\ 0 & 0.19 & -0.26 & 0.24 & -0.05 & 0 \\ 0 & -0.07 & -0.02 & -0.16 & 0.01 & 0 \\ 0 & -0.13 & -0.23 & 0.03 & -0.31 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \mathbf{z}_{t-2} + \varepsilon_{z,t}, \end{aligned} \tag{14b}$$

$$y_t = x'_{t-1} \begin{pmatrix} 0.3 & -0.15 & 0 & 0 \end{pmatrix}' + z'_{t-1} \begin{pmatrix} -0.1 & 0 & 0.02 & 0.4 & 0 & -0.2 \end{pmatrix}' + u_t. \tag{14c}$$

Figure 2: One of the simulated sample paths, $T = 1000$



The most distinct feature of this DGP is that the stationary variables $\mathbf{z}_t$ follows a VAR(2) process with a more complicated dependence structure, which is partly taken from an estimated VAR(2) process from US macroeconomic data (see Martin et al., 2012, p. 485, example 13.19). Therefore, this DGP should well represent data obtained from the real economy. We plot one simulated sample path with size $T = 1000$ in Figure 2. Note that the cointegrating vector between $x_{1,t}$ and $x_{2,t}$ is $(1, -0.5)'$. Panel (b) of Figure 2 shows that $x_{1,t} - 0.5x_{2,t}$ exhibits stationary time series characteristics, so does y_t depicted in panel (c).

Forecasting evaluation is tabulated in Table 2. As seen from Table 2, the main results are quite similar to what we observe from using the previous DGP discussed in Section 5.3. Compared with the OLS benchmark, LASSO$_{min}$ and LASSO$_{1se}$ always generate smaller FMSE and subsequently positive out-of-sample R^2. The mis-specification induced by using historical average or the AR(1) model can be very costly in this case. The FMSE is at least tripled by using the historical average as the one-step-ahead forecast. Results based on the forecasting combination are disappointing, and this is consistent with our expectation given that the forecasting combination completely ignores the presence of the dependence structure

Table 2: Forecasting performance of different model strategies

T	OLS	Average	AR(1)	F-comb	LASSO_{min}	LASSO_{1se}
FMSE						
100	1.6756	7.1397	2.3330	4.4096	**1.5801**	1.6042
200	1.4215	7.6931	2.2635	5.0832	**1.3802**	1.3838
400	1.4536	8.7265	2.4273	6.0144	**1.4460**	1.4469
1000	1.2486	7.9195	2.1575	5.6196	1.2406	**1.2218**
Out-of-sample R^2						
100		-3.2609	-0.3923	-1.6316	**0.0570**	0.0427
200		-4.4120	-0.5924	-2.5759	**0.0291**	0.0265
400		-5.0032	-0.6689	-3.1376	**0.0052**	0.0046
1000		-5.3428	-0.7280	-3.5008	0.0064	**0.0214**

[a] The out-of-sample R^2 is defined as $R^2 = 1 - FMSE_i/FMSE_{OLS}$ for each model i.

among predictors.

The comparison between LASSO_{min} and LASSO_{1se} provides an interesting contrast to the previous DGP. More often than not, LASSO_{min} produces a slightly smaller FMSE than the LASSO_{1se} model, indicating that the in-sample overfitting issue is less manifest for this DGP with a more complex dependence structure.

One advantage of the LASSO approach is that we allow the dimension of the stationary variables z_t, p, to be arbitrarily large. For instance, p could be even larger than the sample size T, in which case traditional OLS estimation is not feasible. In the previous two DGPs, we set $p = 6$ and $k = 4$. We investigate whether a higher dimension of the stationary predictors has any impact on the forecasting performance of LASSO compared with alternative forecasting strategies. For simplicity, we maintain the dynamics of the DGP as outlined by equations 14, and add a simple AR(1) process with autocorrelation 0.3 to the stationary process z_t. These stationary AR(1) processes enter y_t with coefficient 0.1.

We tabulate the FMSE of different model strategies in Table 3 for three different values of p: 12, 20 and 50. The smallest FMSE in each row is in bold letters. The results observed previously are largely preserved when we increase the dimension of the stationary predictors z_t. The two LASSO estimates always produce the most accurate forecast in terms of FMSE. The percentage improvement of LASSO relative to the OLS benchmark is always larger when the sample size is small. Notably, as p increases, all model specifications considered here generate larger forecast errors. The historical average, AR(1) model and forecast combination are all mis-specified and hence lead to much larger forecast errors.

Table 3: FMSE different model strategies for different values of p

T	OLS	Average	AR(1)	F-comb	LASSO$_{min}$	LASSO$_{1se}$
$p = 12$						
100	1.8520	7.8187	2.5614	5.8080	**1.6774**	1.6808
200	1.5505	7.6690	2.3598	5.9267	1.4978	**1.4718**
400	1.5943	8.0177	2.3667	6.4681	1.5811	**1.5467**
1000	1.5004	8.6363	2.5141	7.0918	1.4918	**1.4645**
$p = 20$						
100	2.0887	7.1947	2.3830	5.8983	1.7752	**1.7200**
200	1.7803	8.0662	2.4802	6.8046	1.6843	**1.5966**
400	1.6883	8.0135	2.3819	6.8941	1.6634	**1.6212**
1000	1.6096	8.6388	2.2921	7.4974	1.6017	**1.5720**
$p = 50$						
100	4.1635	8.2801	3.0859	7.6885	2.7298	**2.5970**
200	2.5218	7.7649	2.7705	7.2337	2.2484	**2.1318**
400	2.0956	8.3824	2.5824	7.8424	2.0349	**1.9049**
1000	2.1310	9.0421	2.7149	8.4954	2.1110	**2.0017**

6 Stock return predictability

Stock return predictability has been a controversial subject in the finance literature. Researchers have not only been constantly seeking new predictors, but also exploiting innovative ways of extracting information from the existing set of predictors (see, for example Buncic and Tischhauser, 2016; Rapach and Zhou, 2013; Neely et al., 2014). Welch and Goyal (2008) conducted one of the most influential studies in this area, reviewing the usefulness and stability of many commonly used predictors. We use the dataset examined in Welch and Goyal (2008) and demonstrate how LASSO estimation could improve the predictability of stock returns via exploring cointegration among the predictors.

6.1 Data

We follow the mainstream literature on forecasting equity premium and use the 14 predictors from Welch and Goyal (2008). Table 4 presents a brief description of the predictors as well as their estimated first-order autocorrelation coefficients over the entire sample period from January 1945 to December 2012 with a monthly frequency. The dependent variable of interest is the equity premium y_t, which is defined as the difference between the compounded return on the S&P 500 index and the Treasury bill rate. Most of the predictors are financial or macroeconomic variables. As shown in Table 4, the majority of these predictors are highly persistent, with 11 out of 14 variables having a first order autocorrelation coefficient higher than 0.95.

The three least persistent variables are stock index variance, inflation and government bond return.

In the benchmark case, we estimate the predictive regression model using a 20-year rolling window to allow for possible parameter instability during the sample period 1945:01-2012:12. Hence the out-of-sample prediction of equity premium starts from 1965:01. We conduct a robustness analysis using 10-year and 30-year rolling windows. Figure 3 plots the estimated AR(1) coefficients for each predictor and equity premium (EQ) using 20-year rolling window. It is evident that not only many predictors exhibit a high persistence level but also the time series dependence structure is changing throughout the forecasting period.

Given the salient feature of the predictors supported by Table and Figure 3, we expect to observe certain cointegrating relationships among the persistent variables. For example, the dividend price ratio (d/p) and dividend yield (d/y) should have a stable long-run relationship in theory. If such a cointegrating relationship exists, it would fit with our theoretical framework nicely. Therefore, we conduct the tests proposed by Johansen (1991, 1995) and Poskitt (2000) to examine the possible cointegration among the persistent variables. We exclude dividend payout ratio (d/e) and long-term yield (lty) when estimating the models in order to avoid multi-collinearity. The cointegration tests are performed on the remaining 9 persistent predictors.

Figure 4 shows the number of cointegrating relationships (i.e. cointegrating rank) throughout the sample period using a 20-year window.[8] The blue solid line denotes the cointegrating rank selected using the Johansen test (Johansen, 1991, 1995), and the red dashed line denotes the cointegrating rank selected using the non-parametric method of Poskitt (2000). Johansen test based on estimating a vector error correction model is one of the most widely used tests for cointegration among multiple time series. However, it has also been commonly acknowledged in the literature that this test suffers substantially from the curse of dimensionality (Gonzalo and Pitarakis, 1999). In particular, the Johansen test tends to overestimate the cointegrating rank, which is reflected in Figure 4. Among 9 persistent predictors, the Johansen test consistently finds at least 3 cointegrating relationships, whereas the more conservative Poskitt (2000) test only finds cointegrating rank 1 or 2. When only d/p and d/y are included in the tests, we always conclude that these two variables are cointegrated.

6.2 Forecasting comparison

We estimate the same set of models as in the simulation study in Section 5. The benchmark model is estimated using OLS. Note that the historical average model uses only observations in the given rolling window instead of the entire past history from the beginning of the sample period. The AR(1) model, forecast combination, LASSO$_{min}$ and LASSO$_{1se}$ models

[8]The testing outcome based on 10-year and 30-year windows is qualitatively similar. These results are available upon request.

Table 4: Variable definitions and estimated first order autocorrelation

Predictor	Definition	Auto.corr(1)
d/p	dividend price ratio: the difference between the log of dividends and the log of prices	0.9940
d/y	dividend yield: the difference between the log of dividends and the log of lagged prices	0.9940
e/p	earning price ratio: the difference between the log of earning and the log of prices	0.9908
d/e	dividend payout ratio: the difference between the log of dividends and the log of earnings	0.9854
b/m	book-to-market ratio: the ratio of book value to market value for the Dow Jones Industrial Average	0.9926
$ntis$	net equity expansion: the ratio of 12-month moving sums of net issues by NYSE listed stocks over the total end-of-year market capitalization of NYSE stocks	0.9757
tbl	Treasury bill rates: the 3-month Treasury bill rates	0.9891
lty	long-term yield: long-term government bond yield	0.9935
tms	term spread: the difference between the long term bond yield and the Treasury bill rate	0.9565
dfy	default yield spread: the difference between Moody's BAA and AAA-rated corporate bond yields	0.9720
dfr	default return spread: the difference between the returns of long-term corporate bonds and long-term government bonds	0.9726
$svar$	stock variance: the sum of squared daily returns on the S&P 500 index	0.4716
$infl$	inflation: inflation of the Consumer Price Index for all urban consumers	0.5268
ltr	long-term return: the rate of returns of long term government bonds	0.0476

Reference: Anderson et al. (2015)

Figure 3: AR(1) coefficients over the 20-year rolling window

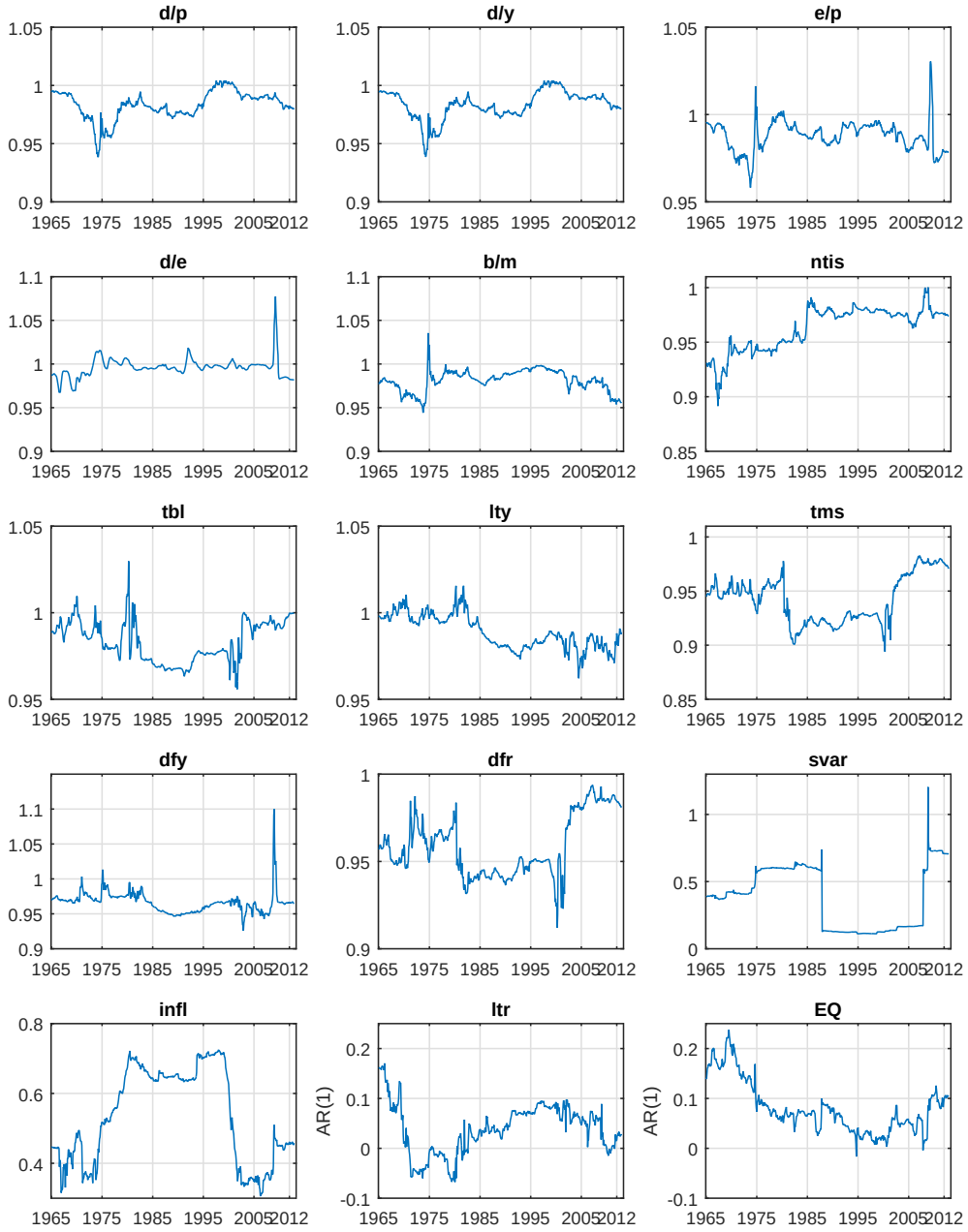
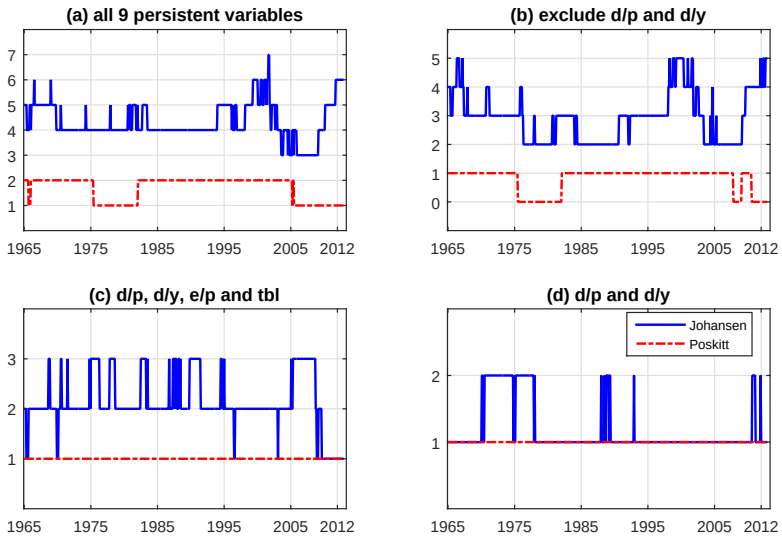


Figure 4: Test for cointegration using a 20-year rolling window



have the same interpretations as before. The FMSE and out-of-sample R^2 are tabulated in Tables 5 and 6, respectively. Following Welch and Goyal (2008), we first consider the generic "kitchen sink" model, where all of the predictors listed in Table 4 are included except d/e and lty. In addition, we estimate a simplified model that contains the three non-persistent predictors ($svar$, $infl$ and ltr) and only two strongly cointegrated persistent predictors d/p and d/y.

Tables 5 and 6 show that in terms of the overall performance, the LASSO$_{min}$ model outperforms all other model specifications. The LASSO$_{min}$ model most often produces the smallest one-step-ahead FMSE and hence the highest out-of-sample R^2 across the 3 different window sizes considered. The improvement in out-of-sample R^2 can be as large as 12% over the kitchen sink OLS model. The LASSO$_{1se}$ model also performs well in the kitchen sink case. However, it often fails to generate better predictions than the OLS benchmark when the persistent predictors other than d/p and d/y are excluded from the model. Forecast combination produces the smallest FMSE using 10-year and 30-year windows in the kitchen sink case but fails to improve forecast accuracy upon the OLS benchmark using 20-year and 30-year windows in the simplified case with two persistent predictors.

We conduct the Giacomini and White (2006) test for conditional predictive ability on the one-step-ahead FMSE against the OLS benchmark, indicating the significance levels in Table 5. At the 10% level, the LASSO$_{min}$ model produces a significantly better one-step-ahead forecast than the corresponding OLS alternative using 10-year and 20-year windows for the kitchen sink specification, also using a 10-year window for the parsimonious two persistent predictors case. The forecast combination and LASSO$_{1se}$ model also perform significantly better than the OLS benchmark 10-year kitchen sink model. The Hansen (2005) test for superior predictive

Table 5: Comparison of FMSE for different model strategies

window	OLS	Average	AR(1)	F-comb	LASSO$_{min}$	LASSO$_{1se}$
Kitchen sink						
10-year	0.002089	0.001884†	0.001900	**0.001803****	0.001831**	0.001859*
20-year	0.002046	0.002050	0.002036	0.001929	**0.001928**†	0.002014
30-year	0.002035	0.002051	0.002042	**0.001956**	0.001968	0.002029
Only d/p and d/y						
10-year	0.001909	0.001884	0.001900	0.001808	**0.001799**†	0.001862
20-year	0.001914	0.002050	0.002036	0.001921	**0.001866**	0.002022
30-year	0.001895	0.002051	0.002042	0.001931	**0.001892**	0.002026

[a] Bold letters indicate the model specification with the smallest forecast mean squared error in each row.
[b] Significance levels of the Giacomini and White (2006) test for conditional predictive ability against the OLS alternative: $\dagger$: 10%, $*$: 5%, $**$: 1%.
[c] Colored blocks represent superior model specifications compared to others using the Hansen (2005) test for superior predictive ability.

Table 6: Comparison of out-of-sample R^2 for different model strategies

window	Average	AR(1)	F-comb	LASSO$_{min}$	LASSO$_{1se}$
Kitchen sink					
10-year	0.098079	0.090310	**0.137113**	0.123380	0.110225
20-year	-0.001966	0.004560	0.056839	**0.057558**	0.015609
30-year	-0.007684	-0.003416	**0.039092**	0.033139	0.002955
Only d/p and d/y					
10-year	0.012927	0.004424	0.052771	**0.057273**	0.024433
20-year	-0.071145	-0.064169	-0.003850	**0.024845**	-0.056772
30-year	-0.082182	-0.077595	-0.019161	**0.001746**	-0.069085

[a] The out-of-sample R^2 is defined as $R^2 = 1 - FMSE_i/FMSE_{OLS}$ for each model i.
[b] Bold letters indicate the model specification that produces the largest out-of-sample R^2 in each row.

ability leads us to similar conclusion. For each forecasting model i, we test the null that

$$H_0 : \ FMSE_i \leq FMSE_j, \qquad \text{for } j \neq i,\ j \in \mathcal{J},$$

where $\mathcal{J}$ contains all models considered here. We highlight in purple blocks the cases, where H_0 cannot be rejected at 5% significance level in Table 5. In all cases we fail to reject that LASSO$_{min}$ model produces a one-step-ahead FMSE which is smaller than or equal to the FMSE generated from all alternative models. In the kitchen sink case, forecast combination is equally good as the LASSO$_{min}$ model in the statistical sense. However, when we keep two persistent predictors in the model, both the OLS and LASSO$_{min}$ models produces more accurate one-step-ahead forecasts than forecast combination using 20-year and 30-year windows. Overall LASSO$_{min}$ model consistently outperforms most of the alternative model specifications considered here.

6.3 Robustness analysis

The tests for cointegration rank shown in Figure 4 suggest that among the 9 persistent predictors, there are usually more than one cointegration relationships. Although our modelling framework assumes a single cointegrating vector, the case of multiple cointegration relationships is easily compatible. Suppose the two normalised cointegrating vectors are δ_1 and δ_2, with the normalisation constants c_1 and c_2. The fact that we are considering one-equation predictive regression means that $\boldsymbol{\beta} = c_1\delta_1 + c_2\delta_2$. In other words, instead of estimating the two cointegrating relationships separately, we can only identify a linear combination of them in $\boldsymbol{\beta}$. The same rule applies to models with higher cointegrating ranks among the persistent predictors.

This is exactly the case in our data set for the stock return prediction. If multiple cointegration relationships exist in a given estimation window, the LASSO estimates of the linear combination of the cointegrating vectors are still consistent. Furthermore, all the asymptotic properties of the LASSO estimates presented in Section 4 still hold. We plot the LASSO estimates of all 12 coefficients using the kitchen sink model with a 20-year rolling window in Figures 5 and 6. An estimated coefficient of 0 indicates that the LASSO$_{min}$ model drops that coefficient from the predictive regression model. Interestingly, the first two panels of Figure 5 show that d/p and d/y do not always have a $(1, -1)$ cointegration relationship due to the existence of other persistent predictors in the model. There is also an evident structural break in early 1996, after which roughly half of the predictors are dropped from the predictive regression model by LASSO. Only the earnings price ratio e/p, stock variance $svar$ and inflation $infl$ appear relevant in most of the windows after early 1996.

Figures 5 and 6 also show that the sign of the estimated coefficients is changing in different parts of the sample. For instance, the earnings price ratio e/p is positively related to the equity

Figure 5: Estimated coefficients using LASSO_{min} with all persistent predictors overlapped with business cycle (20-year window)

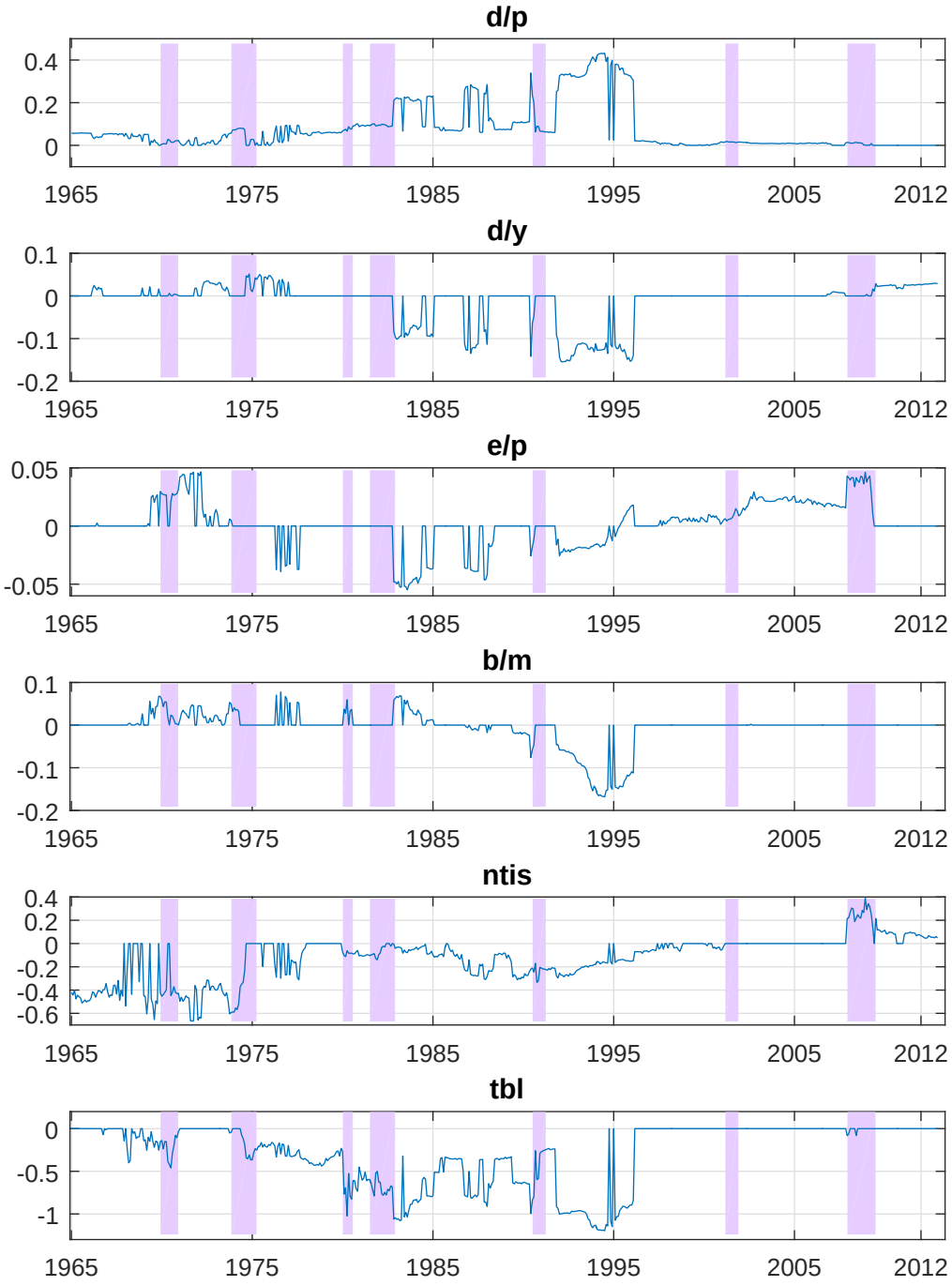


Figure 6: Estimated coefficients using LASSO$_{min}$ with all persistent predictors overlapped with business cycle (20-year window)

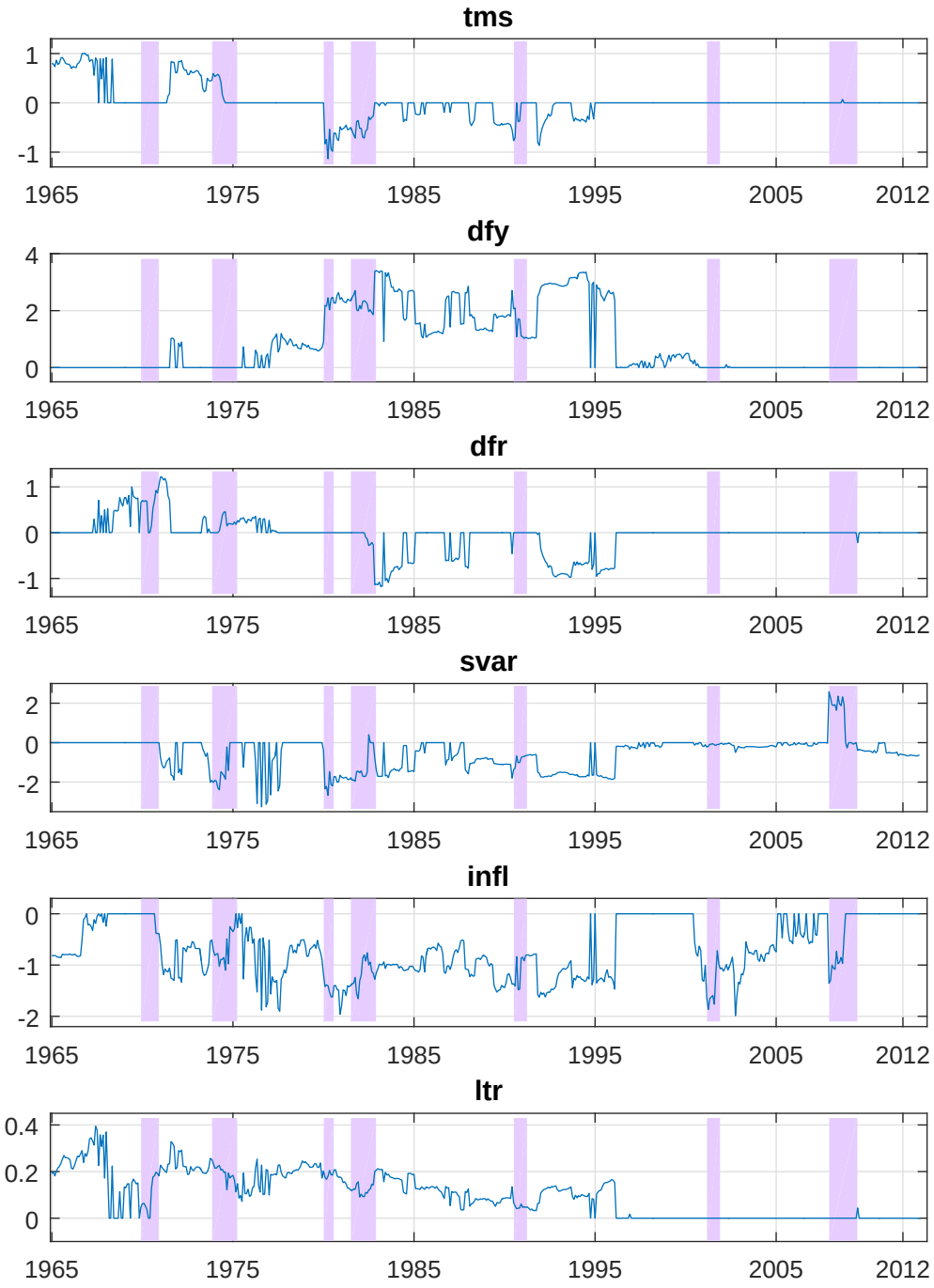


Table 7: Comparison of FMSE for different model strategies using sample 1927:01–2012:12

window	OLS	Average	AR(1)	F-comb	LASSO$_{min}$	LASSO$_{1se}$
Kitchen sink						
10-year	0.002384	0.002146†	0.002196	**0.002056****	0.002116**	0.002127†
20-year	0.001941	0.001892	0.001877	0.001763†	**0.001746****	0.001846
30-year	0.001857	0.001980	0.001956	0.001825	**0.001784**†	0.001939
Only d/p and d/y						
10-year	0.002157	0.002146	0.002196	0.002053	**0.002048**	0.002128
20-year	0.001729	0.001892	0.001877	0.001746	**0.001690**	0.001852
30-year	0.001737	0.001980	0.001956	0.001807	**0.001733**	0.001940

[a] Bold letters indicate the model specification with the smallest forecast mean squared error in each row.
[b] Significance levels of Giacomini and White (2006) test for conditional predictive ability against the OLS alternative: $\dagger$: 10%, $*$: 5%, $**$: 1%.
[c] Colored blocks represent superior model specifications using the Hansen (2005) test for superior predictive ability.

premium prior to 1975 or after 1995, but negatively correlated from 1975 to 1995. This reflects the changing nature of stock return predictability through time. We observe a consistent sign in that a higher Treasury bill rate tbl or inflation $infl$ always leads to a lower equity premium, and a higher dividend price ratio d/p, default yield spread dfy or long-term return ltr leads to higher premium.

We depict the NBER business cycle dates in Figures 5 and 6 to investigate whether the economy being in expansion or recession has any systematic impact on the predictive power of the regressors. The recession dates are shaded in color. There does not appear to be a stable relationship between business cycle and the relevance of each predictor. The results shown above demonstrate the sporadic relationship between equity premium and the predictors commonly used in the literature, which emphasizes the need to use LASSO to "auto-select" the most relevant and important predictors given different estimation windows to achieve the best forecasting performance.

Finally, we examine the robustness of the forecasting results using the entire data set of Welch and Goyal (2008) starting from January 1927. The FMSE of competing models under consideration is tabulated in Table 7. Compared with Table 5, most of the results are qualitatively similar. Overall, the LASSO model with λ chosen as the one that minimizes the cross-validated error (LASSO$_{min}$) performs the best. The Hansen (2005) test for superior predictive ability always ranks LASSO$_{min}$ model as one of the two best model specifications for forecasting equity premium. In general, we can conclude that LASSO has been successful in dealing with the sporadic relationship between the predictors and the prediction objective. Cointegration among persistent predictors is utilized by LASSO in the estimation, which leads to superior forecasting performance.

7 Conclusion

This paper considers the use of LASSO in a predictive regression for stock returns in the presence of cointegration relationships among the persistent predictors. We show that, asymptotically, LASSO yields a consistent estimator of the coefficients, including the cointegrating vector. We also provide the asymptotic distribution of the cointegrating vector, which could be used for statistical inference. Both the simulation studies and an application to the equity premium prediction data show that the proposed LASSO approach leads to superior forecasting performance relative to the commonly used OLS kitchen sink model and the historical average in most cases.

Stock return predictability has long been debated in the finance literature. The LASSO approach we proposed here has a few desirable properties. Firstly, as pointed out by Welch and Goyal (2008), although a large pool of predictors has been suggested by numerous studies, most predictors do not have a stable predictive power over different parts of the sample period. This is not surprising from an econometric point of view, as the stock return displays stationary martingale difference features, while many predictors recommended by the existing literature are highly persistent. We address this issue of unbalanced regression by incorporating cointegration relationships among the persistent predictors. Another challenge many practitioners face is how to select the most relevant predictors from the large set of available variables. This is particularly difficult given the sporadic relationship between the stock return and those predictors. The variable selection function of the LASSO tackles this issue nicely. Finally, our set of assumptions required for the consistency of the LASSO estimator allows the dimension of the stationary predictors to be arbitrarily large. It is still feasible to implement the LASSO when the number of stationary predictors is higher than the sample size, whereas this cannot be done with the traditional OLS. Thus, the "curse of dimensionality" is trivial for the LASSO estimator.

To the best of our knowledge, our paper is the first to study the statistical properties of LASSO estimates of predictive regression in the presence of cointegration. We should emphasize that our methodology does not exclude other ways of improving the predictability of stock returns, for example, bringing in other macroeconomic variables (Buncic and Tischhauser, 2016) and technical indicators (Neely et al., 2014) and imposing constraints on the return forecasts and/or estimated coefficients (Campbell and Thompson, 2008). As documented by Buncic and Tischhauser (2016) and Li and Tsiakas (2016), combining additional information and parameter constraints with LASSO would further improve the predictability of stock returns.

References

Ang, A. and Bekaert, G. (2007), 'Stock return predictability: Is it there?', *Review of Financial studies* **20**(3), 651–707.

Binsbergen, V., Jules, H. and Koijen, R. S. (2010), 'Predictive regressions: A present-value approach', *The Journal of Finance* **65**(4), 1439–1471.

Buncic, D. and Tischhauser, M. (2016), Macroeconomic factors and equity premium predictability, Working paper.

Campbell, J. Y. (1987), 'Stock returns and the term structure', *Journal of financial economics* **18**(2), 373–399.

Campbell, J. Y. and Shiller, R. J. (1988), 'The dividend-price ratio and expectations of future dividends and discount factors', *Review of financial studies* **1**(3), 195–228.

Campbell, J. Y. and Thompson, S. B. (2008), 'Predicting excess stock returns out of sample: Can anything beat the historical average?', *Review of Financial Studies* **21**(4), 1509–1531.

Campbell, J. Y. and Yogo, M. (2006), 'Efficient tests of stock return predictability', *Journal of financial economics* **81**(1), 27–60.

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

Cochrane, J. H. (1991), 'Production-based asset pricing and the link between stock returns and economic fluctuations', *The Journal of Finance* **46**(1), 209–237.

Cremers, K. M. (2002), 'Stock return predictability: A bayesian model selection perspective', *Review of Financial Studies* **15**(4), 1223–1249.

Dow, C. H. (1920), *Scientific stock speculation*, Magazine of Wall Street.

Elliott, G., Gargano, A. and Timmermann, A. (2013), 'Complete subset regressions', *Journal of Econometrics* **177**(2), 357–373.

Fama, E. F. and French, K. R. (1993), 'Common risk factors in the returns on stocks and bonds', *Journal of financial economics* **33**(1), 3–56.

Fama, E. F. and Schwert, G. W. (1977), 'Asset returns and inflation', *Journal of financial economics* **5**(2), 115–146.

Fan, J. and Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**(1), 101.

Giacomini, R. and White, H. (2006), 'Tests of conditional predictive ability', *Econometrica* **74**(6), 1545–1578.

Gonzalo, J. and Pitarakis, J.-Y. (1999), Dimensionality effect in cointegration analysis, *in* R. Engle and H. White, eds, 'Festschrift in Honour of Clive Granger', Oxford University Press, pp. 212–229.

Hansen, B. E. (1992), 'Convergence to stochastic integrals for dependent heterogeneous processes', *Econometric Theory* **8**(04), 489–500.

Hansen, P. R. (2005), 'A test for superior predictive ability', *Journal of Business & Economic Statistics* **23**, 365–380.

Johansen, S. (1991), 'Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models', *Econometrica* **59**, 1551–1580.

Johansen, S. (1995), *Likelihood-based Inference on Cointegrated Vector Auto-regressive Models*, Oxford University Press, Oxford.

Kock, A. B. (2016), 'Consistent and conservative model selection with the adaptive lasso in stationary and nonstationary autoregressions', *Econometric Theory* **32**(01), 243–259.

Koijen, R. S. and Nieuwerburgh, S. V. (2011), 'Predictability of returns and cash flows', *Annual Review of Financial Economics* **3**(1), 467–491.
URL: *https://ideas.repec.org/a/anr/refeco/v3y2011p467-491.html*

Koo, B. and Seo, M. H. (2015), 'Structural-break models under mis-specification: implications for forecasting', *Journal of Econometrics* **188**(1), 166–181.

Lamont, O. (1998), 'Earnings and expected returns', *The journal of Finance* **53**(5), 1563–1587.

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

Li, J. and Tsiakas, I. (2016), A new approach to equity premium prediction: Economic fundamentals matter!, Working paper.

Liao, Z. and Phillips, P. C. (2015), 'Automated estimation of vector error correction models', *Econometric Theory* **31**(03), 581–646.

Ludvigson, S. C. and Ng, S. (2007), 'The empirical risk–return relation: A factor analysis approach', *Journal of Financial Economics* **83**(1), 171–222.

Martin, V., Hurn, S. and Harris, D. (2012), *Econometric Modelling with Time Series*, number 9780521196604 *in* 'Cambridge Books', Cambridge University Press.

Neely, C., Rapach, D., Tu, J. and Zhou, G. (2014), 'Forecasting the equity risk premium: The role of technical indicators', *Management Science* **60**(7), 1772–1791.

Pástor, L. and Stambaugh, R. F. (2009), 'Predictive systems: Living with imperfect predictors', *The Journal of Finance* **64**(4), 1583–1628.

Phillips, P. C. (2014), 'On confidence intervals for autoregressive roots and predictive regression', *Econometrica* **82**(3), 1177–1195.

Phillips, P. C. (2015), 'Halbert white jr. memorial jfec lecture: Pitfalls and possibilities in predictive regression', *Journal of Financial Econometrics* **13**(3), 521–555.

Pontiff, J. and Schall, L. D. (1998), 'Book-to-market ratios as predictors of market returns', *Journal of Financial Economics* **49**(2), 141–160.

Poskitt, D. S. (2000), 'Strongly Consistent Determination of Cointegrating Rank via Canonical Correlations', *Journal of Business and Economic Statistics* **18**, 77–90.

Qian, J., Hastie, T., JeromeFriedman, Tibshirani, R. and Simon, N. (2013), Glmnet for matlab, http://www.stanford.edu/ hastie/glmnetmatlab/.

Rapach, D. E., Ringgenberg, M. C. and Zhou, G. (2016), 'Short interest and aggregate stock returns', *Journal of Financial Economics* **121**(1), 46–65.

Rapach, D. E., Strauss, J. K. and Zhou, G. (2010), 'Out-of-sample equity premium prediction: Combination forecasts and links to the real economy', *Review of Financial Studies* **23**(2), 821–862.

Rapach, D. and Zhou, G. (2013), Forecasting stock returns, Vol. 2, Elsevier, chapter Chapter 6, pp. 328–383.

Rossi, B. (2013), 'Advances in forecasting under instability', *Handbook of Economic Forecasting* (111), 1203–1324.

Stambaugh, R. F. (1999), 'Predictive regressions', *Journal of Financial Economics* **54**(3), 375–421.

Stock, J. H. and Watson, M. W. (1996), 'Evidence on structural instability in macroeconomic time series relations', *Journal of Business & Economic Statistics* **14**(1), 11–30.

Stock, J. H. and Watson, M. W. (2003), 'Forecasting output and inflation: The role of asset prices', *Journal of Economic Literature* **41**(3), 788–829.

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

Tibshirani, R. (2011), 'Regression shrinkage and selection via the lasso: a retrospective', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(3), 273–282.

Timmermann, A. (2006), 'Forecast combinations', *Handbook of economic forecasting* **1**, 135–196.

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

A Proofs

For consistency, we need the following set of assumptions.

Assumptions

A1 The parameter space for γ is compact and $\|\delta\|_1 = 1$ for all δ.

A2 The compatibility condition is satisfied for the set S_0 where S_0 is the support of parameter, $\boldsymbol{\theta}^0$, i.e. $S_0 := \{j : \theta_j^0 \neq 0\}$ for $\boldsymbol{\theta}$. That is, for some positive constant ϕ_0 and all parameters $\boldsymbol{\theta}$ satisfying $\|\boldsymbol{\theta}_{S_0^c}\|_1 \leq 3|\boldsymbol{\theta}_{S_0}\|_1$,

$$\|\boldsymbol{\theta}_{S_0}\|_1^2 \leq \left(\boldsymbol{\theta}'\hat{\Sigma}\boldsymbol{\theta}\right) s_0/\phi^2$$

where $s_0 = |S_0|$ and $\hat{\Sigma} = D'D/n$ is the so-called gram matrix.

A3 Let nonstationary predictors and stationary predictors be denoted by $x_t = \sum_{j=1}^{[nt]} \varepsilon_{nj}$ and z_{tj} for $j = 1, ..., p$, respectively, where $p \to \infty$. Also, let the error process in (1) be $u_t = \sum_{j=1}^{[nt]} v_{nj}$. Then, $\{\varepsilon_{nj}\}$, $\{v_{nj}\}$, and $\{z_{tj}\}_{j=1,...,p}$ be centered α-mixing sequences satisfying the following: for some positive γ_1, $\gamma_2 > 1$, b, and c,

$$\begin{aligned} \alpha(m) &\leq \exp\left(-cm^{\gamma_1}\right) \\ P\left\{\left|z_{tj}\right| > s\right\} &\leq H(s) := \exp\left(1 - (s/b)^{\gamma_2}\right) \quad \text{for all } j \qquad (15) \\ \gamma &= \left(\gamma_1^{-1} + \gamma_2^{-1}\right)^{-1} < 1, \\ \ln p &= O\left(n^{\gamma_1 \wedge (1 - 2\gamma_1(1-\gamma)/\gamma)}\right). \qquad (16) \end{aligned}$$

A4 For some $\varepsilon > 0$, $s_0 p^{1+\varepsilon} \ln p = o(n^{-3/2})$.

Remark 1 *Assumption **A1** is concerned with restrictions on the parameter space. Assumption **A2** is standard for the large sample theory for the LASSO estimation. Assumption **A3** is a technical condition required for Lemmas 1 and 2. Notably, Assumption **A3** implies Assumption 1 in Hansen (1992), which ensures weak convergence of the sum of the product of x_t and u_t in (1) to a stochastic integral.*

A.1 Propositions

Proof of Proposition 1. Since, by definition,

$$\|\mathbf{u} - D(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2^2/n + \lambda\|\hat{\boldsymbol{\gamma}}\|_1 \leq \|\mathbf{u}\|_2^2/n + \lambda\|\boldsymbol{\gamma}^0\|_1,$$

rearranging the above equation yields

$$\begin{aligned}\|D(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2^2/n+\lambda\|\hat{\boldsymbol{\gamma}}\|_1 \leq & 2\mathbf{u}'D(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)/n+\lambda\|\boldsymbol{\gamma}^0\|_1\\ (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)'\frac{D'D}{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)+\lambda\|\hat{\boldsymbol{\gamma}}\|_1 \leq & 2\mathbf{u}'D(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)/n+\lambda\|\boldsymbol{\gamma}^0\|_1. \end{aligned} \tag{17}$$

We start with the first term of the right hand side in (17). Given A1 (i.e. c belongs to a compact parameter space), $\hat{c}=O_p(1)$ and $\mathbf{u}'X/n=n^{-1}\sum_{t=1}^n u_t x_{t-1} \xrightarrow{d} \int_0^1 \mathscr{W}(r)d\mathscr{U}(r)=O_p(1)$. These and Lemma 1 imply that there exists $\tilde{\lambda}$ such that

$$\Pr\left\{2|\mathbf{u}'D|_\infty/n \leq \tilde{\lambda}\right\} \to 1 \text{ w.p.1},$$

$$\mathbf{u}'D/n=\begin{pmatrix} \mathbf{u}'\tilde{X}/n\sqrt{n} & \mathbf{u}'G/n \end{pmatrix}$$

where $\tilde{\lambda}^{-1}=O(\sqrt{n}/\ln p)$ and hence,

$$2|\mathbf{u}'D(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{\mathbf{0}})|_\infty/n \leq \tilde{\lambda}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^{\mathbf{0}}\|_1+O_p(1)\|\hat{\delta}-\delta_0\|_1$$

due to Lemma 1 and Hölder inequality. Therefore, regarding (17), for $\lambda \geq 2\tilde{\lambda}$,

$$2\|D(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2^2/n+2\lambda\|\hat{\boldsymbol{\gamma}}\|_1 \leq \lambda\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\|_1+2\lambda\|\boldsymbol{\gamma}^0\|_1+O_p(1)\|\hat{\delta}-\delta_0\|_1$$

Note that due to the triangle inequality,

$$\|\hat{\boldsymbol{\gamma}}\|_1=\|\hat{\boldsymbol{\gamma}}_{S_0}\|_1+\|\hat{\boldsymbol{\gamma}}_{S_0^c}\|_1 \geq \|\boldsymbol{\gamma}^0_{S_0}\|_1-\|\hat{\boldsymbol{\gamma}}_{S_0}-\boldsymbol{\gamma}^0_{S_0}\|_1+\|\hat{\boldsymbol{\gamma}}_{S_0^c}\|_1$$

where S_0 is the support of parameter variables, $\boldsymbol{\gamma}^0$, i.e. $S_0 := \{j : \gamma_j^0 \neq 0\}$ for $\boldsymbol{\gamma}$ and by construction,

$$\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}^0\|_1=\|\hat{\boldsymbol{\gamma}}_{S_0}-\boldsymbol{\gamma}^0_{S_0}\|_1+\|\hat{\boldsymbol{\gamma}}_{S_0^c}\|_1.$$

Then,

$$2\|D(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2^2/n+\lambda\|\hat{\boldsymbol{\gamma}}_{S_0^c}\|_1 \leq 3\lambda\|\hat{\boldsymbol{\gamma}}_{S_0}-\boldsymbol{\gamma}^0_{S_0}\|_1+O_p(n^{-1/2})\|\sqrt{n}(\hat{\delta}-\delta_0)\|_1 \leq 3\lambda\|\hat{\boldsymbol{\theta}}^0_{S_*}-\boldsymbol{\theta}^0_{S_*}\|_1, \tag{18}$$

where S_* is the index set combining S_0 and indexes for δ.

Given the compatibility condition (A2), there exists some positive constant ϕ_0 such that

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}_{S_0}-\boldsymbol{\theta}^0_{S_0}\|_1^2 &\leq \|D(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2^2 s_0(\lambda,\boldsymbol{\theta})/n\phi_0^2\\ &= (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)'\frac{D'D}{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)s_0(\lambda,\boldsymbol{\theta})/\phi_0^2. \end{aligned} \tag{19}$$

where $s_0(\lambda, \boldsymbol{\theta})$ denotes cardinality of the index set S_* for $\boldsymbol{\theta}$, i.e. $s_0 = |S_*|$. Then,

$$\begin{aligned}
&2\|D(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2^2/n + \lambda\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_1 \\
=&2\|D(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2^2/n + \lambda\|\hat{\boldsymbol{\theta}}_{S_0} - \boldsymbol{\theta}^0_{S_0}\|_1 + \lambda\|\hat{\boldsymbol{\theta}}_{S_0^c}\|_1 \\
\leq&4\lambda\|\hat{\boldsymbol{\theta}}_{S_0} - \boldsymbol{\theta}^0_{S_0}\|_1 \\
\leq&4\lambda\sqrt{s_0(\lambda, \boldsymbol{\theta})}\|D(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2/\sqrt{n}\phi_0 \\
\leq&\|D(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2^2/n + 4\lambda^2 s_0(\lambda, \boldsymbol{\theta})/\phi_0^2
\end{aligned}$$

The first inequality comes from (18), the second inequality is due to (19) and finally, the last inequality uses $4uv \leq u^2 + 4v^2$. Rearranging the inequality,

$$\|D(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2^2/n + \lambda\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\|_1 \leq 4\lambda^2 s_0(\lambda, \boldsymbol{\delta})/\phi_0^2 \tag{20}$$

Next, we refine the convergence rate of $\hat{\delta}$. Recall that

$$\frac{D'D}{n} = \begin{pmatrix} \tilde{X}'\tilde{X}/n^2 & \tilde{X}'G/n\sqrt{n} \\ G'\tilde{X}/n\sqrt{n} & G'G/n \end{pmatrix},$$

and note that

$$\begin{aligned}
\tilde{X}'\tilde{X}/n^2 &= n^{-2}\sum_{t=1}^{n}\tilde{x}_{t-1}\tilde{x}_{t-1}' \xrightarrow{d} (\sigma_\varepsilon c^0)^2\int_0^1 \mathscr{W}(r)\mathscr{W}(r)'dr = O_p(1) \\
G'G/n &= n^{-1}\sum_{t=1}^{n} g_t g_t' \xrightarrow{d} Eg_t g_t' \\
\tilde{X}'G/n\sqrt{n} &= O_p(n^{-1/2+q})
\end{aligned}$$

where $\mathscr{W}(\cdot)$ is the standard Brownian motion, $\tilde{x}_t = x_t\hat{c}$, $g_t = (x_{t-1}'\delta^0, z_{t-1}')'$, and q is any positive number due to Lemma 2. Now, put this observation and the convergence rate of $\hat{\gamma}$, we just obtained, into the first inequality in (18) to deduce that

$$c_1 n\|\hat{\delta} - \delta_0\|_1 \leq O_p(1)\|\hat{\delta} - \delta_0\|_1$$

for some positive c_1 by choosing q small, which yields the desired rate for $\hat{\delta}$ ■

Proof of Proposition 2. Let

$$S_n(\gamma, \delta) = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - g_t'\gamma - x_t'\delta\right)^2.$$

It is clear from Proposition 1 that

$$\sup_{\delta} |S_n(\hat{\gamma}, \delta) - S_n(\gamma_0, \delta)| = o_p(n^{-1}),$$

where the supremum is taken over any n^{-1} neighborhood of δ_0. Therefore, we consider the weak convergence of the stochastic process, i.e.,

$$\begin{aligned} D_n(b) &= n\left(S_n(\delta_0 + n^{-1}b) - S_n(\delta_0)\right), \\ &= -\frac{2c_0}{n}\sum_{t=1}^{n} u_t x_t' b + \frac{c_0^2}{n} b' \sum_{t=1}^{n} x_t x_t' b \\ &\Rightarrow -2c_0 b'\left[\int_0^1 \mathscr{W}(r) d\mathscr{U}(r) + \Lambda\right] + c_0^2 b' \int_0^1 \mathscr{W}(r)\mathscr{W}(r)' dr b \\ &:= D(b) \end{aligned}$$

over any given compact space for b, where $S_n(\delta) = S_n(\gamma_0, \delta)$.

Furthermore, the estimator $\hat{\delta}$ is the minimizer of $S_n(\hat{\gamma}, \delta)$ under the constraint that $\|\delta\|_1 = 1$. The consistency of $\hat{\delta}$ means the sign-consistency as well. Thus,

$$\begin{aligned} 0 &= \|\hat{\delta}\|_1 - \|\delta_0\|_1 \\ &= \sum_j (n^{-1}\text{sgn}(\delta_{0j}))n(\hat{\delta}_j - \delta_{0j}) \\ &= \text{sgn}(\delta_0)' b \end{aligned}$$

with probability approaching 1. Therefore, we can conclude that

$$n(\hat{\delta} - \delta_0) \to^d \underset{b:\text{sgn}(\delta_0)'b=0}{\arg\min} D(b),$$

by the argmax continuous mapping theorem. ■

A.2 Lemmata

Lemma 1 *Let* $\{z_{tj}\}$, $j = 1, ..., p$, *be centered* α*-mixing sequences satisfying the following: for some positive* γ_1, $\gamma_2 > 1$, b, *and* c,

$$\begin{aligned} \alpha(m) &\leq \exp(-cm^{\gamma_1}) \\ P\{|z_{tj}| > s\} &\leq H(s) := \exp\left(1 - (s/b)^{\gamma_2}\right) \quad \textit{for all } j \qquad (21) \\ \gamma &= \left(\gamma_1^{-1} + \gamma_2^{-1}\right)^{-1} < 1, \\ \ln p &= O\left(n^{\gamma_1 \wedge (1 - 2\gamma_1(1-\gamma)/\gamma)}\right). \qquad (22) \end{aligned}$$

Then,

$$\max_{j\leq p}\left|\sum_{t=1}^{n} z_{tj}\right| = O_p\left(\sqrt{n}\ln p\right).$$

Proof of Lemma 1. Consider the sums of truncated variables first. Consider $z_{tj}^M = z_{tj}1\left\{\left|z_{tj}\right| \leq M\right\} + M1\left\{\left|z_{tj}\right| > M\right\}$ with $M = \left(1+n^{\gamma_1}\right)^{1/\gamma_2}$. Due to Proposition 2 in Merlevede et al. (2011), we have

$$E\left(\exp\left(n^{-1/2}\sum_{t=1}^{n}\left(z_{tj}^M - Ez_{tj}^M\right)\right)\right) = O\left(1\right),$$

uniformly in j and thus

$$E\max_{j\leq p}\left|n^{-1/2}\sum_{t=1}^{n}\left(z_{tj}^M - Ez_{tj}^M\right)\right| = O\left(\ln p\right). \tag{23}$$

For the remainder term, note that

$$\begin{aligned} E\max_{j\leq p}\sum_{t=1}^{n}\left|z_{tj} - z_{tj}^M + Ez_{tj}^M\right| &\leq npE\left|z_{tj} - z_{tj}^M + Ez_{tj}^M\right| \\ &\leq 2np\int_M^\infty H\left(x\right)dx \\ &= O\left(np\exp\left(-n^{\gamma_1}\right)\right), \end{aligned}$$

where the last equality is obtained by bounding the integral by $MH(M) = \exp\left(-n^{\gamma_1}\right)$ in general since $H\left(x\right)$ decays at an exponential rate.

Thus,

$$\begin{aligned} E\max_{j\leq p}\left|\sum_{t=1}^{n} z_{tj}\right| &\leq E\max_{j\leq p}\left|\sum_{t=1}^{n} z_{tj}^M - Ez_{tj}^M\right| + E\max_{j\leq p}\left|\sum_{t=1}^{n}\left[z_{tj} - \left(z_{tj}^M - Ez_{tj}^M\right)\right]\right| \\ &= O\left(\sqrt{n}\ln p\right) + O\left(np\exp\left(-n^{\gamma_1}\right)\right). \end{aligned}$$

Finally, recall that $\ln p = o\left(n^{\gamma_1}\right)$ by (22). ■

Next, we derive another maximal inequality involving integrated processes by means of martingale approximation. (e.g. Hansen 1992)

Lemma 2 *Let $\{x_t\}$ be an integrated process such that Δx_t is centered and $x_0 = O_p\left(1\right)$. Let $\{\Delta x_t\}$ and $\left\{z_{tj}\right\}$, $j = 1, ..., p$, be α-mixing sequences satisfying the conditions in Lemma 1. Also let $p = O\left(n^r\right)$ for some $r < \infty$. Then,*

$$\max_{1\leq j\leq p}\left|\sum_{t=1}^{n} z_{tj}x_t\right| = O_p\left(\left(n\vee p\right)^{1+\delta}\right),$$

for any $\delta > 0$.

Proof of Lemma 2.

We apply the martingale difference approximation to z_{tj} following Hansen (1992). Let $\mathcal{F}_t$ denote the generated sigma field from all $z_{t-s,j}$ and x_{t+1-s}, $s, j = 1, 2, ...$ and $E_t(\cdot)$ the conditional expectation based on $\mathcal{F}_t$. Define

$$\epsilon_{tj} = \sum_{s=0}^{\infty} \left(E_t z_{t+s,j} - E_{t-1} z_{t+s,j} \right), \quad \zeta_{tj} = \sum_{s=1}^{\infty} E_t z_{t+s,j}.$$

Then, it is clear that

$$z_{tj} = \epsilon_{tj} + \zeta_{t-1,j} - \zeta_{t,j}$$

and ϵ_{tj} is a martingale difference sequence.

We begin with deriving a moment bound for the sum of a martingale difference array $\left\{ \varepsilon_t = \epsilon_{tj} x_t / \sqrt{n} \right\}$. Consider a truncated sequence

$$\varepsilon_t^C = \varepsilon_t 1\{|\varepsilon_t| < C\} + C1\{\varepsilon_t \geq C\} - C1\{\varepsilon_t \leq -C\}$$

and its centered version

$$\tilde{\varepsilon}_t^C = \varepsilon_t^C - E\left(\varepsilon_t^C | \mathcal{F}_{t-1} \right).$$

Then, by construction $\tilde{\varepsilon}_t^C$ is a martingale difference sequence. Furthermore,

$$\left| \tilde{\varepsilon}_t^C \right| \leq \left| \varepsilon_t^C \right| + E\left(\left| \varepsilon_t^C \right| | \mathcal{F}_{t-1} \right) \leq 2C$$

by the triangle and Jensen's inequalities and by the fact that $\left| \varepsilon_t^C \right| \leq C$. Furthermore, let

$$r_t^C = \varepsilon_t - \tilde{\varepsilon}_t^C = \left(\varepsilon_t - \varepsilon_t^C \right) - E\left(\left(\varepsilon_t - \varepsilon_t^C \right) | \mathcal{F}_{t-1} \right),$$

where the last equality following from the fact that $\{\varepsilon_t\}$ is an MDS and thus

$$\varepsilon_t = \tilde{\varepsilon}_t^C + r_t^C.$$

By Azuma's (1967) inequality,

$$\Pr\left\{ \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{\varepsilon}_t^C \right| \geq c \right\} \leq 2 \exp\left(-\frac{c^2}{4C^2} \right).$$

On the other hand, since $r_t^C = 0$ if $|\varepsilon_t| \leq C$, we have for any $c > 0$,

$$\Pr\left\{ \left| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} r_t^C \right| > c \right\} \leq \Pr\left\{ \max_t |\varepsilon_t| > C \right\}.$$

However,

$$\begin{aligned}\Pr\left\{\max_t |\varepsilon_t| > C\right\} &\leq \Pr\left\{\max_t |\epsilon_{tj}| \max_t \left|\frac{x_t}{\sqrt{n}}\right| > C\right\} \\ &\leq \sum_t \Pr\left\{|\epsilon_{tj}| > \sqrt{C}\right\} + \Pr\left\{\max_t \left|\frac{x_n}{\sqrt{n}}\right| > \sqrt{C}\right\} \\ &\leq C^{-q} \sum_{t=1}^{n} E\left|\epsilon_{tj}\right|^{2q} + 2n \exp\left(-\frac{(nC)^{\gamma/2}}{C_1}\right) + \exp\left(-\frac{C}{C_2}\right),\end{aligned}$$

where the first term comes from Markov inequality and the second and third terms are given by Merlevede et al.'s (2011) equation (1.11).

Next, we derive some moments bounds for the sums of ζ_{tj} and ϵ_{tj}. By (A.1) of Hansen (1992), if z_{tj} is α-mixing of size $-q\beta/(q-\beta)$ with $E\left|z_{tj}\right|^q < \infty$,

$$\left(E\left|\zeta_{tj}\right|^\beta\right)^{1/\beta} \leq 6 \sum_{k=1}^{\infty} \alpha_k^{1/\beta - 1/q} \left(E\left|z_{tj}\right|^q\right)^{1/q}.$$

And, by the triangle inequality,

$$\left(E\left|\epsilon_{tj}\right|^q\right)^{1/q} = \left(E\left|z_{tj} + \zeta_{tj} - \zeta_{t-1,j}\right|^q\right)^{1/q} \leq \left(E\left|z_{tj}\right|^q\right)^{1/q} + 2\left(E\left|\zeta_{tj}\right|^q\right)^{1/q}.$$

Under our moment condition and mixing decay rate in Lemma 1, these quantities are bounded for any β and q such that $q > \beta$.

Then, for any a_n,

$$\begin{aligned}&\Pr\left\{\left|\frac{1}{a_n\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t\right| \geq c\right\} \\ &\leq \Pr\left\{\left|\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{\varepsilon}_t^C\right| \geq a_n \frac{c}{2}\right\} + \Pr\left\{\max_t |\varepsilon_t| > C\right\} \\ &\leq 2\exp\left(-\frac{a_n^2 c^2}{4C^2}\right) + C^{-q} \sum_{t=1}^{n} E\left|\epsilon_{tj}\right|^{2q} + 2n \exp\left(-\frac{(nC)^{\gamma/2}}{C_1}\right) + \exp\left(-\frac{C}{C_2}\right)\end{aligned}$$

Given this bound, we can deduce from the union bound that

$$\begin{aligned}&\Pr\left\{\max_{1\leq j\leq p} \left|\frac{1}{a_n\sqrt{n}} \sum_{t=1}^{n} \varepsilon_{jt}\right| > c_1\right\} \\ &\leq p \Pr\left\{\left|\frac{1}{a_n\sqrt{n}} \sum_{t=1}^{n} \varepsilon_{jt}\right| > c_1\right\} \\ &\leq p\left(2\exp\left(-\frac{a_n^2 c_1^2}{4C^2}\right) + C^{-q} \sum_{t=1}^{n} E\left|\epsilon_{tj}\right|^{2q} + 2n \exp\left(-\frac{(nC)^{\gamma/2}}{C_1}\right) + \exp\left(-\frac{C}{C_2}\right)\right).\end{aligned}$$

Since $E\left|\epsilon_{tj}\right|^{2q} < \infty$, we begin with choosing C such that $pnC^{-q} \to 0$. This choice of C makes

the last two terms degenerate and for any $c_1 > 0$, the first term also degenerates as long as $a_n \geq b_n (pn)^{\frac{1}{2q}} \ln 2p$ for some $b_n \to \infty$. Since q is arbitrary and p is at most a polynomial of n, we can conclude that

$$\max_{j \leq p} \left| \frac{1}{n} \sum_{t=1}^{n} \epsilon_{tj} x_t \right| = O_p \left((p \vee n)^{\frac{1}{q}} \right),$$

for any q.

Next, note that

$$\begin{aligned}
&\max_j \left| \sum_{t=1}^{n} \left(\zeta_{t-1,j} - \zeta_{t,j} \right) x_t \right| \\
&\leq \max_j \left| \zeta_{0,j} \right| |x_1| + \max_j \left| \zeta_{n,j} \right| |x_n| + \max_j \left| \sum_{t=2}^{n} \left(\zeta_{t-1,j} \Delta x_t - E\zeta_{t-1,j} \Delta x_t \right) \right| + n \max_j E\zeta_{t-1,j} \Delta x_t \\
&= O_p \left(\sqrt{n} \log p \right) + O_p \left(\sqrt{n} \log p \right) + O(n),
\end{aligned}$$

by the observation that $\sup_t |x_t| / \sqrt{n} = O_p(1)$, $\max_j \left| \zeta_{0,j} \right| = O_p \left(\sqrt{\log p} \right)$, and by Lemma 1 with $\{|xy| > c\} \subset \{|x| > \sqrt{c}\} \cup \{|y| > \sqrt{c}\}$. ■