Journal of Financial Economics 145 (2022) 154–177

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

Market efficiency in the age of big data☆

Ian W.R. Martin[a], Stefan Nagel[b,*]

[a] London School of Economics, Houghton Street, London WC2A 2AE, UK
[b] University of Chicago, Booth School of Business, 5807 S Woodlawn Ave, Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Received 11 July 2020
Revised 13 May 2021
Accepted 6 June 2021
Available online 27 November 2021

JEL classification:
G14
G12
C11

Keywords:
Bayesian learning
High-dimensional prediction problems
Return predictability
Out-of-sample tests

ABSTRACT

Modern investors face a high-dimensional prediction problem: thousands of observable variables are potentially relevant for forecasting. We reassess the conventional wisdom on market efficiency in light of this fact. In our equilibrium model, N assets have cash flows that are linear in J characteristics, with unknown coefficients. Risk-neutral Bayesian investors learn these coefficients and determine market prices. If J and N are comparable in size, returns are cross-sectionally predictable ex post. In-sample tests of market efficiency reject the no-predictability null with high probability, even though investors use information optimally in real time. In contrast, out-of-sample tests retain their economic meaning.

1. Introduction

Machine learning methods have proved useful in forecasting problems with huge numbers of predictor variables. High-dimensional prediction problems of this kind are faced not only by data scientists studying data as outside observers, but also by economic decision-makers in the marketplace. Many forward-looking economic decisions require predictions for which large numbers of variables could potentially be relevant, but the exact relationship between predictors and forecast target is unknown and must be learned from observed data. In this paper, we argue that to understand market outcomes in such settings, it is important to take into account the high-dimensional nature of decision-makers' prediction problem.

We demonstrate this in an asset-pricing setting. We show that properties of asset prices are strongly affected by the dimensionality of investors' prediction problem. Conventional notions of how to test market efficiency and how to interpret pricing anomalies break down in the high-dimensional case.

To price risky assets such as stocks, investors must forecast the future cash flows generated by these assets. In our model, cash-flow growth rates of a cross-section of N firms are a linear function of J firm characteristics that are fixed over time. Investors are Bayesian, homogeneous, and risk neutral; they price stocks based on the predictive dis-

☆ G. William Schwert was the editor for this article. We are grateful for the comments of Stéphane Bonhomme, Svetlana Bryzgalova, John Campbell, Kent Daniel, Gene Fama, Cam Harvey, Ralph Koijen, Sendhil Mullainathan, Lubos Pastor, Andrew Patton, Paul Schneider, Andrei Shleifer, Allan Timmerman, Laura Veldkamp, seminar participants at Berkeley, City University of Hong Kong, Federal Reserve Bank of Atlanta, Harvard, Rochester, and the University of Chicago, and conference participants at the ABFR webinar, Joint Statistical Meetings, and WFA meetings. We thank Tianshu Lyu for excellent research assistance. Martin thanks the ERC for support under Starting Grant 639744. Nagel gratefully acknowledges financial support from the Center for Research in Security Prices at the University of Chicago Booth School of Business.

* Corresponding author.
E-mail addresses: i.w.martin@lse.ac.uk (I.W.R. Martin), stefan.nagel@chicagobooth.edu (S. Nagel).

https://doi.org/10.1016/j.jfineco.2021.10.006

tribution of cash flows. Realized asset returns in this setting are simply equal to investors' forecast errors. If investors knew the coefficients of the predictive model, they could form expectations of cash-flow growth, and hence price assets, in such a way that returns would not be predictable in the cross-section. This is the conventional rational expectations (RE) equilibrium that is the foundation for typical market efficiency tests. Similarly, if J is small relative to N, investors could estimate the parameters of their cash-flow forecasting model with great precision, leading to asset prices that are close to those in the RE equilibrium.

In reality, however, investors face a myriad of potential predictor variables that could be relevant in constructing such forecasts. In other words, J is not small relative to N. As technology has improved, the set of available and potentially valuation-relevant predictor variables has expanded enormously over time. Textual analysis, satellite imagery, social media data, and many other new data sources yield a wealth of information. But in order to use these sources of information in a forecasting model, investors must estimate the relation between these signals and future cash flows. This is a high-dimensional learning problem. The number of potential predictor variables could easily surpass the number of assets whose cash flow data is available to estimate this relation.

Machine learning methods handle this issue by imposing some regularization on the estimation, for example by shrinking parameter estimates towards a fixed target or by searching for a sparse model representation that includes only a small subset of variables from a much larger set of potential predictors. With the goal of optimizing out-of-sample forecasting performance, regularization lets the learner trade off the costs of downweighting certain pieces of information against the benefit of reduced parameter-estimation error. In a Bayesian interpretation, shrinkage reflects informative prior beliefs: when forecasters know, based on economic plausibility considerations, that forecasting model parameters cannot have arbitrarily large magnitudes, their posterior beliefs are shrunk towards the prior mean.

Shrinkage ameliorates, but does not eliminate, the effects of parameter uncertainty on asset prices in the high-dimensional case. Relative to the RE equilibrium, asset prices are distorted by two components. First, noise in the past cash-flow growth observations that investors learn from will have, by chance, some correlation with the J predictor variables. This induces error in investors' parameter estimates, and hence an asset price distortion, that is correlated with the J predictor variables. Shrinkage downweights this estimation error component, but it also gives rise to a second distortion component because shrinkage implies underweighting the predictive information in the J predictors. Naturally, this second component, too, is correlated with the J predictor variables.

To stack the deck against return predictability, we endow investors with prior beliefs that are objectively correct in the sense that the coefficients of the cash flow–generating model are drawn from this prior distribution. Investors also know that this model is linear. With this objective prior, the optimal amount of shrinkage exactly balances the two components in such a way that investors' forecast errors, and hence also asset returns, are unpredictable out-of-sample.

The fact that returns are not predictable out-of-sample, however, does not imply that there is no in-sample predictability. An econometrician conducting an in-sample predictability test uses data that had not been available to investors in real time when they priced assets. In an RE setting, this would not matter, because investors would already have perfect knowledge of model parameters. Approximately, the same would be true in a low-dimensional setting with small J and large N, where investors would be able to estimate forecasting-model parameters with high precision. But in a high-dimensional setting, the econometrician's ability to see data realized ex post, after investors' pricing decisions are made, gives her a substantial advantage.

To show this, we consider an econometrician who collects asset price data from our model economy ex post and runs in-sample regressions to test whether the J firm characteristics cross-sectionally predict returns. When J is vanishing in size relative to N, there is almost no predictability: with $N \to \infty$ and J fixed, the predictability test would reject the null with test size close to the chosen significance level (e.g., 0.05). In contrast, in high-dimensional asymptotics, where $N, J \to \infty$ jointly, with their ratio J/N converging to a fixed number, the econometrician would reject the no-predictability null hypothesis, in the limit, with probability one. In simulations, we show that these high-dimensional asymptotic results are a good approximation for the case of finite N with J comparable in size to N. We obtain rejection probabilities close to one in these simulations, too. Importantly, the high rejection rates in-sample tests are not caused by distortions in the sampling properties of the econometrician's test statistics. Instead, the high rejection rates correctly reflect the fact that equilibrium asset prices contain in-sample predictable components that are large in a high-dimensional setting.

The situation is different for out-of-sample tests. In our model economy, a portfolio formed based on the econometrician's predictive regression estimates up to period t, with positive weights for stocks with positive predicted returns and negative weights for stocks with negative expected returns, has an average return of zero in the subsequent period $t+1$. In other words, returns are not predictable out-of-sample. This is true, too, in the high-dimensional asymptotic case. Intuitively, since Bayesian investors optimally use information available to them and price assets such that returns are not predictable under their predictive distribution, an econometrician who is restricted to constructing return forecasts using only data that had been available to investors in real time is not able to predict returns out-of-sample either.

These results illustrate forcefully that the economic content of the (semi-strong) market efficiency notion that prices "fully reflect" all public information (Fama, 1970) is not clear in this high-dimensional setting, even though we abstract from the joint hypothesis problem by assuming

that investors are risk-neutral.[1] Does "fully reflect" mean that investors know the parameters of the cash flow–prediction model? Or does "fully reflect" mean that investors employ Bayesian updating when they learn from data about these parameters? The null hypothesis in a vast empirical literature in asset pricing, including return predictability regressions, event studies, and asset-pricing model estimation based on orthogonality conditions, is the former version of the market efficiency hypothesis. Our results show that testing and rejecting this version has little economic content in a high-dimensional setting. An apparent rejection of market efficiency might simply represent the unsurprising consequence of investors not having precise knowledge of the parameters of a data-generating process that involves thousands of predictor variables.

Empirical discoveries of new cross-sectional return predictors that are statistically significant according to conventional in-sample tests are therefore less interesting in a high-dimensional world. From the perspective of our model, it is not surprising that the technology-driven explosion in the number of predictor variables available to investors has coincided with an explosion in the number of return predictors that are found significant in asset-pricing studies (Cochrane, 2011; Harvey et al., 2016). Even without *p*-hacking, multiple testing, or data mining (Lo and MacKinlay, 1990; Harvey et al., 2016; Chordia et al., 2019), evidence of cross-sectional return predictability from in-sample regressions does not tell us much about the expected returns that investors perceived ex ante at the time they priced assets. Thus, out-of-sample tests (such as those in McLean and Pontiff, 2016) gain additional importance in the age of Big Data.

Researchers are often skeptical of out-of-sample tests. In the case where a fixed underlying process is generating returns (as would be the case in many RE models), in- and out-of-sample methods test the same hypothesis, and in-sample tests are more powerful because they use the available data to the fullest extent. As a consequence, it is not clear why one would want to focus on out-of-sample tests (Inoue and Kilian, 2005; Campbell and Thompson, 2008; Cochrane, 2008; Hansen and Timmermann, 2015). In contrast, if investors face a high-dimensional learning problem, substantial in-sample predictability can coexist with absence of out-of-sample predictability. In-sample and out-of-sample tests examine fundamentally different hypotheses in this case. This provides a clear motivation for out-of-sample testing.

We show that out-of-sample portfolio returns can isolate predictable components of returns that reflect risk premia or behavioral biases. Unlike in-sample estimates, the out-of-sample estimates are not distorted by investors' learning-induced forecast errors. We also show that if the econometrician applies shrinkage similar to ridge regression in the in-sample return prediction regression, with the penalty hyperparameter estimated via cross-validation, then the portfolio return based on these shrinkage estimates can be equivalent to an out-of-sample portfolio return. Within our Bayesian learning setting, this result therefore provides an economic interpretation of approaches that use cross-validation (Kozak et al., 2020) or closely related methods (Chinco et al., 2021) to estimate prior beliefs for cross-sectional return prediction.

We illustrate the different perspectives provided by in- and out-of-sample tests with an empirical example. In the cross-section of U.S. stocks, we consider each stock's history of monthly simple and squared returns over the previous 120 months as a set of return predictors. Running a ridge regression over a full five decade sample, the in-sample coefficient estimates pick up the most prominent past return-based anomalies in the literature, including momentum (Jegadeesh and Titman, 1993; Novy-Marx, 2012), long-term reversals (DeBondt and Thaler, 1985), and momentum seasonality (Heston and Sadka, 2008). In other words, there is substantial in-sample predictability. In terms of out-of-sample predictability, the picture looks very different. Using rolling regressions over 20-year windows to estimate prediction model coefficients and then using those to predict returns in subsequent periods, we find that predictability is generally much weaker out-of-sample than in-sample. Moreover, there is substantial decay over time. While some out-of-sample predictability exists in the early decades of the sample, it is basically nil in recent years. This suggests that there may be little reason to seek risk-based or behavioral explanations of the cross-sectional predictability that shows up in the in-sample analysis.

One potential explanation for the out-of-sample predictability in the earlier parts of the sample may be that investors several decades ago were not able to process the information in each stock's price history as effectively as investors today. One can think of this as bounded rationality that induces excessive shrinkage or sparsity of investors' forecasting models, along the lines of Sims (2003), Gabaix (2014), and Molavi et al. (2020). We show in our simulations that sparsity or shrinkage beyond the level called for by objectively correct Bayesian priors leads to positive out-of-sample return predictability.

Overall, our results suggest that in-sample cross-sectional return predictability tests are ill-suited for uncovering return premia that require explanations based on priced risk exposures or behavioral biases. This is not to say that all of the documented patterns in the literature are explainable with learning and will not persist out-of-sample. But it is important to obtain other supporting evidence beyond in-sample predictability tests. If predictability associated with a predictor variable persists out-of-sample, if there is a compelling theoretical motivation, or if other types of data point to a risk or behavioral bias explanation (e.g., economic risk exposures, data on investor expectations), the case for a risk premium or a persistent behavioral bias is much stronger.

The insight in our analysis that learning can induce in-sample return predictability relates to an earlier literature that studies learning-induced return predictability in low-dimensional time-series settings with few return predictors (e.g., Timmermann, 1993; Lewellen and Shanken, 2002; Collin-Dufresne et al., 2016). These earlier time-

[1] The joint hypothesis problem (Fama, 1970) refers to the problem that the econometrician studying asset prices does not know the model that determines risk premia required by risk-averse investors.

series analyses do not address the question of how learning affects asset prices and the properties of in- and out-of-sample return predictability tests in a high-dimensional cross-sectional setting. This is the focus of our paper.

Our approach has antecedents elsewhere in the literature. Aragones et al. (2005) and Al-Najjar (2009) treat decision-makers as statisticians who have to learn from observed data in a non-Bayesian high-dimensional setting. Their focus is on conditions under which disagreement between agents can persist in the long run. Klein (2019) and Calvano et al. (2018) focus on strategic interaction of machine learning pricing algorithms in product markets. Investors in our setting face a simpler learning problem within a Bayesian linear framework and without strategic interactions. Even in this simple setting, important pricing implications emerge.

All proofs are in the appendix.

2. Bayesian pricing in a high-dimensional setting

Consider an economy in discrete time, $t \in \{1, 2, \ldots\}$, with N assets. Each asset is associated with a vector of J firm characteristics observable to investors that we collect in the $N \times J$ matrix $\boldsymbol{X}$. The assets pay dividends, collected in the vector $\boldsymbol{y}_t$, whose growth $\Delta \boldsymbol{y}_t = \boldsymbol{y}_t - \boldsymbol{y}_{t-1}$ is partly predictable based on $\boldsymbol{X}$:

Assumption 1.

$$\Delta \boldsymbol{y}_t = \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_e), \\ \operatorname{rank}(\boldsymbol{X}) = J, \quad \frac{1}{NJ} \operatorname{tr} \boldsymbol{X}'\boldsymbol{X} = 1. \tag{1}$$

The set of characteristics is potentially very large, but for simplicity we assume $J < N$. It would be relatively straightforward to extend the framework to allow for $J \geq N$, but the main points can be seen more clearly in the simpler $J < N$ setting. The set of characteristics in $\boldsymbol{X}$ exhausts the set of variables that investors can condition on. Due to technological change, this set could change as previously unavailable predictors become available, so we will be concerned with the behavior of prices for various values of J.

The assumption that $\frac{1}{NJ} \operatorname{tr} \boldsymbol{X}'\boldsymbol{X} = 1$ is a normalization that defines a natural scale for the characteristics. For example, it holds if characteristics are scaled to have unit norm (i.e., if $\frac{1}{N}\sum_{n=1}^{N} x_{nj}^2 = 1$ for every characteristic j), as then

$$\frac{1}{NJ} \operatorname{tr} \boldsymbol{X}'\boldsymbol{X} = \frac{1}{J}\sum_{j=1}^{J} \frac{1}{N}\sum_{n=1}^{N} x_{nj}^2 = 1.$$

We assume that the characteristics associated with a firm are constant over time for simplicity. In reality, firms' characteristics change. But as long as investors know the firm's characteristics at every point in time, one can accommodate this in our setting by thinking of $\boldsymbol{y}_t$ as a vector of payoffs for hypothetical characteristics-constant firms. We would have to reshuffle firms each period so that each element of $\boldsymbol{y}_t$ is always associated with the same characteristics.

We further make the following assumption:

Assumption 2. Investors are risk-neutral and the interest rate is zero.

By abstracting from risk premia, we intentionally make it easy for an econometrician to test market efficiency in our setting. With risk-neutral investors, there is no joint hypothesis problem due to unknown risk–pricing models. Yet, as we will show, interpretation of standard market efficiency tests is still tricky.

We focus on the pricing of one-period dividend strips so that $\boldsymbol{p}_t$ represents the vector of prices, at time t, of claims to dividends paid at time $t+1$. We think of one period in this model as a long time span, say a decade, so that the errors in $\boldsymbol{e}_t$ are actually the averages of the errors one would find by sampling at higher frequencies over many shorter subperiods. With this interpretation in mind, we can then think of the dividend strip payoff as a long-lived stock's cash flows compressed into a single cash flow at the typical duration of a stock (e.g., perhaps a decade).

The price vector is then equal to the vector of next-period expected dividends,

$$\boldsymbol{p}_t = \widetilde{\mathbb{E}}_t \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \widetilde{\mathbb{E}}_t \Delta \boldsymbol{y}_{t+1} = \boldsymbol{y}_t + \widetilde{\mathbb{E}}_t (\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1}).$$

This formulation encompasses a range of possible assumptions about the process by which investors' expectations $\widetilde{\mathbb{E}}_t[\cdot]$ are formed.

In a rational expectations model, for example, investors know $\boldsymbol{g}$, so $\widetilde{\mathbb{E}}_t(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1}) = \boldsymbol{X}\boldsymbol{g}$. The dividend strip price is therefore $\boldsymbol{p}_t = \boldsymbol{y}_t + \boldsymbol{X}\boldsymbol{g}$, and realized price changes $\boldsymbol{r}_{t+1} = \boldsymbol{y}_{t+1} - \boldsymbol{p}_t = \Delta \boldsymbol{y}_{t+1} - \boldsymbol{X}\boldsymbol{g} = \boldsymbol{e}_{t+1}$ are unpredictable with $\boldsymbol{X}$. This is the usual null hypothesis that underlies tests based on orthogonality conditions and Euler equations.

However, we focus on the realistic case where investors do not know $\boldsymbol{g}$. They therefore face a learning problem in pricing assets. They can learn about $\boldsymbol{g}$ by observing the realizations of $\{\Delta \boldsymbol{y}_s\}_1^t$ and the characteristics $\boldsymbol{X}$. (We assume that investors know $\boldsymbol{\Sigma}_e$.) We then have $\widetilde{\mathbb{E}}_t(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1}) = \boldsymbol{X}\tilde{\boldsymbol{g}}_t$, where $\tilde{\boldsymbol{g}}_t$ represents investors' posterior mean of $\boldsymbol{g}$ at time t, after learning from historical data.

If J is close to (or perhaps even larger than) N, running an ordinary least squares (OLS) regression to estimate $\boldsymbol{g}$ would not give investors useful forecasts. For example, with $J = N$, a cross-sectional regression of $\Delta \boldsymbol{y}_t$ on $\boldsymbol{X}$ exactly fits $\Delta \boldsymbol{y}_t$ in sample. Then $\widetilde{\mathbb{E}}_t(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1}) = \Delta \boldsymbol{y}_t$ so that $\boldsymbol{p}_t = \boldsymbol{y}_t + \Delta \boldsymbol{y}_t$ and $\boldsymbol{r}_{t+1} = \Delta \boldsymbol{y}_{t+1} - \Delta \boldsymbol{y}_t$. The forecast mean squared error (MSE) is then $\operatorname{var}(\boldsymbol{e}_{t+1} - \boldsymbol{e}_t)$, i.e., twice the variance of the truly unpredictable $\boldsymbol{e}_{t+1}$.

For comparison, the naive "random walk" forecast that sets $\mathbb{E}_t \Delta \boldsymbol{y}_{t+1} = \boldsymbol{0}$ would result in $\boldsymbol{p}_t = \boldsymbol{y}_t$ and hence $\boldsymbol{r}_{t+1} = \Delta \boldsymbol{y}_{t+1}$. In this case, the forecast MSE is $\operatorname{var}(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1})$. If a relatively small component of cash-flow growth is predictable, that is, if $\operatorname{var}(\boldsymbol{X}\boldsymbol{g}) \ll \operatorname{var} \boldsymbol{e}_{t+1}$, then the random walk forecast MSE may be substantially lower than the OLS forecast MSE.

2.1. Priors and posteriors

The problem with least-squares regression forecasts is that the prior implicit in the least-squares estimator is economically unreasonable. The posterior mean equals the generalized least squares (GLS) estimator if investors' prior

for $\boldsymbol{g}$ is diffuse. But a diffuse prior for $\boldsymbol{g}$ is not a plausible assumption. Economic reasoning should lead investors to realize that the amount of predictable variation in $\Delta \boldsymbol{y}_{t+1}$ must be limited. It does not make economic sense for investors to believe that arbitrarily large values for $\boldsymbol{g}$ are just as likely as values that give rise to moderate predictable variation in $\Delta \boldsymbol{y}_{t+1}$. While they might not have very precise prior knowledge of $\boldsymbol{g}$, it reasonable to assume that the distribution representing investors' prior beliefs about $\boldsymbol{g}$ is concentrated around moderate values of $\boldsymbol{g}$.

We therefore make the following specification of prior beliefs.

Assumption 3. Before seeing data, investors hold prior beliefs

$$\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_g)\,.$$

That prior beliefs are centered around zero means that investors a priori do not know which characteristics predict cash-flow growth and by how much. But they know that magnitudes of $\boldsymbol{g}$ elements cannot be too big. Economic restrictions on $\boldsymbol{\Sigma}_g$ that restrict the likely magnitudes of $\boldsymbol{g}$ elements will play an important role later on in our analysis.

Proposition 1. After investors have observed dividend growth in a single period, $\Delta \boldsymbol{y}_1$, their posterior distribution of $\boldsymbol{g}$ is $\boldsymbol{g}|\Delta \boldsymbol{y}_1, \boldsymbol{X} \sim N(\tilde{\boldsymbol{g}}_1, \boldsymbol{D}_1)$, where

$$\begin{aligned}\tilde{\boldsymbol{g}}_1 &= \boldsymbol{D}_1 \boldsymbol{d}_1,\\ \boldsymbol{D}_1^{-1} &= \boldsymbol{\Sigma}_g^{-1} + \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X},\\ \boldsymbol{d}_1 &= \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\Delta \boldsymbol{y}_1.\end{aligned}$$

After observing data for t periods, the posterior mean is $\tilde{\boldsymbol{g}}_t = \boldsymbol{D}_t \boldsymbol{d}_t$ and

$$\begin{aligned}\boldsymbol{D}_t^{-1} &= \boldsymbol{\Sigma}_g^{-1} + t\boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X},\\ \boldsymbol{d}_t &= t\boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\overline{\Delta \boldsymbol{y}}_t,\end{aligned}$$

where $\overline{\Delta \boldsymbol{y}}_t = \frac{1}{t}\sum_{s=1}^{t}\Delta \boldsymbol{y}_s$. Therefore

$$\tilde{\boldsymbol{g}}_t = \left[\frac{1}{t}\boldsymbol{\Sigma}_g^{-1} + \boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\overline{\Delta \boldsymbol{y}}_t. \tag{2}$$

The posterior mean $\tilde{\boldsymbol{g}}_1$ takes the form of a Tikhonov-regularized (i.e., ridge) regression estimator, where the inverse of $\boldsymbol{D}_1$ is "stabilized" by adding $\boldsymbol{\Sigma}_g^{-1}$ to $\boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X}$. Thus our Bayesian framework connects to a large literature in machine learning in which Tikhonov regularization is used to deal with high-dimensional prediction problems. For example, see Shalev-Shwartz and Ben-David (2014).[2]

We first simplify the setup by making

Assumption 4.

$$\begin{aligned}\boldsymbol{\Sigma}_e &= \boldsymbol{I},\\ \boldsymbol{\Sigma}_g &= \frac{\theta}{J}\boldsymbol{I}, \quad \theta > 0.\end{aligned}$$

Our results go through for a general (nonsingular) covariance matrix $\boldsymbol{\Sigma}_e$, i.e., with a factor structure in residuals, though at the cost of some extra notational complexity. By assuming $\boldsymbol{\Sigma}_e = \boldsymbol{I}$, we are making the learning problem easy for investors. With uncorrelated residuals, investors achieve a given posterior precision with a smaller J than if residuals were correlated.

By assuming that $\boldsymbol{\Sigma}_g$ is proportional to the identity, we put all the predictor variables on an equal footing from the prior perspective; and it is essential that the variance of the elements of $\boldsymbol{g}$ should decline with J in order to consider sensible asymptotic limits. To see this, note that the covariance matrix of $\boldsymbol{X}\boldsymbol{g}$ is $\boldsymbol{X}\boldsymbol{\Sigma}_g\boldsymbol{X}'$, so the cross-sectional average prior variance of the predictable component of cash-flow growth satisfies

$$\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{X}\boldsymbol{\Sigma}_g\boldsymbol{X}'\right)_{ii} \overset{(A4)}{=} \frac{\theta}{JN}\sum_{i=1}^{N}\sum_{j=1}^{J}x_{ij}^2 \overset{(A1)}{=} \theta, \tag{3}$$

using Assumptions 1 and 4.

To understand (2), it will be convenient to think in terms of the principal components of the data matrix $\boldsymbol{X}$. Specifically, we form the eigendecomposition

$$\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'. \tag{4}$$

Here $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues $\lambda_j > 0$ of $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$ along its diagonal, and $\boldsymbol{Q}$ is an orthogonal matrix whose columns are the corresponding eigenvectors of $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$. These columns are the principal components of $\boldsymbol{X}$. (It is possible to eigendecompose $\boldsymbol{X}$ in this way because $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$ is symmetric and positive definite.) Lastly, the normalization $\mathrm{tr}\,\boldsymbol{X}'\boldsymbol{X} = NJ$ (Assumption 1) implies that the average eigenvalue equals one:

$$\frac{1}{J}\sum_{i=1}^{J}\lambda_i = \frac{1}{J}\,\mathrm{tr}\,\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = 1, \tag{5}$$

using the fact that the sum of the eigenvalues of a matrix equals its trace.

Combining Assumption 4 with Eq. (2), we obtain

$$\tilde{\boldsymbol{g}}_t = \left[\frac{J}{\theta t}\boldsymbol{I} + \boldsymbol{X}'\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\overline{\Delta \boldsymbol{y}}_t = \boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\overline{\Delta \boldsymbol{y}}_t, \tag{6}$$

where

$$\boldsymbol{\Gamma}_t = \boldsymbol{Q}\left(\boldsymbol{I} + \frac{J}{N\theta t}\boldsymbol{\Lambda}^{-1}\right)^{-1}\boldsymbol{Q}' \tag{7}$$

is a symmetric matrix. The posterior mean $\tilde{\boldsymbol{g}}_t$ shrinks the naive OLS estimate (from a regression of $\overline{\Delta \boldsymbol{y}}_t$ onto the columns of $\boldsymbol{X}$) along the principal components. Shrinkage is a consequence of the informative prior for $\boldsymbol{g}$. The prior's influence on the posterior is stronger if the observed data is less informative relative to the prior. To see explicitly what the degree of shrinkage depends on, note

[2] To interpret the posterior mean in terms of standard regression estimators, we can use the Woodbury identity to write

$$\begin{aligned}\tilde{\boldsymbol{g}}_1 &= \boldsymbol{\Sigma}_g\left\{\boldsymbol{\Sigma}_g + \left(\boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X}\right)^{-1}\right\}^{-1}\tilde{\boldsymbol{g}}_{GLS,1}\\ \tilde{\boldsymbol{g}}_{GLS,1} &= \left(\boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_e^{-1}\Delta \boldsymbol{y}_1.\end{aligned}$$

This shows that the posterior mean is a weighted average of the prior mean (zero) and the GLS regression estimator $\tilde{\boldsymbol{g}}_{GLS,1}$.

that $\left(\boldsymbol{I}+\frac{J}{N\theta t}\boldsymbol{\Lambda}^{-1}\right)^{-1}$ is diagonal with elements

$$\frac{\lambda_j}{\lambda_j+\frac{J}{N\theta t}}$$

along its diagonal. Thus, shrinkage is strong if t or θ are small, or J/N is large, or along principal components with small eigenvalues.

We are now in a position to characterize the behavior of realized returns in equilibrium.

Proposition 2. With assets priced based on $\tilde{\boldsymbol{g}}_t$, realized returns are

$$\boldsymbol{r}_{t+1}=\boldsymbol{y}_{t+1}-\boldsymbol{p}_t=\boldsymbol{X}(\boldsymbol{I}-\boldsymbol{\Gamma}_t)\boldsymbol{g}-\boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t+\boldsymbol{e}_{t+1}, \tag{8}$$

where $\bar{\boldsymbol{e}}_t=\frac{1}{t}\sum_{s=1}^{t}\boldsymbol{e}_s$. Hence expected returns satisfy $\mathbb{E}\boldsymbol{r}_{t+1}=0$, and the covariance matrix satisfies

$$\mathbb{E}\boldsymbol{r}_{t+1}\boldsymbol{r}_{t+1}'=\frac{\theta}{J}\boldsymbol{X}(\boldsymbol{I}-\boldsymbol{\Gamma}_t)\boldsymbol{X}'+\boldsymbol{I}.$$

Realized returns thus have three components. The first term on the right-hand side of (8) reflects the effect of "underreaction," due to shrinkage, to the fundamental information in $\boldsymbol{X}$. If investors had an uninformative prior ($\theta\to\infty$ and hence $\boldsymbol{\Gamma}_t\to\boldsymbol{I}$) as in many conventional low-dimensional Bayesian learning models, this term would not be present. But as we have argued, such an uninformative prior would imply that investors entertain an unreasonable amount of predictable variation in dividend growth. Under investors' informed prior beliefs, this part is still zero in expectation because the prior mean of $\boldsymbol{g}$ is zero; but for a given draw of $\boldsymbol{g}$ that generates the data an econometrician would study, it is not zero.

The second term represents the effect of noise on investors' posterior mean. To the extent that the unpredictable shocks in $\bar{\boldsymbol{e}}_t$ in a given sample line up, by chance, with columns of $\boldsymbol{X}$, this induces estimation error that tilts investors' cash flow–growth forecast away from $\boldsymbol{X}\boldsymbol{g}$. Shrinkage via $\boldsymbol{\Gamma}_t$ reduces this component, at the cost of generating the first term. Under Bayesian learning, $\boldsymbol{\Gamma}_t$ optimally trades off the pricing error arising from these two components.

The third term is the unpredictable shock $\boldsymbol{e}_{t+1}$. In the rational expectations case where $\boldsymbol{g}$ is known to investors, the realized return would simply be equal to $\boldsymbol{e}_{t+1}$ and the first two terms would not exist.

In the Bayesian learning case, however, the first two terms are not zero, and, as a consequence, returns contain highly persistent components correlated with the columns of $\boldsymbol{X}$. Even though $\mathbb{E}\boldsymbol{r}_{t+1}=0$ when we integrate over the distribution of $\boldsymbol{g}$ under investors' prior beliefs in Proposition 2, the expected returns are not zero conditional on a given draw of $\boldsymbol{g}$ and $\bar{\boldsymbol{e}}_t$. Unconditionally, returns are therefore more volatile, in the sense that $\mathbb{E}\boldsymbol{r}_{t+1}\boldsymbol{r}_{t+1}'-\boldsymbol{I}$ is positive definite (so that every asset and every portfolio of assets have higher volatility than they would in the rational expectations case in which, by Assumption 4, the covariance matrix equals $\boldsymbol{I}$). As we now show, the presence of these components may induce certain forms of return predictability.

3. Asymptotic analysis

We now analyze the properties of asset prices in high-dimensional asymptotic analysis when $N, J\to\infty$, where $J/N\to\psi>0$, where ψ is a fixed number.[3] This differs from the usual low-dimensional large N, fixed J (or large T, fixed N, and fixed J) asymptotics that underlie most econometric methods in asset pricing. The high-dimensional asymptotics are intended to provide a tractable approximation for the case where J and N are finite and J is not small relative to N (just as conventional large N–fixed J asymptotics provide an approximation for the small J/N case).

3.1. In-sample predictability

We consider an econometrician who studies these realized returns with the usual tools of frequentist statistics. The econometrician looks for return predictability by regressing $\boldsymbol{r}_{t+1}$ on the variables in $\boldsymbol{X}$ using OLS.[4] By allowing the econometrician to see all the predictor variables used by investors, we make things as easy as possible for the econometrician. In the Online Appendix, we modify our main results to allow the econometrician to see only a subset of the investors' predictors, and show that this modification does not affect our main results in any substantive way.

As the number of predictor variables increases, $J\to\infty$, we are potentially in the realm of Big Data. But the mere fact that empiricists have access to a lot of data is not enough, as some of the data could be (asymptotically) redundant. Our next assumption can therefore be thought of as formalizing the notion of a Big Data environment.

Assumption 5 (Big Data). The eigenvalues λ_j of $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$ satisfy $\lambda_j>\varepsilon$ for all j, where $\varepsilon>0$ is a uniform constant as $N\to\infty$.

To understand this assumption, note that if $\boldsymbol{X}'\boldsymbol{X}$ has eigenvalues that are very close to zero, then the columns of $\boldsymbol{X}$ are roughly collinear. To find a linear combination $\boldsymbol{v}\in\mathbb{R}^J$ of columns of $\boldsymbol{X}$ with the property that $\boldsymbol{X}\boldsymbol{v}$ is small (where $\boldsymbol{v}$ is a unit vector, $\boldsymbol{v}'\boldsymbol{v}=1$), we can choose $\boldsymbol{v}$ to be a unit eigenvector of $\boldsymbol{X}'\boldsymbol{X}$ with minimal eigenvalue $\lambda_{\min}$ so that $(\boldsymbol{X}\boldsymbol{v})'(\boldsymbol{X}\boldsymbol{v})=\lambda_{\min}\approx 0$. Thus if there are eigenvalues close to zero, then some characteristics are approximately spanned by other characteristics.

[3] See, for example, Anatolyev (2012) and Dobriban and Wager (2018) for recent examples from the econometrics and statistics literature on high-dimensional regression that use this type of asymptotic analysis. This literature focuses on the asymptotic properties of estimators and statistical tests given an underlying data-generating model that stays fixed as N and J change. In contrast, in our case the nature of investors' learning problem changes as N and J change, and hence the properties of the data that the econometrician analyzes change as well.

[4] Given our assumption that $\boldsymbol{\Sigma}_e=\boldsymbol{I}$, OLS and GLS coincide. The econometrician could also use shrinkage methods like ridge regression, effectively imposing a prior that the coefficients in the return predictability regression cannot be too big. To the extent that the implied prior distribution of the coefficients is roughly in line with the true distribution of the coefficients, using such methods would strengthen the in-sample return predictability. We do not show formal results for this case, but we have explored the issue in simulations.

We emphasize, however, that even if a small number of principal components capture much of the variation in the data, it does not necessarily follow that the Big Data assumption is violated. In Section 3.2, we will exhibit a natural benchmark case (namely, the case in which the characteristics in $\boldsymbol{X}$ are drawn in an IID random way) in which the assumption holds even though there are indeed a few principal components that capture much of the variation (and many more that contribute relatively little variation). It is ultimately an empirical question whether we live in a Big Data world in the sense of Assumption 5. As we will now show, the qualitative properties of standard econometric tests are starkly different depending on whether or not the assumption holds.

The econometrician regresses $\boldsymbol{r}_{t+1}$ on $\boldsymbol{X}$, obtaining a vector of cross-sectional regression coefficients

$$\boldsymbol{h}_{t+1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{r}_{t+1}. \tag{9}$$

In our baseline asymptotic analysis, we focus on the case in which the econometrician uses all characteristics jointly as return predictors. The online appendix shows that we obtain similar results in the case where the econometrician only observes a subset of the firm characteristics known to investors. Our simulations in Section 4 below also consider the case where many econometricians analyze single-predictor regressions.

Following the logic of rational expectations econometrics, which assumes that investors price assets under knowledge of $\boldsymbol{g}$, the econometrician entertains $\boldsymbol{r}_{t+1} = \boldsymbol{e}_{t+1}$ as the no-predictability null hypothesis. Given that the elements of $\boldsymbol{e}_{t+1}$ are distributed $N(0,1)$, it would follow, under this null, that[5]

$$\sqrt{N}\boldsymbol{h}_{t+1} \sim N\left(0, N(\boldsymbol{X}'\boldsymbol{X})^{-1}\right). \tag{10}$$

Hence, under the econometrician's rational expectations null hypothesis, we would have

$$\boldsymbol{h}_{t+1}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{h}_{t+1} \sim \chi^2_J. \tag{11}$$

As we want to characterize the properties of the econometrician's test under asymptotics where $N, J \to \infty$ and $J/N \to \psi > 0$, it is more convenient to let the econometrician consider a scaled version of this test statistic:

$$T_{re} \equiv \frac{\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} - J}{\sqrt{2J}}. \tag{12}$$

Under the econometrician's rational expectations null, we would have, asymptotically,

$$T_{re} \xrightarrow{d} N(0,1) \quad \text{as } N, J \to \infty,\ J/N \to \psi > 0. \tag{13}$$

But the actual asymptotic distribution of T_{re} is influenced by the components of returns involving $\tilde{\boldsymbol{e}}_t$ and $\boldsymbol{g}$ in (8). These components alter the asymptotic distribution and may lead the rejection probabilities of a test using $N(0,1)$ critical values based on (13), or χ^2 critical values based on (11), to differ from the nominal size of the test.

Our first result characterizes the properties of this test statistic under the true model, according to which returns follow Eq. (8). In this analysis, we assume that $\boldsymbol{g}$ is drawn from the prior distribution. This assumption is conservative in the sense that investors' prior beliefs about the distribution of $\boldsymbol{g}$ are objectively correct. If investors' priors were to deviate from this distribution, this would be another source of return predictability.

To assess the performance of the rational expectations econometrician's test statistic in our setting, it is helpful to write

$$\boldsymbol{\Sigma}_{re} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \quad \text{and} \quad \boldsymbol{\Sigma}_b = \mathbb{E}\left(\boldsymbol{h}_{t+1}\boldsymbol{h}_{t+1}'\right)$$

for the covariance matrices of the predictive coefficient estimates under the (incorrect) rational expectations null hypothesis and the true model, respectively. When returns are generated under the true model (8), the rational expectations econometrician will use inappropriately small standard errors, in the sense that $\boldsymbol{\Sigma}_b\boldsymbol{\Sigma}_{re}^{-1} - \boldsymbol{I}$ is positive definite.[6]

The first two moments of the eigenvalues of $\boldsymbol{\Sigma}_b\boldsymbol{\Sigma}_{re}^{-1}$ control the asymptotic behavior of T_{re}. These eigenvalues $(\zeta_{i,t})$ can be written explicitly in terms of the eigenvalues (λ_j) of $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$ as

$$\zeta_{i,t} = 1 + \frac{1}{t + \frac{J}{N\theta\lambda_i}}. \tag{14}$$

We write the limiting mean and variance of the eigenvalues as

$$\mu = \lim_{J\to\infty}\frac{1}{J}\sum_{i=1}^{J}\zeta_{i,t} \quad \text{and} \quad \sigma^2 = \lim_{J\to\infty}\frac{1}{J}\sum_{i=1}^{J}\zeta_{i,t}^2 - \mu^2.$$

By the "Big Data" Assumption 5, we have $1 < \mu < 2$ and $1 < \sqrt{\mu^2+\sigma^2} < 2$ for all $t \geq 1$. (Without the assumption, we would have $\mu = 1$ and $\sigma = 0$ if $\lambda_i \to 0$ and hence $\zeta_{i,t} \to 1$.)

Proposition 3. If returns are generated according to (8), then in the large N, J limit

$$\frac{\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} - \sum_{i=1}^{J}\zeta_{i,t}}{\sqrt{2\sum_{i=1}^{J}\zeta_{i,t}^2}} \xrightarrow{d} N(0,1).$$

It follows that the test statistic T_{re} satisfies

$$\frac{T_{re}}{\sqrt{\mu^2+\sigma^2}} - \frac{\mu-1}{\sqrt{2\left(\mu^2+\sigma^2\right)}}\sqrt{J} \xrightarrow{d} N(0,1)$$

where $1 < \mu < 2$ and $1 < \sqrt{\mu^2+\sigma^2} < 2$.

We can therefore think of T_{re} as a multiple of a standard Normal random variable plus a term of order $\sqrt{J}$:

$$T_{re} \approx \sqrt{\mu^2+\sigma^2}\,N(0,1) + \frac{\mu-1}{\sqrt{2}}\sqrt{J}. \tag{15}$$

[5] Recall that we assume $\boldsymbol{\Sigma}_e = \boldsymbol{I}$ (Assumption 4). More generally, if the econometrician has to estimate $\boldsymbol{\Sigma}_e$, this can be done based on the regression residual $\boldsymbol{\xi}_{t+1} = \boldsymbol{r}_{t+1} - \boldsymbol{X}\boldsymbol{h}_{t+1} = \left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{e}_{t+1}$, used to estimate the variance as $(\boldsymbol{X}'\boldsymbol{X})^{-1}\frac{1}{N-J}\boldsymbol{\xi}_{t+1}'\boldsymbol{\xi}_{t+1}$, which would estimate the variance consistently under conventional large-N, fixed-J. However, if the econometrician does not impose the null hypothesis $\boldsymbol{h}_{t+1} = 0$ in this variance estimation, the estimated variance is an additional source of test size distortions; see Anatolyev (2012) for more details.

[6] As the proof of Proposition 3 shows, $\boldsymbol{\Sigma}_b\boldsymbol{\Sigma}_{re}^{-1} - \boldsymbol{I} = \frac{N\theta}{J}\boldsymbol{Q}\left(\boldsymbol{I} + \frac{N\theta t}{J}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}\boldsymbol{Q}'$, where $\boldsymbol{Q}$ is orthogonal and $\boldsymbol{\Lambda}$ is diagonal with positive entries. It is therefore a symmetric matrix with positive eigenvalues, and so is positive definite.

(For comparison, the rational expectations econometrician thinks T_{re} is asymptotically standard Normal, as in Eq. (13).) Thus the rejection probability tends rapidly to one as N and J tend to infinity.

These results apply as J tends to infinity with t held finite. If we also allow t to tend to infinity, then agents learn the model and μ shrinks toward 1; in this case, Eq. (15) shows that the key question is whether μ tends to 1 faster than J tends to infinity. Under the Big Data Assumption 5, we have $\mu - 1 > \frac{1}{t+\frac{J}{N\theta\varepsilon}}$, so the final term in (15) will create problems for the econometrician if $\sqrt{J}$ grows faster than t. Since one period in our analysis should correspond to the time length of the return sample that an econometrician might use in an asset-pricing test, the length of one period should be at least a decade. Therefore, t will typically be much smaller than $\sqrt{J}$ in cross-sectional asset-pricing settings. An asymptotic analysis in which $\sqrt{J}$ grows faster than t then likely provides a better approximation for typical finite-sample properties than one in which t grows faster. For this reason, we continue to focus on the fixed-t case.

Proposition 4. In a test of return predictability based on the rational expectations null (13), we would have, for any critical value c_α and at any time t,

$$\mathbb{P}(T_{re} > c_\alpha) \to 1 \text{ as } N, J \to \infty, J/N \to \psi.$$

More precisely, for any fixed $t > 0$, the probability that the test fails to reject declines exponentially fast as N and J increase, at a rate that is determined by μ, σ, and ψ:

$$\lim_{N\to\infty} -\frac{1}{N}\log \mathbb{P}(T_{re} < c_\alpha) = \frac{(\mu-1)^2\psi}{4\left(\mu^2+\sigma^2\right)}, \tag{16}$$

for any critical value c_α.

Thus, when J is not small relative to N, in-sample predictability tests lose their economic meaning, in the sense that the usual interpretation of in-sample return predictability evidence is not warranted. The typical conclusion from rejections of the no-predictability null in studies of the cross-section of stock returns is that models of risk premia or mispricing due to imperfectly rational investors are needed to explain the evidence. Our model points to a third possibility: investors are Bayesian, but in-sample return predictability arises even for large t because investors' forecasting problem is high-dimensional.

Instead of testing for predictability by studying the size of coefficients in predictive regressions via $\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1}$, as above, we might imagine trying to construct a trading strategy with weights proportional to in-sample predicted returns,

$$\boldsymbol{w}_{IS,t} = \frac{1}{N}\boldsymbol{X}\boldsymbol{h}_{t+1}.$$

As the predictive coefficients satisfy $\boldsymbol{h}_{t+1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{r}_{t+1}$, the return on the strategy is

$$\begin{aligned}\boldsymbol{r}_{t+1}'\boldsymbol{w}_{IS,t} &= \frac{1}{N}\boldsymbol{r}_{t+1}'\boldsymbol{X}\boldsymbol{h}_{t+1}\\ &= \frac{1}{N}\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1}.\end{aligned}$$

Thus the two approaches are equivalent.

3.2. A benchmark example: $\boldsymbol{X}$ a random matrix

With some additional assumptions on the matrix of firm characteristics $\boldsymbol{X}$, we can explicitly calculate the limit moments that appear in our propositions. We can then see, for example, how these limit moments depend on ψ, the limiting ratio of the number of predictors, J, to the number of observations, N.

Suppose that characteristics are determined at random, so that $\boldsymbol{X}$ has IID entries x_{ij} with mean zero, unit variance, and finite fourth moment.[7] Nature generates this matrix once before investors start learning, and it stays fixed thereafter. As before, investors know $\boldsymbol{X}$, and it stays fixed when we imagine an econometrician repeatedly sampling data by rerunning the economy.

We can use results from random matrix theory to characterize the distribution of the eigenvalues λ_i. In particular, the eigenvalue distribution converges to the Marchenko-Pastur distribution as $N, J \to \infty$ with $J/N = \psi$. For ψ close to one, this distribution features substantial probability mass on eigenvalues close to zero, indicating that many of the columns of $\boldsymbol{X}$ are close to being collinear. Nonetheless, the results of Yin et al. (1988) and Bai and Yin (1993) ensure that all the eigenvalues lie in a bounded interval that does not contain the origin: $\lambda_j \in \left[\left(1-\sqrt{\psi}\right)^2, \left(1+\sqrt{\psi}\right)^2\right]$ for all j. Thus Assumption 5 is satisfied.

Figure 1 shows histograms of the eigenvalue distributions in examples with $\boldsymbol{X}$ drawn randomly with $N(0,1)$ entries, setting $N = 1000$ and $J = 10$, 100, 500, and 900. Solid red lines in the figures illustrate the limiting Marchenko-Pastur distribution for the eigenvalues λ_j in each case; we also calculate the corresponding asymptotic distribution of the eigenvalues $\zeta_{j,t}$ by change of variable, using Eq. (A.6) in the appendix. When $\psi = J/N$ is close to one, there is considerable mass near zero, implying that there are many approximately collinear relations between the columns of $\boldsymbol{X}$. This is a realistic property that one would also find in actual empirical data if one assembled a huge matrix of firm characteristics.

Our next result characterizes the limiting cross-sectional mean, μ, and variance, σ^2, of the distribution of the eigenvalues $\zeta_{j,t}$.

Proposition 5. The cross-sectional moments of $\zeta_{j,t}$ satisfy

$$\mu = 1 + \frac{\psi + \theta t(\psi+1) - \sqrt{[\psi+\theta t(\psi+1)]^2 - 4\theta^2 t^2\psi}}{2\theta t^2\psi} \tag{17}$$

and

$$\begin{aligned}\sigma^2 = {} & \frac{\theta^2t^2\psi - (\theta t+\psi)^2}{2\theta^2t^4\psi^2}\\ & + \frac{\theta t\psi\left(\theta^2t^2(\psi-2) - \theta t\psi + \psi^2\right) + (\theta t+\psi)^3}{2\theta^2t^4\psi^2\sqrt{[\psi+\theta t(\psi+1)]^2 - 4\theta^2t^2\psi}}.\end{aligned} \tag{18}$$

[7] Then Assumption 1 holds asymptotically, as $\frac{1}{NJ}\operatorname{tr}\boldsymbol{X}'\boldsymbol{X} = \frac{1}{NJ}\sum_{n,j}x_{n,j}^2 \to 1$ as $N, J \to \infty$ by the strong law of large numbers.

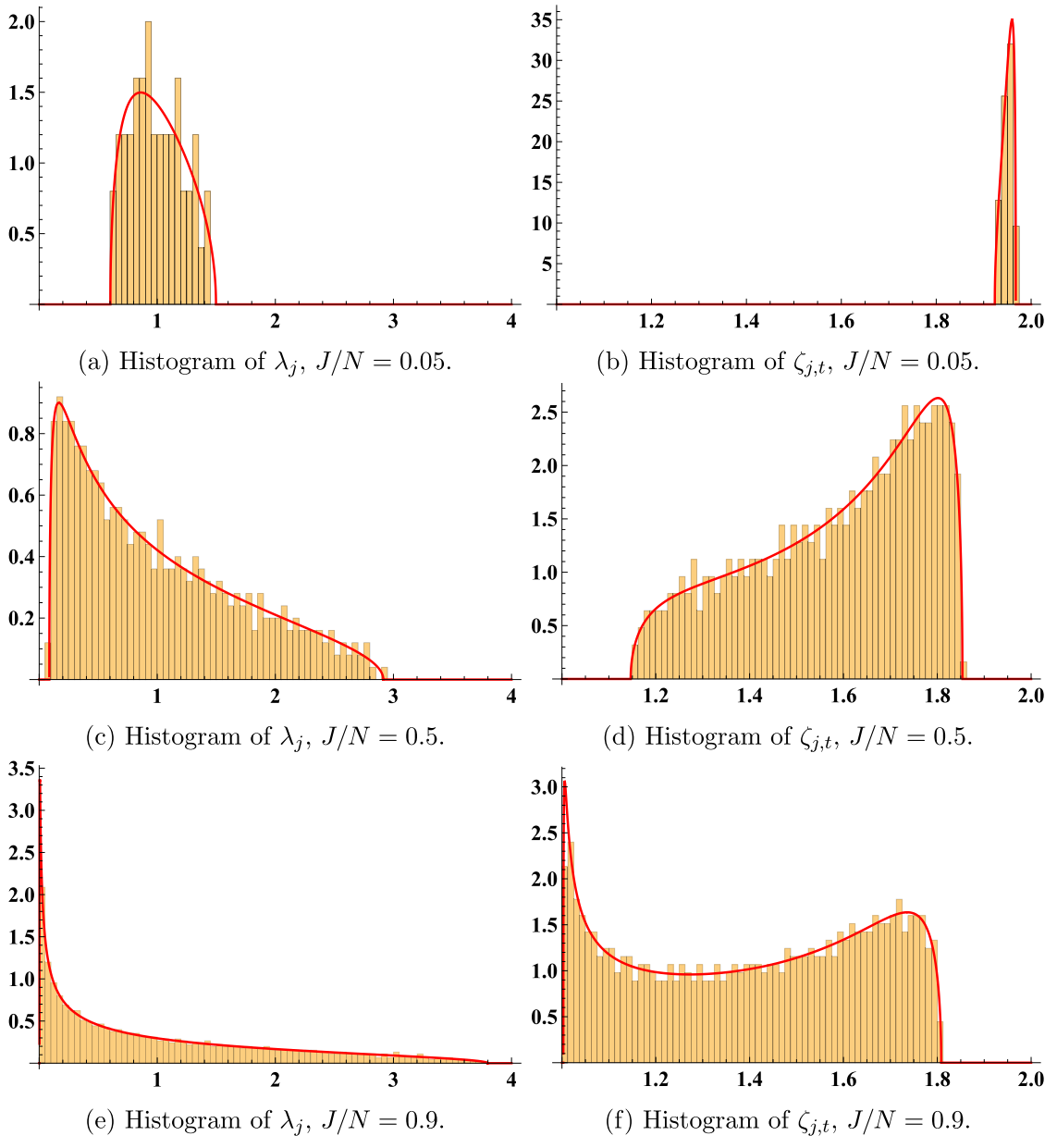


Fig. 1. Histograms of eigenvalue distributions in examples with $\theta = 1$, $t = 1$, $N = 1000$ and $J = 50, 500, 900$. The asymptotic distribution is shown as a solid line in each panel.

This result allows explicit calculation of the limit moments that appear in Propositions 3 and 4. For example, Proposition 4 shows that the probability of rejecting the null of no predictability declines exponentially fast as N increases, and derives the rate of the exponential decay. Proposition 5 can be used to derive an exact analytical expression for the rate in terms of the model primitives θ, ψ, and t.

Figure 2 shows how the rate function (16) depends on ψ for $\theta = 1$ and $t = 1$. For $\psi > 0.4$, the rate is higher than 0.015, indicating that the probability of not rejecting the null is on the order of $\exp(-0.015N)$, which is a tiny number even for relatively small cross-sections of, say, $N \geq 300$.

3.3. (Absence of) out-of-sample return predictability

The situation looks very different with regard to out-of-sample predictability. We now consider a trading strategy that holds stocks in period $t+1$ with weights proportional to predicted returns based on regression coefficients $\boldsymbol{h}_{s+1}$,

$$r_{OOS,t+1} = \boldsymbol{w}'_{OOS,s+1}\boldsymbol{r}_{t+1}, \qquad \boldsymbol{w}_{OOS,s+1} = \frac{1}{N}\boldsymbol{X}\boldsymbol{h}_{s+1} \tag{19}$$

where $s \neq t$ such that the trading strategy is out-of-sample in the sense that the returns used to obtain the coefficient estimates $\boldsymbol{h}_{s+1}$ do not overlap with the returns $\boldsymbol{r}_{t+1}$ used in the evaluation of this strategy.

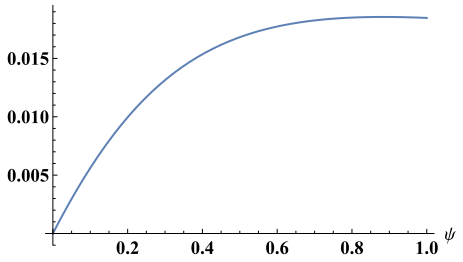


Fig. 2. The rate function given in equation (16).

In the forward out-of-sample case $s < t$, $r_{OOS,t+1}$ is the return on a trading strategy that would be implementable at the end of period $s+1$ based on the econometrician's regression coefficients at that time. We also analyze the backward out-of-sample case where $s > t$. In this case, $r_{OOS,t+1}$ does not represent the return on a tradable strategy, but it is still interesting for econometric evaluation of cross-sectional return predictability where researchers sometimes do go back in time and evaluate trading strategies on new, previously unavailable historical data from earlier time periods.

We obtain the following result for the asymptotic distribution of $r_{OOS,t+1}$:

Proposition 6. If returns are generated according to (8) and $r_{OOS,t+1}$ is calculated as in (19) with $s \neq t$, then

$$\mathbb{E}r_{OOS,t+1} = 0,$$

and, in the large N, J limit,

$$\frac{r_{OOS,t+1}}{\sqrt{\sum_{i=1}^{J} \zeta_{i,s}\zeta_{i,t}}} \xrightarrow{d} N(0,1).$$

In the forward prediction case $t > s$, using last period's estimated coefficients to form the portfolio, this is a natural result. Investors are Bayesian, so the econometrician cannot "beat" investors in predicting returns as long as the econometrician is put on the same footing as investors in terms of the data that are available at the time of making the prediction. Hence, the expected value of $r_{OOS,t+1}$ is zero.

That the result also applies backwards in time, with $t < s$, is more surprising. The result suggests that the econometrician could conduct backwards out-of-sample tests. The fact that many cross-sectional asset-pricing anomalies do not hold up in backwards out-of-sample tests (Linnainmaa and Roberts, 2018) could therefore be a consequence of investor learning, even without data-snooping on the part of researchers who published the original anomaly studies.

While the forward result is likely a general property of Bayesian learning (with objectively correct prior), the backwards result might be somewhat specific to the environment we have set up here (e.g., the assumption that cash-flow growth is IID over time). To what extent one can generalize the backwards result is an interesting question for future research.

The absence of backwards out-of-sample predictability is also interesting because it suggests that the common practice of using cross-validation for evaluating prediction models is justified in an environment like ours where investors face a learning problem. Many recent papers in the emerging machine learning literature in empirical asset pricing use cross-validation (e.g., Feng et al., 2020; Kozak et al., 2020; Bryzgalova et al., 2019). In cross-validation, the data is partitioned repeatedly into estimation and validation samples. For example, several blocks of calendar years are used for model estimation, and other blocks are held for out-of-sample model validation. Then the blocks are switched and the procedure is repeated multiple times. The time ordering of blocks does not matter in this procedure. Some validation blocks can therefore temporally precede some of the estimation blocks. This brings up the concern that cross-validation could pick up learning-induced in-sample predictability. Our backwards result in Proposition 6 suggests that this is not the case. Backwards prediction, forward prediction, and combinations of the two (as in cross-validation) are equivalent. Further below, we explore whether cross-validation approaches could also be useful in assessing the magnitude of out-of-sample predictable cross-sectional variation in returns.

While the result in Proposition 6 that out-of-sample trading strategies with weights based on return forecasts have zero expected returns is straightforward, empirical testing of this prediction is not. The econometrician could try to empirically approximate the asymptotic standard deviation in the denominator of the ratio in Proposition 6 from the time-series standard deviation of the r_{OOS} observations. As we suggested earlier, one could think of one period in this model as roughly a decade and could imagine the econometrician observing intra-period r_{OOS} realizations that are sampled at higher frequency. Would a simple t-test that compares the estimated mean of intraperiod returns to the standard error calculated from intraperiod returns provide a valid test of the $\mathbb{E}r_{OOS,t+1} = 0$ hypothesis?

Unfortunately, the answer is no. The reason is that the expectation in $\mathbb{E}r_{OOS,t+1} = 0$ is unconditional in the sense that it integrates over the distribution of $\boldsymbol{g}$. However, the econometrician only observes one sample for a given draw of $\boldsymbol{g}$. Nature draws $\boldsymbol{g}$ before any data are generated, and the econometrician cannot rerun history for a different value of $\boldsymbol{g}$. Moreover, if the econometrician starts sampling the out-of-sample strategy returns at the beginning of period $t+1$, the sample is also conditioned on the path of $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_t$.

In this one sample available to the econometrician, the distribution of $r_{OOS,t+1}$ has mean $\mathbb{E}\left[r_{OOS,t+1}|\boldsymbol{g}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_t\right]$, which is generally not zero. In the same way, a sample variance of $r_{OOS,t+1}$ in this one sample would estimate $\text{var}(r_{OOS,t+1}|\boldsymbol{g}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_t)$, not the unconditional variance under the square root in the denominator of the expression in Proposition 6. As a consequence, a t-statistic based on this sample mean and variance would not yield a correctly sized test of the Bayesian learning null hypothesis.

Heuristically, at least, there is a relatively simple solution. Since $\mathbb{E}r_{OOS,t+1} = 0$, we have $\text{var}(r_{OOS,t+1}) = \mathbb{E}r^2_{OOS,t+1}$. Moreover, $\mathbb{E}r^2_{OOS,t+1} = \mathbb{E}\mathbb{E}[r^2_{OOS,t+1}|\boldsymbol{g}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_t]$ and so

$\text{var}(r_{OOS,t+1}) = \mathbb{E}\mathbb{E}[r^2_{OOS,t+1}|\boldsymbol{g}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_t]$. Therefore, if the econometrician uses the mean of squared intraperiod returns rather than the sample variance of returns as an estimate of $\text{var}(r_{OOS,t+1})$, this estimate is on average, across samples with different $\boldsymbol{g}$ and $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_t$, equal to the unconditional variance. The reason is that the mean of squared returns captures the deviations of $r_{OOS,t+1}$ from zero due to $\mathbb{E}\left[r_{OOS,t+1}|\boldsymbol{g}, \boldsymbol{e}_1, \ldots, \boldsymbol{e}_t\right] \neq 0$, while in a sample variance calculation these components are removed through demeaning.

3.4. Out-of-sample moment conditions for risk premia estimation

In our model so far, we have abstracted from risk premia or the possibility that investors' subjective beliefs may deviate from the prescriptions of Bayesian updating with objectively correct priors. Risk premia or belief distortions would induce additional predictable components in asset returns beyond those that we have (8). And an empiricist may want to estimate to what extent characteristics $\boldsymbol{X}$ are associated with risk premia or belief distortions, while allowing, at the same time, for the possibility that investors may be learning about the predictive role of $\boldsymbol{X}$ for asset cash flows.

Suppose the characteristics are associated with a predictable component $\boldsymbol{X\gamma}$ for some vector $\boldsymbol{\gamma}$ that represents risk premia (for brevity, we use the label "risk premia" from now on, but with the understanding that nonzero elements in $\boldsymbol{\gamma}$ could arise from belief distortions or frictions, too). In this case, adding this component to the returns in (8), we get

$$\boldsymbol{r}_{t+1} = \boldsymbol{X\gamma} + \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{X\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\tilde{\boldsymbol{e}}_t + \boldsymbol{e}_{t+1}. \quad (20)$$

How can the econometrician estimate the risk premium component $\boldsymbol{X\gamma}$? Under RE, with the second and third term on the right-hand side of (20) absent, the econometrician could just regress $\boldsymbol{r}_{t+1}$ on $\boldsymbol{X}$ in an in-sample regression and estimate $\boldsymbol{X\gamma}$ with the fitted value from this regression. But when investors are learning about $\boldsymbol{g}$, this does not work because the second and third terms are not zero and are correlated with $\boldsymbol{X}$. An OLS regression of $\boldsymbol{r}_{t+1}$ on $\boldsymbol{X}$ would be distorted by the presence of these terms. Along the same lines as in tests of the no-predictability null that we analyzed earlier, in-sample tests of hypotheses about $\boldsymbol{\gamma}$ would be distorted by the presence of the learning-induced components in returns.

We can solve this problem by focusing on the out-of-sample return $r_{OOS,t+1}$. Given the returns (20) and using the results from Proposition 6, it is straightforward to show that

$$\frac{1}{N}\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X\gamma} = \mathbb{E}\boldsymbol{r}_{OOS,t+1}. \quad (21)$$

Therefore, the econometrician could estimate $\frac{1}{N}\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X\gamma}$ by the sample analog of the expectation on the right-hand side. This does not identify the risk premia associated with individual characteristics, but it allows the econometrician to estimate the overall contribution of risk premia to return predictability.

Of course, statistical inference in this case runs into the same problem that we discussed for the out-of-sample test in the previous subsection. Analogous to our earlier discussion, the econometrician could use squared intraperiod realizations of $r_{OOS,t+1}$ to estimate $\text{var}(r_{OOS,t+1})$. In this case, with risk premia components in returns, the squared returns would also incorporate the squared risk premia. As a consequence, $\mathbb{E}r^2_{OOS,t+1} > \text{var}(r_{OOS,t+1})$. Standard errors based on these estimates of $\text{var}(r_{OOS,t+1})$ may therefore be somewhat conservative in the sense that they overstate the sampling variation on average.

3.5. Comparison with cross-validated penalized regression

Since the above estimation procedure amounts to looking for return components that are predictable out-of-sample, one may wonder whether penalized regression with penalty choice based on cross-validation could be an alternative route towards a suitable estimator of quantities like $\frac{1}{N}\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X\gamma}$ that provide an assessment of the magnitude of risk premia. As we show now, there is indeed a close connection between the estimator based on a sample analog of (21) and penalized regression with penalty choice based on cross-validation.

Consider the penalized criterion for an in-sample regression in period t,

$$\boldsymbol{b}_t = \arg\min_{\boldsymbol{b}_t}\left[(\boldsymbol{r}_t - \boldsymbol{X}\boldsymbol{b}_t)'(\boldsymbol{r}_t - \boldsymbol{X}\boldsymbol{b}_t) + \xi\boldsymbol{b}_t'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_t\right] \quad (22)$$

for a given penalty parameter ξ. This second term in this criterion penalizes parameter estimates $\boldsymbol{b}_t$ that imply a large in-sample predictable component of returns, $\boldsymbol{b}_t'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_t$, or, equivalently, a high return of the in-sample portfolio that invests with weights proportional to $\boldsymbol{X}\boldsymbol{b}_t$. This penalty specification is closely related to the approach in Kozak et al. (2020) that estimates stochastic discount factors with a maximum squared Sharpe ratio penalty. It is also similar to ridge regression, but with the difference that standard ridge regression would penalize simply the sum of squared coefficients $\boldsymbol{b}_t'\boldsymbol{b}_t$.

The solution to the problem (22) is

$$\boldsymbol{b}_t = \frac{1}{1+\xi}\boldsymbol{h}_t, \quad (23)$$

or, in other words, simply the OLS regression coefficients shrunk towards zero by a scalar factor that depends on the penalty parameter ξ. Cross-validation then seeks the ξ that minimizes the out-of-sample residual sum of squares, namely,

$$\xi^* = \arg\min_{\xi}\mathbb{E}\left[\left\{\boldsymbol{r}_{t+1} - \frac{1}{1+\xi}\boldsymbol{X}\boldsymbol{h}_{s+1}\right\}' \left\{\boldsymbol{r}_{t+1} - \frac{1}{1+\xi}\boldsymbol{X}\boldsymbol{h}_{s+1}\right\}\right]. \quad (24)$$

Given the optimal ξ, we can then calculate a cross-validated portfolio return with weights based on the shrunk coefficients $\frac{1}{1+\xi^*}\boldsymbol{h}_{s+1}$:

$$r_{CV,s+1} = \boldsymbol{w}_{CV,s+1}'\boldsymbol{r}_{s+1}, \qquad \boldsymbol{w}_{CV,s+1} = \frac{1}{1+\xi^*}\boldsymbol{X}\boldsymbol{h}_{s+1}. \quad (25)$$

This is an in-sample portfolio return where period $s+1$ returns are weighted based on regression coefficients estimated from the same returns, but with shrinkage toward zero induced by $\frac{1}{1+\xi^*}$.

Taking the first-order condition of problem (24), and comparing with the definition of r_{CV} in (25) and r_{OOS} in (19), one can see that it implies

$$\mathbb{E}r_{CV,s+1} = \mathbb{E}r_{OOS,s+1}. \tag{26}$$

Hence, even though $r_{CV,s+1}$ is an in-sample portfolio return, the cross-validated ξ exerts the right amount of shrinkage to remove the learning-induced in-sample predictable variation in returns and isolate $\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\gamma}$, just as the out-of-sample portfolio return $r_{OOS,t+1}$ does according to (21).

Since penalization can be interpreted as shrinkage induced by informative prior beliefs, we can interpret this result as showing that the econometrician can use cross-validation to empirically back out prior beliefs that remove the tendency of in-sample regressions to overstate how much return predictability there really is out-of-sample. Within our Bayesian learning setting, this result therefore provides an economic interpretation of approaches that use cross-validation (Kozak et al., 2020) or related methods (Chinco et al., 2021) to estimate prior beliefs for cross-sectional return prediction.

In our empirical application below, we compare r_{OOS} with a slightly simplified version of r_{CV} that is based on a standard ridge regression, i.e., with penalty proportional to $\boldsymbol{b}_t'\boldsymbol{b}_t$ instead of $\boldsymbol{b}_t'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_t$.

4. Finite-sample analysis: simulations

In this section, we report the results of finite-sample simulations. These simulations provide some insight on the extent to which the asymptotic results provide a good approximation in a setting with realistic J and N. We set $N = 1000$ and let J vary from 1 to close to 1000. We draw the elements of $\boldsymbol{X}$ from a standard Normal distribution.

For the purpose of this numerical analysis, we also need to set the parameter θ that pins down the share of predictable variation in cash-flow growth through $\boldsymbol{\Sigma}_g = \frac{\theta}{J}\boldsymbol{I}$. What matters here is not the total level of cash-flow variance but rather the share that is predictable. For this reason, we normalize, as before, $\boldsymbol{\Sigma}_e = \boldsymbol{I}$. We then look for a value of θ that yields a plausible amount of predictable variation in cash-flow growth relative to this normalized residual variance.

Based on our data-generating process for cash-flow growth in (1), annualized growth rates over a horizon of T periods are

$$\frac{1}{T}\sum_{t=1}^{T}\Delta\boldsymbol{y}_t = \boldsymbol{X}\boldsymbol{g} + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{e}_t.$$

We now evaluate the share that is predictable, given knowledge of $\boldsymbol{g}$. This is an upper bound on the share that investors learning about $\boldsymbol{g}$ may be able to predict. The annualized variance of the predictable component (first term on the right-hand side) is constant with respect to T, while the variance of the residual component (second term on the right-hand side) shrinks at the rate $1/T$. As we indicated earlier, we think of one period in the model as representing roughly one decade. In this case, at a horizon of one decade, i.e., $T = 1$, the variance of the forecastable component is

$$\frac{1}{N}\mathbb{E}\left[\boldsymbol{g}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{g}\right] = \frac{1}{N}\operatorname{tr}\left(\boldsymbol{X}'\boldsymbol{X}\right)\frac{\theta}{J} \approx \theta$$

so the ratio of forecastable to residual variance also equals θ.

We can do a back-of-the-envelope calculation to compare this to the growth rate evidence for various revenue and profit measures in Chan et al. (2003). When stocks are sorted based on IBES analysts' forecasts (their Table IX), Chan et al. find an interdecile spread of slightly around ten percentage points (pp) for annualized growth rates over the next one, three, and five years. Extrapolating to ten years, we would have ten pp also at a ten-year horizon.[8] When they sort stocks instead based on their ex post realized annualized growth rates over a one-year horizon, they find an interdecile range of around 50 pp. Assuming normal distributions, these estimates imply a ratio of about 0.04 of forecastable to residual variance at a one-year horizon. An IID data-generating process for cash flows, as in our model, would imply that the residual variance shrinks at rate $1/T$ and hence the ratio of forecastable to residual variance at a ten-year horizon is 0.4. This share of forecastable variance represents a lower bound, as analysts can predict only a less than full share of the total potentially predictable variance. For this reason, it seems reasonable to set the ratio of maximally predictable to residual variance somewhat higher than 0.4 in our simulations. Accordingly, we set $\theta = 1$.

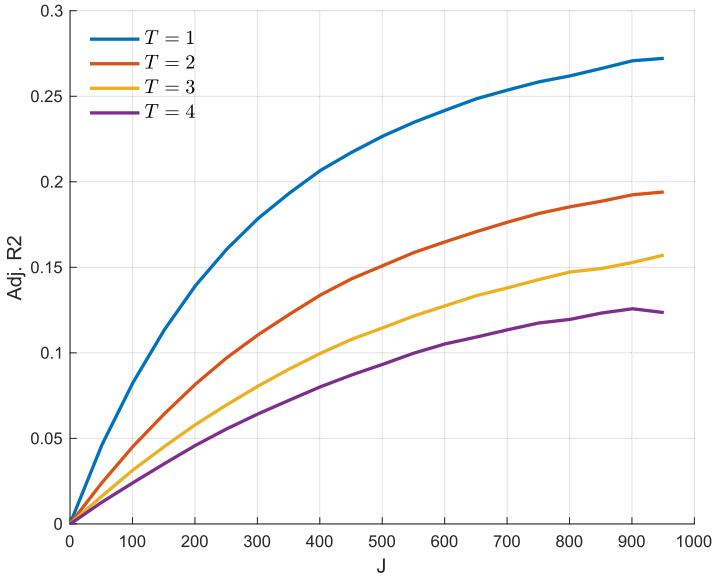
4.1. Return prediction with many predictors

We now simulate cash flows and, based on investors' Bayesian updating and pricing, the returns on the $N = 1000$ assets. We then consider an econometrician who samples these returns ex post and runs regressions of $\boldsymbol{r}_{T+1}$ on $\boldsymbol{X}$ after investors have learned about $\boldsymbol{g}$ for T periods. Figure 3a presents the (in-sample) adjusted R^2 from these regressions. As J increases towards N, the adjusted R^2 also increases; hence, returns become more predictable in-sample. If investors have learned for more periods, return predicability gets weaker.

Figure 3b looks at the properties of a standard Wald test of the no-predictability null hypothesis, testing whether the coefficients on the J predictor variables are jointly equal to zero. The plot shows the rejection probabilities (actual size) from a χ^2–test based on the null distribution in (10). The dotted line shows the nominal size of 5% that the test would have, asymptotically, if investors priced assets under rational expectations with perfect knowledge of $\boldsymbol{g}$. The figure shows that the actual rejection probabilities can be far higher than 5%. The rejection probabilities go to one as J grows towards N. The in-

[8] For this analysis, they do not report percentiles at sufficient detail to calculate the interdecile spread, but they report means for quintile bins, and the spread between top and bottom quintile bin mean should correspond approximately to the interdecile range.

(a) Adjusted R^2



(b) Rejection probability of no-return-predictability null

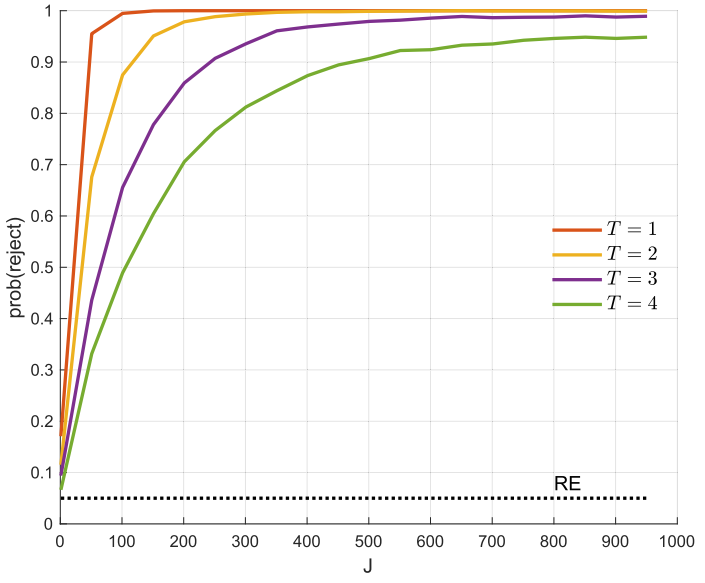


Fig. 3. In-sample return predictability tests. Based on cross-sectional regressions with $N = 1000$ assets and J predictor variables, predicting the last return in a sample of size $T + 1$ and where investors have learned about $\boldsymbol{g}$ from a sample of size T. The test in panel (b) is a joint χ^2–test using all J predictors. It has an asymptotic 5% rejection probability under the rational expectations null hypothesis (where investors know $\boldsymbol{g}$). The solid lines show the actual rejection probabilities when this test is applied in a setting where Bayesian investors with an objectively correct informative prior estimate $\boldsymbol{g}$.

crease is slower if investors have learned for more periods, but even with $T = 4$, the rejection probability exceeds 90% when $J > N/2$. Thus, the simulations confirm that the asymptotic result of rejection probabilities going to one is, indeed, a good approximation for the large J/N case with finite N and J.

4.2. Return prediction with single predictors and multiple testing

In our analysis so far, we assumed that the econometrician runs a kitchen-sink regression using all variables in $\boldsymbol{X}$ as predictors. This may seem unrealistic, as individual

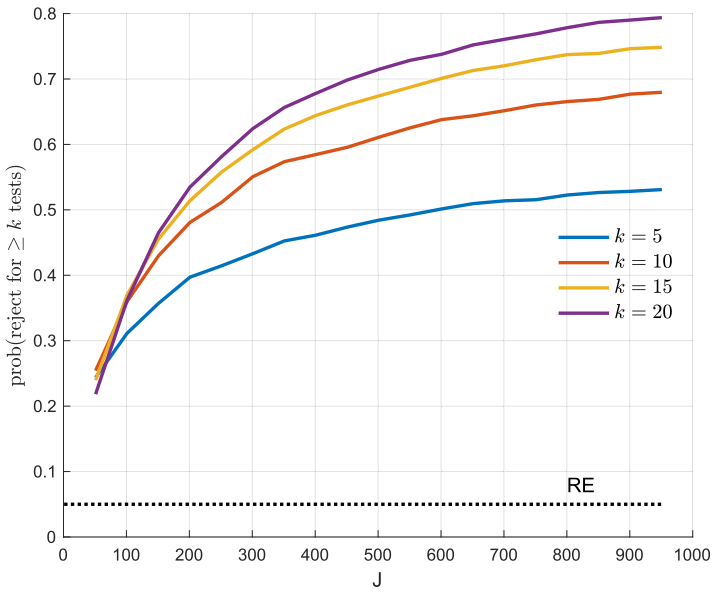


Fig. 4. In-sample return predictability tests with single predictors and correction for multiple testing. Based on J cross-sectional regression with $N = 1000$ assets and one out of J predictor variables in each regression, predicting the last return in a sample of size $T + 1$, and where investors have learned about $\boldsymbol{g}$ from a sample of size $T = 4$. Critical values are adjusted so that the probability that k of the J hypotheses are rejected would be 5% in the rational expectations case (where investors know $\boldsymbol{g}$). The solid lines show the actual rejection probabilities when this test is applied in a setting where Bayesian investors with an objectively correct informative prior estimate $\boldsymbol{g}$.

empirical studies often use individual predictors or small subsets of the universe of predictors observable to the econometrician. Researchers collectively examined a large number of predictors, but not necessarily a large number jointly in individual studies. However, as we show now, the finding of excessive rejection of the no-predictability null is not exclusive to the kitchen-sink setup. Similar conclusions emerge if we consider the possibility that a collection of individual researchers each examine single predictors.

We now imagine that J econometricians, indexed by $j = 1, 2, \ldots, J$, run regressions of $\boldsymbol{r}_{t+1}$ on a single characteristic x_j (i.e., column j of $\boldsymbol{X}$) and each of them tests the hypothesis H_j that $\boldsymbol{x}_j$ does not predict returns. How many of them will reject the no-predictability null? Does high dimensionality of investors' learning problem lead to more rejections?

To obtain an interpretable benchmark, we apply a multiple testing adjustment in which we adjust critical values using the approach of Guo and Romano (2007) such that we control the k-familywise error rate (k-FWER). Specifically, we adjust critical values such that the probability of rejection of at least k of these J hypotheses would be equal to 5%, i.e.,

$$k\text{-FWER} = P(\text{reject at least } k \text{ hypotheses } H_j) = 0.05$$

if asset prices were generated under rational expectations. We then look at actual probability of rejection of at least k hypotheses when asset prices are generated under investor learning.

Figure 4 presents the results for several different values of k and $T = 4$. As in the kitchen-sink regression case, we see a general increase in rejection probabilities relative to the RE benchmark value of 5% when J gets bigger relative to N. For example, with $k = 20$ and $J/N = 0.05$, the actual rejection probability is 25%, compared with a k-FWER–controlled rejection probability of 5% under RE. With J/N close to unity, the actual rejection probability rises to 80%, while it is still 5% under RE. Therefore, similar to the kitchen-sink regression case, high dimensionality leads to likely rejections of the no-predictability null, even if econometricians correctly use a multiple-testing adjustment to evaluate their joint evidence from the J individual tests.

5. Sparsity

So far we have assumed a setting in which shrinkage of coefficients towards zero is the optimal way for investors to deal with the large number of cash-flow predictors. But investors do not impose sparsity (i.e., some coefficients of exactly zero) on the forecasting model. The absence of sparsity was a consequence of the normal prior distribution of $\boldsymbol{g}$. If, instead, investors' prior is that the elements of $\boldsymbol{g}$ are drawn from a Laplace distribution and investors price assets based on the mode rather than the mean of the posterior distribution (i.e., a maximum-a-posteriori estimator), then asset prices reflect sparse cash flow forecasts in which some columns of $\boldsymbol{X}$ are multiplied with coefficients of zero. Their forecasts can then be represented as the fitted values from a Lasso regression (Tibshirani, 1996).

That investors use the posterior mode in pricing is a deviation from the fully Bayesian framework. Our simulations will shed light on how much of a deviation this is in terms

of how much additional return predictability results from it.

With a Laplace prior, elements g_j are IID with the distribution

$$f(g_j) = \frac{1}{2b} \exp\left(-\frac{|g_j|}{b}\right).$$

The variance is $2b^2$. To keep the variance the same as in the normal prior case, we set $2b^2 = \frac{\theta}{J}$. This Laplace distribution not only represents investors' prior, but we now also draw the elements of $\boldsymbol{g}$ in our simulations from this distribution. So the prior is again objectively correct, as in the normal prior case we considered earlier.

Figure 5 shows that the results in the Laplace prior case are extremely similar to those in Fig. 3 for the normal prior case. In terms of how in-sample predictability strengthens with increasing J/N, it does not make much difference whether investors shrink prediction model coefficients with or without sparsity. For the sake of brevity, most of our results in the paper therefore focus on the normal prior case.

6. Excess shrinkage or sparsity

In our analysis up to this point, shrinkage or sparsity was purely due to prior knowledge reflected in investors' prior beliefs. Aside from such statistical optimality considerations, there could be other reasons for investors to shrink coefficients and impose sparsity on their forecasting models. For example, if variables are costly to observe, investors might prefer to discard a variable if it only offers a weak signal about cash flows. Relative to the frictionless Bayesian benchmark with objectively correct prior, such a model would be excessively sparse, but the reduction in forecast performance may be justified by the cost savings from model sparsity. Relatedly, Sims (2003) and Gabaix (2014) show that shrinkage or sparsity can be used to represent boundedly rational decision-making if attention to a variable generates an actual or psychological cost (that shrinkage or sparsity helps avoid).

Such variants of the model with excessive shrinkage can still be mapped into a Bayesian updating scheme, but the prior beliefs are concentrated more tightly around zero than in the case with objectively correct prior (where the prior distribution agrees with distribution that we draw $\boldsymbol{g}$ from in generating the data). For this reason, we label the benchmark case with objectively correct prior as DGP-consistent shrinkage or sparsity.

Figure 6 shows the consequences of excessive shrinkage or sparsity for out-of-sample return predictability. In all cases shown in the figure, the cash-flow data are generated, as before, with $\theta = 1$. However, investors' prior beliefs are now based on a different value of θ. In the excessive shrinkage and sparsity cases, we let investors form beliefs based on $\theta = 0.5$, which means that they have a prior distribution for the elements of $\boldsymbol{g}$ that is more tightly concentrated around zero than the actual distribution of $\boldsymbol{g}$ that generates the data. For comparison, we also consider a case where shrinkage is insufficient. In this case, investors assume $\theta = 2$. This can be interpreted as investors having a lack of confidence, and, hence, excessively wide dispersion, in their prior beliefs about $\boldsymbol{g}$. In all cases, we show results for $T = 4$, which means that investors have learned for four periods up to the beginning of the period in which we measure the return on the out-of-sample portfolio.

The figure shows the out-of-sample return of a portfolio that weights assets by their predicted expected return as in (19). Panel (a) shows the results in the normal prior/ridge regression case. As expected from Proposition 6, the average out-of-sample return in the DGP-consistent case is zero. In contrast, when shrinkage is excessive, investors end up downweighting too strongly the information in $\boldsymbol{X}$ that predicts cash flows. As a result, an econometrician sampling returns from this economy is able to forecast returns out-of-sample. And the effect gets stronger with higher J. Forming a portfolio that weights assets based on their estimated expected returns from predictive regressions on data up to time T earns a predictably positive return in period $T + 1$.

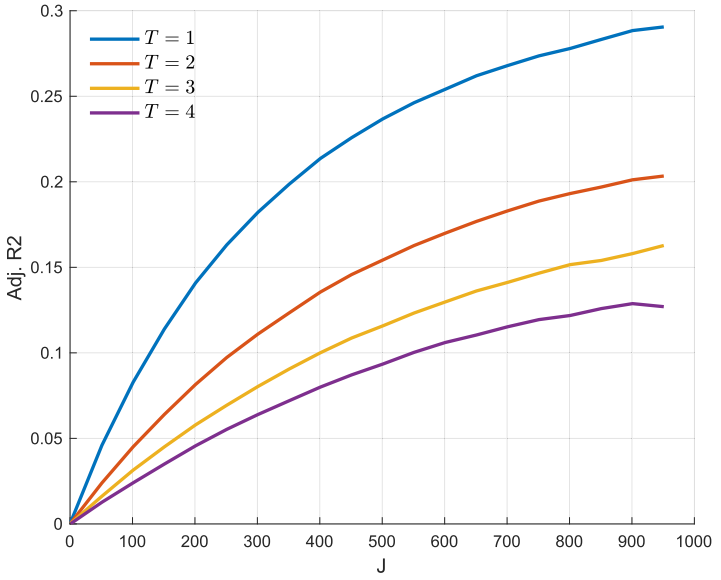
Insufficient shrinkage also results in an out-of-sample average return that differs from zero, but the sign is negative. This means that assets that would be predicted to have positive expected returns, based on the econometrician's predictive regression estimates from data up to time T, actually end up having negative returns in $T + 1$ and vice versa. With insufficient shrinkage, the component $-\boldsymbol{X}\boldsymbol{\Gamma}_t\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t$ in the expression for $\boldsymbol{r}_{t+1}$ in (8) plays a bigger role than under DGP-consistent shrinkage. As a consequence, its negative covariance with the estimation error component $\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}_t$ of the fitted value $\boldsymbol{X}\boldsymbol{h}_t$ from the predictive regression in (9) dominates, which means the forecasts based on $\boldsymbol{X}\boldsymbol{h}_t$ tend to have the wrong sign out-of-sample.

As panel (b) shows, the results in the Laplace prior/lasso regression case are similar. One difference is that the average out-of-sample return in the DGP-consistent case is somewhat negative instead of zero. This is a consequence of the assumption that investors in this case use the mode of the posterior distribution of $\boldsymbol{g}$ (which induces sparsity) to price assets rather than the posterior mean (which would imply a forecasting model that is not sparse).

The takeaway from all this is that out-of-sample return predictability evidence can help shed light on whether investors apply excessive shrinkage or sparsity in their cash flow–forecasting models. For example, if actual or psychological costs of complex models induce additional sparsity, this should show up in the data as out-of-sample predictable returns. Similarly, if investors' prior distribution of $\boldsymbol{g}$ assumed a distribution of coefficients that was too tightly concentrated around zero compared with the true distribution that generated the data, this would show up as out-of-sample predictability.[9]

[9] Note that this would not mean that investors were irrational. Rational Bayesian reasoning does not require that prior beliefs be consistent with the true distribution of $\boldsymbol{g}$, which would be unknown to investors. Existence of out-of-sample return predictability evidence would be consistent not only with the bounded rationality explanation of excess shrinkage or sparsity, but also with a tight-prior explanation. Changes over time in out-of-sample predictability might allow us to disentangle the two.

(a) Adjusted R^2



(b) Rejection probability of no-return-predictability null

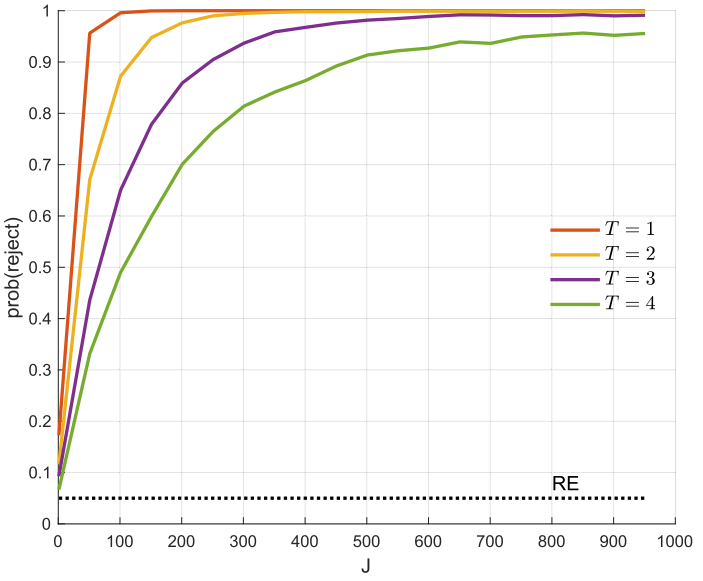
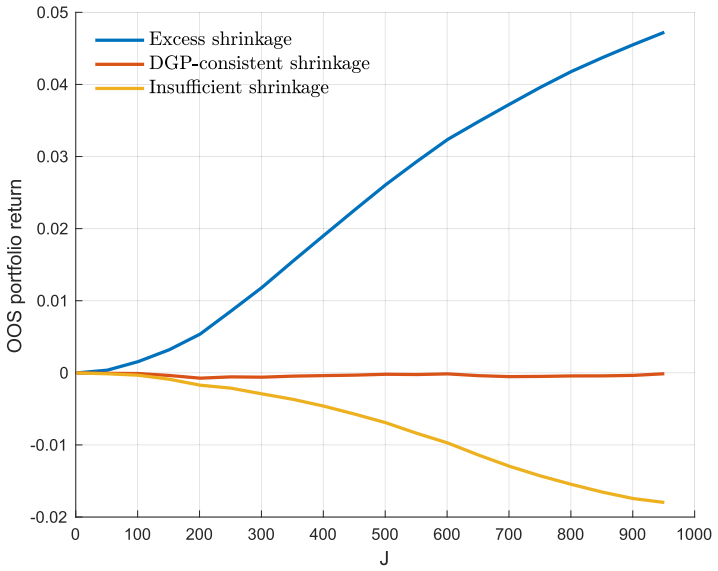


Fig. 5. Lasso: in-sample return predictability tests. Based on cross-sectional regressions with $N = 1000$ assets and J predictor variables, predicting the last return in a sample of size $T + 1$ and where investors have learned about $\boldsymbol{g}$ from a sample of size T. The test in panel (b) is a joint χ^2–test using all J predictors. It has an asymptotic 5% rejection probability under the rational expectations null hypothesis (where investors know $\boldsymbol{g}$). The solid lines show the actual rejection probabilities when this test is applied in a setting where investors estimate $\boldsymbol{g}$ (which is drawn from a Laplace distribution) with Lasso.

This analysis also provides a perspective on the likely effects of technological progress in data construction and data analysis on the return predictability that is observed in empirical analyses. Many studies of the cross-section of stock returns use data that go back to time periods when data availability and analysis were much more constrained than they are today. A researcher today can construct many variables (say, through automated textual analysis of corporate filings) that were inaccessible to investors until not very long ago. In this sense, the forecasting models that investors used when they priced stocks several decades ago may have been excessively sparse relative to the model

(a) Ridge



(b) Lasso

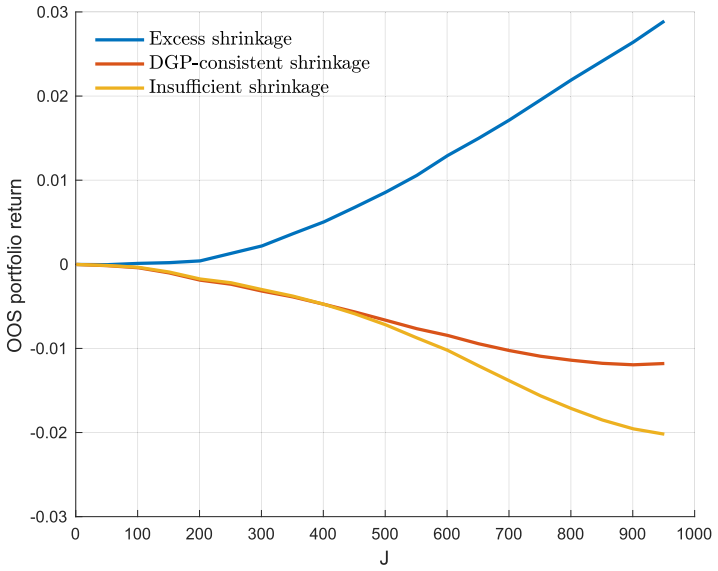


Fig. 6. Out-of-sample portfolio returns when investors apply excess shrinkage or sparsity. Based on cross-sectional regressions with $N = 1000$ assets and J predictor variables, predicting the last return in a sample of size $T + 1$, and where investors have learned about $\boldsymbol{g}$ from sample of size $T = 4$. Cash-flow data are always generated with $\theta = 1$. In the DGP-consistent prior case, investors' prior is based on $\theta = 1$. In the excess shrinkage (or sparsity) case, investors' prior is based on $\theta = 0.5$. In the insufficient shrinkage (or sparsity) case, investors' prior is based on $\theta = 2$. The DGP in all cases always features $\theta = 1$.

that an empirical researcher could work with today. It is to be expected, therefore, that a researcher today can construct variables, or use combinations of large numbers of variables, that predict returns in the earlier years of stock return data sets, even in (pseudo-) out-of-sample tests in which the researcher reconstructs investors' learning process, without taking into account the additional model complexity constraints that investors faced in real time.

7. Empirical application: predicting stock returns with past returns

To illustrate how our model provides an interpretation of discrepancies between in-sample and out-of-sample stock return predictability, we look at an empirical application. The key prediction is that investor learning in a high-dimensional setting should lead to a substantial wedge

between in-sample and out-of-sample predictability. The contribution of risk premia, or mispricing induced by behavioral biases or frictions, to cross-sectional variation in expected returns is revealed by out-of-sample portfolio returns, not in-sample estimates.

For this exercise, we seek a large set of predictor variables that were, at least in principle, consistently available to investors over a long period of time. Many predictors that are based on accounting variables do not satisfy this criterion because they became available in Compustat data only in later decades. Furthermore, to stay close to our setting in the model where the econometrician studies a given set of predictor variables only once, without specification searching and multiple testing, we do not want a set of predictors that may already be the product of data mining efforts by earlier researchers. For example, the set of published predictors in the academic literature likely includes some that have been data mined ex post. To address both concerns, we use each stock's full price history to generate our predictor variables. More precisely, we use the monthly returns and monthly squared returns in the past 120 months as predictors for the next month's stock return. Including squared returns allows the relation between past returns and future returns to be nonlinear.

This price history was, at least in principle, available to investors even in the early parts of the sample. This eliminates the possibility that return predictability could show up ex post simply because a variable that we can construct today was not available to investors in real time when they priced stocks. Of course, the fact that the price history was available in principle does not necessarily mean that investors throughout the sample always had the ability to integrate all of these variables into their forecasting models. If they could not, we might find out-of-sample return predictability in parts of the sample where technological constraints may have prevented investors from doing so, which would be consistent with our excess shrinkage results in the previous section.

Focusing on price history–based predictors, without preselecting particular subsets of them based on earlier evidence of predictability, allows us to sidestep, for the most part, the influence of earlier researchers' data mining. The only potential remaining problem is that our choice of considering price history–based predictors as a class could be influenced by existing evidence that subsets of these seem to have predictive power (e.g., momentum, long-run reversal). On the other hand, the class of price history–based predictors would surely be a natural candidate even in absence of any existing evidence, given that weak-form efficiency is the most basic market efficiency notion (Fama, 1970).

A drawback of price history–based predictors is that they do not perfectly map into our model. In our theoretical analysis, we worked with an exogenous cash-flow predictor matrix $\boldsymbol{X}$. In contrast, past returns are an equilibrium outcome. One could, however, imagine an extension of the model in which cash flow–growth shocks could have persistent components at various lags. In this case, investors' set of potential cash-flow predictors would include the history of past cash flow–growth shocks and lagged returns would be correlated with these. In this sense, the distance from our model is not that big. In any case, the purpose of the empirical analysis is not to provide a formal test of the model but rather to illustrate, in a simple setting with a large number of predictors, the wedge between in-sample and out-of-sample predictability.

We use all U.S. stocks in the CRSP database except small stocks that have market capitalization below the 20th NYSE percentile or a price lower than one dollar at the end of month $t-1$. To avoid picking up microstructure related issues, we skip the most recent month in our construction of the set of predictor variables. Thus, we use simple and squared returns in months $t-2$ to $t-120$ (i.e., a total of 238 predictor variables) to predict returns in month t in a panel regression. We demean the dependent variable and all explanatory variables month by month to focus purely on cross-sectional variation. In addition, we cross-sectionally standardize all predictor variables to unit standard deviation each month. We weight the observations each month such that the panel regression gives equal weight to each month in the sample.

As a first step, to demonstrate that a regression with shrinkage delivers meaningful estimates with such a large number of predictor variables, we examine an in-sample panel ridge regression to predict monthly returns from the beginning of 1971 until end of June 2019. We show that the ridge regression automatically recovers many prominent predictability patterns that have been documented in the existing literature for roughly this sample period or parts of it. We pick the penalty hyperparameter that determines the strength of shrinkage through leave-one-year-out cross-validation.[10]

Panel (a) in Fig. 7 presents the regression coefficients for each of the 119 simple return explanatory variables. It shows that a single ridge regression recovers several major anomalies related to past returns: the positive coefficients up to lags of 12 months capture momentum as in Jegadeesh and Titman (1993); the plot also shows that continuation of recent returns is concentrated in lags seven to 12, as pointed out in Novy-Marx (2012); the mostly negative coefficients for lags beyond 12 months reflect long-term reversals as in DeBondt and Thaler (1985); the positive coefficients at lags equal to multiples of 12 reflect the momentum seasonality reported by Heston and Sadka (2008). Panel (b) reports the regression coefficients for the 119 lagged squared returns. At shorter lags, there is no clear pattern. But at longer lags beyond lag 50, the coefficients are predominantly positive, indicating a positive association of long-run lagged individual stock return volatility and future returns.

Figure 8 presents three series of monthly portfolio returns based on different versions of these regressions within 20-year rolling windows. The in-sample portfolio return, r_{IS}, is based on a single OLS regression within a

[10] We compute the estimates using all but one year of the sample, we calculate the implied predicted returns in the year left out of the estimation, and we record the resulting R^2 in the left-out year. We then repeat with a different left-out year, again record the R^2 in the left-out years, and repeat until each year of the sample has been left out once. At the end, we average the R^2 across all left-out years, and we search for a penalty value that maximizes this cross-validated R^2.

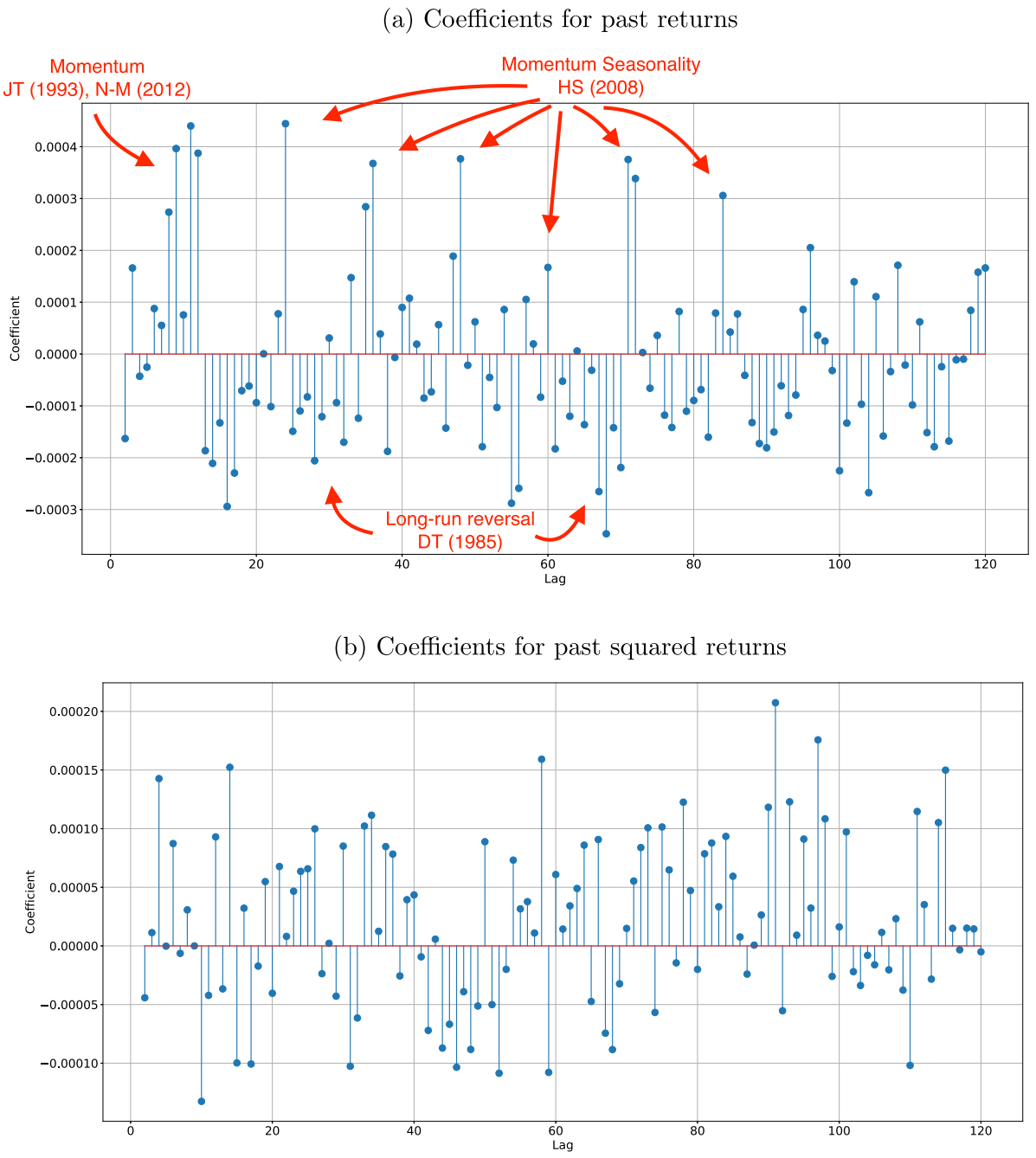


Fig. 7. Ridge regression coefficient estimates.

given window. The portfolio weights are proportional to the predicted returns based on the OLS estimates as in (3.1). Note that the portfolio weights are not normalized to a certain dollar amount long or short. To interpret the magnitude without this normalization, recall that r_{IS} is equal to the in-sample explained return variation of the OLS regression. So, as shown in the figure, r_{IS} of around 0.006%, compared with a monthly cross-sectional return variance of around 0.6%, means that the OLS R^2 is around 1%.

To illustrate the empirical gap between r_{IS} and an out-of-sample portfolio return, the figure also plots r_{OOS}. To construct out-of-sample portfolio returns, we run rolling OLS regressions in 20-year estimation windows. We use the OLS estimates $\boldsymbol{h}_t$ of the window ending in month t to forecast returns in month $t+1$. Given the 20-year estimation window and the up to 10-year lag of the predictors, the first month in which we have a prediction is January 1956, 30 years after the start of the CRSP database. We use the estimates $\boldsymbol{h}_t$ to calculate the out-of-sample port-

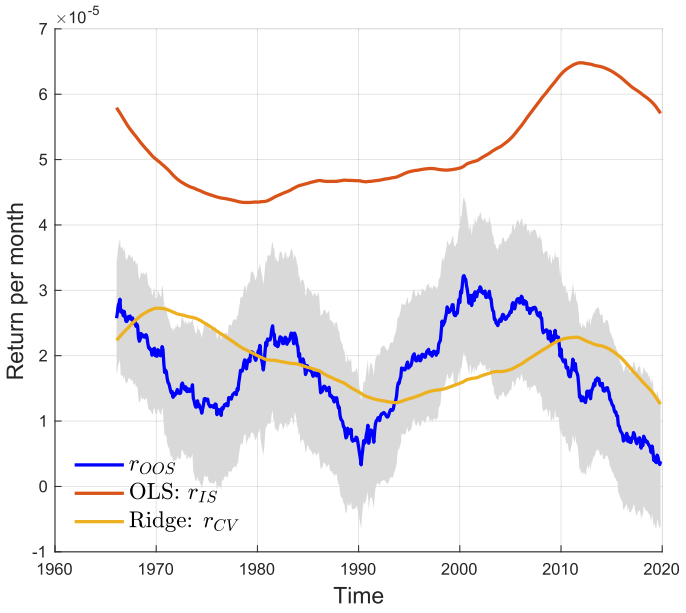


Fig. 8. Rolling estimates of the risk premium/mispricing component of returns. The risk premium component $\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\gamma}$ is estimated by the out-of-sample portfolio return $r_{OOS,t+1} = \boldsymbol{r}'_{t+1}\boldsymbol{X}\boldsymbol{h}_t$. We obtain the OLS estimates $\boldsymbol{h}_t$ from regressions of individual stock returns on lagged returns and lagged squared returns in backwards 20-year moving windows up to month t and use them to construct weights applied to month $t+1$ returns. Stocks with market capitalization below the 20th NYSE percentile and lagged price lower than one dollar are excluded. The blue line in the figure shows $r_{OOS,t+1}$ averaged in 10-year moving windows. The shared area indicates two–standard error bands. For comparison, the figure also shows the in-sample return based on cross-validated ridge regression estimates, r_{CV}, and OLS estimates, r_{IS}. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

folio return in month $t+1$ with weights proportional to predicted returns as in (19). The figure shows this out-of-sample portfolio return averaged over 10-year moving windows.

As the figure shows, r_{IS} is very stable across time and reliably above zero. In contrast, the moving average of the out-of-sample return r_{OOS} is much lower and frequently close to zero or below. The shaded area of the figure shows two–standard error bands obtained from the mean of squared monthly portfolio returns within the 10-year moving windows, in line with our discussion following Proposition 6 on estimation of $\text{var}(r_{OOS,t+1})$. Towards the end of the sample, in the 10-year windows ending between 2014 and 2019, the out-of-sample portfolio return has been less than two standard errors above zero. As we noted in Section 3.4, the average r_{OOS} can be interpreted as an estimate of $\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\gamma}$, the contribution of risk premia (or return premia induced by behavioral biases or frictions) to cross-sectional variation in expected returns. The big wedge between r_{IS} and r_{OOS} shown in this figure underscores the message from our model that in-sample cross-sectional return predictability evidence is not a good motivation for seeking risk-based or behavioral economic explanations. Much of the in-sample predictability here does not carry over into out-of-sample predictability and hence does not reflect risk premia demanded by investors ex ante or persistent belief distortions.

The figure also plots r_{CV}, the in-sample return on a portfolio with weights proportional to predicted returns from rolling cross-validated ridge regression estimates. Recall that our result in (26) suggested $\mathbb{E}r_{OOS} = \mathbb{E}r_{CV}$. This result was based on cross-validated shrinkage with a penalty $\boldsymbol{b}_t'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_t$ in (22), while here we use standard ridge regression shrinkage with a penalty on $\boldsymbol{b}_t'\boldsymbol{b}_t$. Nevertheless, as the figure shows, this equality approximately holds in the empirical data even with a somewhat different penalty specification. Therefore, applying cross-validated ridge regression shrinkage to the portfolio weights is a useful alternative route to obtaining an estimate of $\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\gamma}$.

There are a number of concerns one might have about this analysis. First, would it not be possible to pick, ex post, a much smaller number of lags of simple and/or squared returns that happen to do better, in-sample and in the (pseudo-) out-of-sample tests from the rolling regressions? This may well be true. But what would be the ex ante justification to pick those specific lags? Why not others? Pursuing this avenue would inevitably introduce data snooping and multiple testing concerns that cloud the interpretation of the evidence. As an example, consider that the literature on momentum has shifted from emphasizing the returns in months $t-2$ to $t-12$, as in Jegadeesh and Titman (1993), to highlighting returns in months $t-7$ to $t-12$ (Novy-Marx, 2012). It is not clear that one should seek deep economic reasons for in-sample predictability results that reflect such ex post data-driven specification changes. Simply including all lags up to a certain point and and letting the shrinkage take care of preventing overfitting minimizes this data-snooping problem.

Second, out-of-sample tests may have low power to detect return predictability. This point has been made by Inoue and Kilian (2005), Campbell and Thompson (2008), and Hansen and Timmermann (2015) in a time-series set-

ting. However, their arguments are based on a rational expectations framework in which investors know the parameters of the data-generating process and only the econometrician faces the problem of recovering its parameters from observed data. In their case, in-sample and out-of-sample methods test the same economic hypothesis. But if investors are learning about parameters, especially in settings where the number of potentially relevant predictor variables is huge, the situation is fundamentally different. There is no fixed-parameter model that generates each period's returns. Instead, the properties of the data evolve over time as investors learn. As a consequence, in-sample and out-of-sample methods test different economic hypotheses. In-sample predictability tests basically lose the economic meaning they have in a rational expectations setting because they cannot discriminate between predictability induced by learning and predictability induced by risk premia or behavioral biases. Only out-of-sample tests can do so. For this reason, even if out-of-sample tests have low power, in-sample tests are simply not a viable alternative method, because they test a different hypothesis without clear economic interpretation.

8. Conclusion

Our analysis provides a new perspective on markets in which decision-makers face high-dimensional prediction problems. Learning how to translate observed predictor variables into forecasts is hard when the number of predictors is comparable in size to the number of observations. To an econometrician studying these forecasts ex post or the equilibrium prices that reflect these forecasts the forecast errors look predictable. However, they are not predictable to the decision-maker in real time. We developed this analysis in a cross-sectional asset-pricing application, but the issue may be relevant more broadly in settings in which large numbers of variables are potentially relevant for forecasting.

In the cross-sectional asset-pricing setting, in-sample tests of return predictability lose their economic meaning when investors are faced with many possible predictors of asset cash flows. The usual economic interpretation that in-sample predictable returns represent priced risks or the effects of investors' behavioral biases does not apply in this case. This is not a statistical problem with the sampling properties of the econometrician's predictability tests. Instead, it is a problem with the null hypothesis in these tests. As investors' learning problem becomes harder with increasing dimensionality of the set of potential predictors, the true properties of equilibrium prices change. Even in the absence of risk premia and behavioral biases, the usual null hypothesis that returns are unpredictable need not apply. Investors' learning of the cash flow–forecasting model parameters leaves in-sample predictable components in returns that reflect investors' real-time estimation error and the shrinkage they optimally apply to reduce it in a high-dimensional setting. This is true even though we kept investors' learning problem very simple: the potentially predictable component of future cash-flow growth is linear in predictors, and investors know this linear functional form. If investors also had to entertain that the functional form could be nonlinear, this would further magnify the dimensionality of the prediction problem they face.

In contrast to in-sample tests, out-of-sample tests retain their economic meaning in the high-dimensional case. Our argument in favor of out-of-sample tests is different from those usually discussed in the econometrics literature. The usual case for out-of-sample tests motivates them as remedies against distortions of the sampling properties of in-sample tests or against data mining. As Inoue and Kilian (2005), Campbell and Thompson (2008), Cochrane (2008), and Hansen and Timmermann (2015) have pointed out, the arguments in favor of out-of-sample testing are questionable in settings where the null hypothesis is a population model with truly unpredictable returns. Our point is that when investors face a high-dimensional forecasting problem, this is not an economically interesting null hypothesis. Absence of risk premia and behavioral biases implies absence of out-of-sample predictability but not of in-sample predictability. As investors arguably face a high-dimensional prediction problem in the real world, researchers should give more emphasis to out-of-sample testing.

Our results offer a novel interpretation of the fact that in-sample return predictability tests in the literature have produced hundreds of variables that appear to predict returns in the cross-section. As the number of predictor variables that are available to researchers and investors has grown enormously, it is to be expected, even with fully rational Bayesian investors, that returns should be predictable in hindsight from the perspective of an econometrician running in-sample regressions. Our results show that many such variables do indeed show up as in-sample statistically significant cross-sectional return predictors. But it is not clear, in the absence of a clear theoretical motivation for a predictor variable, or collection of variables, that one should look for risk-based explanations or behavioral explanations for their in-sample predictive power.

A number of extensions of our work could be interesting. Our setting is a purely cross-sectional one with firm characteristics that are constant over time. But a similar learning problem also exists in the time dimension, e.g., at the aggregate stock market level. A huge number of macro variables could, jointly, be relevant for predicting aggregate stock market fundamentals. Furthermore, to keep the model simple and transparent, we have focused on learning about exogenous fundamentals with homogeneous investors. It would be interesting to extend this to a setting with heterogeneous investors. Balasubramanian and Yang (2020) make some progress in this direction by considering privately informed investors who are uncertain about each others' priors in a high-dimensional environment. More generally, investor heterogeneity can generate a role for endogenous price-based signals from which investors can extract information about not only asset fundamentals but also the trading behavior of other investors.

Appendix A. Proofs

We first recall some notation and some basic facts that we will exploit throughout this appendix. We will use the

eigendecomposition

$$\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}', \tag{A.1}$$

and the definition

$$\boldsymbol{\Gamma}_t = \boldsymbol{Q}\left(\boldsymbol{I} + \frac{J}{N\theta t}\boldsymbol{\Lambda}^{-1}\right)^{-1}\boldsymbol{Q}'. \tag{A.2}$$

It will be convenient for future use to note that (A.1) and (A.2) imply that

$$\boldsymbol{I} - \boldsymbol{\Gamma}_t = \boldsymbol{Q}\left(\boldsymbol{I} + \frac{N\theta t}{J}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{Q}' \tag{A.3}$$

and

$$\frac{\theta t}{J}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{\Gamma}_t. \tag{A.4}$$

We will repeatedly exploit the fact that $\boldsymbol{Q}$ is orthogonal (that is, $\boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$) and $\boldsymbol{\Lambda}$ is diagonal. Note further that diagonal matrices commute. Lastly, Assumption 4 implies that (i) $\boldsymbol{\Sigma}_g = (\theta/J)\boldsymbol{I}$, (ii) $\mathbb{E}\tilde{\boldsymbol{e}}_t\boldsymbol{e}_t' = \mathbb{E}\boldsymbol{e}_t\tilde{\boldsymbol{e}}_t' = \mathbb{E}\tilde{\boldsymbol{e}}_t\tilde{\boldsymbol{e}}_t' = \frac{1}{t}\boldsymbol{I}$, (iii) $\mathbb{E}\boldsymbol{e}_t\boldsymbol{e}_t' = \boldsymbol{I}$; and, for $s < t$, (iv) $\mathbb{E}\tilde{\boldsymbol{e}}_t\boldsymbol{e}_{s+1}' = (1/t)\boldsymbol{I}$, (v) $\mathbb{E}\boldsymbol{e}_{t+1}\tilde{\boldsymbol{e}}_s' = 0$, and (vi), $\mathbb{E}\tilde{\boldsymbol{e}}_t\tilde{\boldsymbol{e}}_s' = (1/t)\boldsymbol{I}$.

Proof of Proposition 1. The result follows on making Assumptions 1 and 3 in the Bayesian linear model of Lindley and Smith (1972). □

Proof of Proposition 2. The realized return is

$$\boldsymbol{r}_{t+1} = \boldsymbol{y}_{t+1} - \boldsymbol{p}_t = \boldsymbol{y}_{t+1} - \left[\boldsymbol{y}_t + \widetilde{\mathbb{E}}_t(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1})\right] = \Delta\boldsymbol{y}_{t+1} - \boldsymbol{X}\tilde{\boldsymbol{g}}_t.$$

From Assumption 1 and Eq. (6), this becomes

$$\begin{aligned}\boldsymbol{r}_{t+1} &= \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1} - \boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\overline{\Delta\boldsymbol{y}}_t \\ &= \boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1} - \boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\boldsymbol{X}\boldsymbol{g} + \bar{\boldsymbol{e}}_t] \\ &= \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t + \boldsymbol{e}_{t+1}.\end{aligned}$$

It follows that $\mathbb{E}\boldsymbol{r}_{t+1} = 0$ and

$$\begin{aligned}\mathbb{E}\boldsymbol{r}_{t+1}\boldsymbol{r}_{t+1}' &= \frac{\theta}{J}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)^2\boldsymbol{X}' + \frac{1}{t}\boldsymbol{X}\boldsymbol{\Gamma}_t\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{\Gamma}_t\boldsymbol{X}' + \boldsymbol{I} \\ &\overset{(A.4)}{=} \frac{\theta}{J}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)^2\boldsymbol{X}' + \frac{\theta}{J}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{\Gamma}_t\boldsymbol{X}' + \boldsymbol{I} \\ &= \frac{\theta}{J}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{X}' + \boldsymbol{I},\end{aligned}$$

using (A.4) in the second line. □

Proof of Proposition 3. We form $\boldsymbol{h}_{t+1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{r}_{t+1}$ and then look at the in-sample return $\boldsymbol{r}_{t+1}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{r}_{t+1} = \boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1}$. To think about the distribution of this quantity, we first need to understand $\boldsymbol{h}_{t+1}$ itself. It is a zero mean Normal random vector, and as $\mathbb{E}\boldsymbol{r}_{t+1}\boldsymbol{r}_{t+1}' = \frac{\theta}{J}\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{X}' + \boldsymbol{I}$ by Proposition 2, we have

$$\begin{aligned}\mathbb{E}\boldsymbol{h}_{t+1}\boldsymbol{h}_{t+1}' &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}\boldsymbol{r}_{t+1}\boldsymbol{r}_{t+1}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \frac{\theta}{J}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \frac{\theta}{J}(\boldsymbol{I} - \boldsymbol{\Gamma}_t) + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned}$$

Using Eqs. (A.1) and (A.3), this can be rewritten as

$$N\mathbb{E}\boldsymbol{h}_{t+1}\boldsymbol{h}_{t+1}' = \underbrace{\boldsymbol{Q}\left[\left(\frac{J}{N\theta}\boldsymbol{I} + t\boldsymbol{\Lambda}\right)^{-1} + \boldsymbol{\Lambda}^{-1}\right]}_{\boldsymbol{\Omega}_t}\boldsymbol{Q}'.$$

If we define $\boldsymbol{u}_{t+1} = \sqrt{N}\boldsymbol{\Omega}_t^{-1/2}\boldsymbol{Q}'\boldsymbol{h}_{t+1}$, then $\sqrt{N}\boldsymbol{h}_{t+1} = \boldsymbol{Q}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{u}_{t+1}$ and $\boldsymbol{u}_{t+1}$ is standard Normal: $\mathbb{E}\boldsymbol{u}_{t+1}\boldsymbol{u}_{t+1}' = N\boldsymbol{\Omega}_t^{-1/2}\boldsymbol{Q}'\frac{1}{N}\boldsymbol{Q}\boldsymbol{\Omega}_t\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\Omega}_t^{-1/2} = \boldsymbol{I}$ (using orthogonality of $\boldsymbol{Q}$). We can then write

$$\begin{aligned}\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} &= \underbrace{\boldsymbol{u}_{t+1}'\boldsymbol{\Omega}_t^{1/2}\boldsymbol{Q}'}_{\sqrt{N}\boldsymbol{h}_{t+1}'}\underbrace{\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'}_{\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}}\underbrace{\boldsymbol{Q}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{u}_{t+1}}_{\sqrt{N}\boldsymbol{h}_{t+1}} \\ &= \boldsymbol{u}_{t+1}'\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Lambda}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{u}_{t+1} \\ &= \boldsymbol{u}_{t+1}'\boldsymbol{\Omega}_t\boldsymbol{\Lambda}\boldsymbol{u}_{t+1}.\end{aligned}$$

The last line exploits the fact that $\boldsymbol{\Omega}^{1/2}$ and $\boldsymbol{\Lambda}$ commute, as they are diagonal.

It follows that

$$\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} = \sum_{i=1}^{J}\zeta_{i,t}u_i^2, \tag{A.5}$$

where $\zeta_{i,t}$ are the diagonal entries of the diagonal matrix $\boldsymbol{\Omega}_t\boldsymbol{\Lambda}$ and u_i are independent $N(0,1)$ random variables (the entries of $\boldsymbol{u}_{t+1}$). Explicitly,

$$\zeta_{i,t} = \omega_{i,t}\lambda_i = \frac{\lambda_i}{t\lambda_i + \frac{J}{\theta N}} + 1. \tag{A.6}$$

As $\lambda_i > 0$ by positive definiteness of $\boldsymbol{X}'\boldsymbol{X}$, it follows that for $t \geq 1$, $\zeta_{i,t} \in (1,2)$. Moreover, as $\lim_{J,N\to\infty}\frac{J}{N} = \psi > 0$ and (by Assumption 5) $\lambda_i > \varepsilon$, $\zeta_{i,t}$ is uniformly bounded away from 1 and 2. It follows that $\mu \in (1,2)$ and $\sqrt{\mu^2 + \sigma^2} \in (1,2)$.

We will apply Lyapunov's version of the central limit theorem to $\sum_{i=1}^{J}\zeta_{i,t}u_i^2$, which here requires that for some $\delta > 0$,

$$\lim_{NJ\to\infty}\frac{1}{s_J^{2+\delta}}\sum_{i=1}^{J}\zeta_{i,t}^{2+\delta}\mathbb{E}\left[\left|u_i^2 - 1\right|^{2+\delta}\right] = 0 \quad \text{where}$$

$$s_J^2 = 2\sum_{i=1}^{J}\zeta_{i,t}^2.$$

It is enough to show that this holds for $\delta = 2$. But as $\mathbb{E}\left[(u_i^2 - 1)^4\right] = 60$ and $\zeta_{i,t} \in (1,2)$,

$$\frac{1}{s_J^4}\sum_{i=1}^{J}\zeta_{i,t}^4\mathbb{E}\left[\left|u_i^2 - 1\right|^4\right] = \frac{60\sum_{i=1}^{J}\zeta_{i,t}^4}{\left(2\sum_{i=1}^{J}\zeta_{i,t}^2\right)^2} \leq \frac{960J}{4J^2} \to 0 \quad \text{as} J \to \infty,$$

as required. Therefore the central limit theorem applies for $\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} = \sum_{i=1}^{J}\zeta_{i,t}u_i^2$ after appropriate standardization by mean and variance, which (as the u_i are IID standard Normal) are $\sum_{i=1}^{J}\zeta_{i,t}$ and $2\sum_{i=1}^{J}\zeta_{i,t}^2$, respectively. Thus we have

$$T_b \equiv \frac{\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} - \sum_{i=1}^{J}\zeta_{i,t}}{\sqrt{2\sum_{i=1}^{J}\zeta_{i,t}^2}} \xrightarrow{d} N(0,1).$$

The remaining results follow immediately. □

Proof of Proposition 4. The first statement follows from the second. To prove the second, note that Proposition 3 implies that

$$\mathbb{P}(T_{re} < c_\alpha) = \mathbb{P}\left(\frac{T_{re}}{\sqrt{\mu^2+\sigma^2}} - \frac{\mu-1}{\sqrt{2\left(\mu^2+\sigma^2\right)}}\sqrt{J} < \frac{c_\alpha}{\sqrt{\mu^2+\sigma^2}} - \frac{\mu-1}{\sqrt{2\left(\mu^2+\sigma^2\right)}}\sqrt{J} \right) \to \Phi\left(\frac{c_\alpha}{\sqrt{\mu^2+\sigma^2}} - \frac{\mu-1}{\sqrt{2\left(\mu^2+\sigma^2\right)}}\sqrt{J} \right),$$

where $\Phi(\cdot)$ denotes the standard Normal cumulative distribution function. The result follows from the well-known inequalities $\frac{e^{-x^2/2}}{\left|x+\frac{1}{x}\right|\sqrt{2\pi}} < \Phi(x) < \frac{e^{-x^2/2}}{|x|\sqrt{2\pi}}$, which hold for $x < 0$. □

Proof of Proposition 5. When $t > 0$, the cross-sectional moments of $\zeta_{j,t}$ can be computed using Eq. (14) and the fact that the eigenvalues λ_j follow (in the asymptotic limit) the Marchenko-Pastur distribution, whose probability density function $f_\lambda(x)$ takes the form

$$f_\lambda(x) = \frac{1}{2\pi}\frac{\sqrt{\left[\left(1+\sqrt{\psi}\right)^2 - x\right]\left[x - \left(1-\sqrt{\psi}\right)^2\right]}}{\psi x}$$

if $\left(1-\sqrt{\psi}\right)^2 \le x \le \left(1+\sqrt{\psi}\right)^2$, and $f_\lambda(x) = 0$ elsewhere. The relevant integrals can be calculated explicitly, giving Eqs. (17) and (18); and we can calculate the probability density function of $\zeta_{j,t}$ in the asymptotic limit by change of variable using the relation between $\zeta_{j,t}$ and λ_j given in Eq. (A.6). □

Proof of Proposition 6. We first show that $\mathbb{E}r_{OOS,t+1} = 0$. As $\boldsymbol{h}_{s+1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{r}_{s+1}$, we have

$$\mathbb{E}\left[\boldsymbol{r}_{t+1}(\boldsymbol{X}\boldsymbol{h}_{s+1})'\right] = \mathbb{E}\left[\boldsymbol{r}_{t+1}\boldsymbol{r}_{s+1}'\right]\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'.$$

We will show that $\mathbb{E}\left[\boldsymbol{r}_{t+1}\boldsymbol{r}_{s+1}'\right] = 0$ when $s \neq t$; in other words, all non-contemporaneous autocorrelations and cross-correlations are zero. Henceforth we assume that $s < t$ without loss of generality. From Eq. (8),

$$\mathbb{E}\left[\boldsymbol{r}_{t+1}\boldsymbol{r}_{s+1}'\right] = \frac{\theta}{J}\boldsymbol{X}(\boldsymbol{I}-\boldsymbol{\Gamma}_t)(\boldsymbol{I}-\boldsymbol{\Gamma}_s)\boldsymbol{X}' + \frac{1}{t}\boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\Gamma}_s\boldsymbol{X}' - \frac{1}{t}\boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'.$$

This expression can be rearranged as

$$\mathbb{E}\left[\boldsymbol{r}_{t+1}\boldsymbol{r}_{s+1}'\right] = \frac{1}{t}\boldsymbol{X}\left[\frac{\theta t}{J}(\boldsymbol{I}-\boldsymbol{\Gamma}_t)\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{\Gamma}_t\right](\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{I}-\boldsymbol{\Gamma}_s)\boldsymbol{X}'.$$

As $\frac{\theta t}{J}(\boldsymbol{I}-\boldsymbol{\Gamma}_t)\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{\Gamma}_t$ by Eq. (A.4), the term in square brackets on the right-hand side vanishes, and the result follows.

We now turn to the asymptotic distribution. As $\boldsymbol{r}_{t+1}'\boldsymbol{X}\boldsymbol{h}_{s+1} = \boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{s+1}$, we want to understand the behavior of $\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{s+1}$ where $s < t$. Defining $\boldsymbol{u}_{t+1} = \sqrt{N}\boldsymbol{\Omega}_t^{-1/2}\boldsymbol{Q}'\boldsymbol{h}_{t+1}$, as in the proof of Proposition 3, we have $\sqrt{N}\boldsymbol{h}_{t+1} = \boldsymbol{Q}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{u}_{t+1}$ and $\mathbb{E}\boldsymbol{u}_{t+1}\boldsymbol{u}_{t+1}' = \boldsymbol{I}$. We also have $\mathbb{E}\boldsymbol{u}_{s+1}\boldsymbol{u}_{t+1}' = 0$ whenever $s \neq t$ (because, as shown above, $\mathbb{E}\boldsymbol{r}_{s+1}\boldsymbol{r}_{t+1}' = 0$ and hence $\mathbb{E}\boldsymbol{h}_{s+1}\boldsymbol{h}_{t+1}' = 0$). Thus $\boldsymbol{u}_{t+1}$ and $\boldsymbol{u}_{s+1}$ are independent standard Normal random vectors. We have

$$\begin{aligned}\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{s+1} &= \boldsymbol{u}_{t+1}'\boldsymbol{\Omega}_t^{1/2}\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\Omega}_s^{1/2}\boldsymbol{u}_{s+1}\\ &= \boldsymbol{u}_{t+1}'\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Lambda}\boldsymbol{\Omega}_s^{1/2}\boldsymbol{u}_{s+1}.\end{aligned}$$

As $\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Lambda}\boldsymbol{\Omega}_s^{1/2}$ is a $J \times J$ diagonal matrix with ith diagonal entry $\sqrt{\zeta_{i,t}\zeta_{i,s}}$, we can write

$$\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{s+1} = \sum_{i=1}^{J}\sqrt{\zeta_{i,t}\zeta_{i,s}}w_i,$$

where w_i denotes the product of the ith entries of $\boldsymbol{u}_{t+1}$ and $\boldsymbol{u}_{s+1}$. The w_i are independent of each other, and each is the product of two independent standard Normal random variables. Therefore each w_i has zero mean and unit variance.

We wish to apply Lyapunov's version of the central limit theorem to $\sum_{i=1}^{J}\sqrt{\zeta_{i,t}\zeta_{i,s}}w_i$, which here requires that for some $\delta > 0$,

$$\lim_{NJ\to\infty}\frac{1}{s_J^{2+\delta}}\sum_{i=1}^{J}\mathbb{E}\left[\left|\sqrt{\zeta_{i,t}\zeta_{i,s}}w_i\right|^{2+\delta}\right] = 0$$

where $s_J^2 = \sum_{i=1}^{J}\zeta_{i,t}\zeta_{i,s}$.

It is enough to show that this holds when $\delta = 2$. In this case, as the fourth moment of a standard Normal random variable equals 3, we have $\mathbb{E}\left[w_i^4\right] = 9$, and so indeed

$$\frac{1}{s_J^4}\sum_{i=1}^{J}\mathbb{E}\left[\left|\sqrt{\zeta_{i,t}\zeta_{i,s}}w_i\right|^4\right] = \frac{9\sum_{i=1}^{J}\zeta_{i,t}^2\zeta_{i,s}^2}{\left(\sum_{i=1}^{J}\zeta_{i,s}\zeta_{i,t}\right)^2} \le \frac{144J}{J^2} \to 0 \quad \text{as } J\to\infty.$$

(The inequality follows because $\zeta_{i,t} \in (1,2)$ for all i and $t \ge 1$.) Hence the central limit theorem applies for $\sum_{i=1}^{J}\sqrt{\zeta_{i,t}\zeta_{i,s}}w_i$ after appropriate standardization by mean and variance, which are 0 and $\sum_{i=1}^{J}\zeta_{i,t}\zeta_{i,s}$, respectively. Thus

$$\frac{\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{s+1}}{\sqrt{\sum_{i=1}^{J}\zeta_{i,t}\zeta_{i,s}}} \xrightarrow{d} N(0,1).$$

□

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jfineco.2021.10.006.

References

Al-Najjar, N.I., 2009. Decision makers as statisticians: diversity, ambiguity, and learning. Econometrica 77, 1371–1401.

Anatolyev, S., 2012. Inference in regression models with many regressors. J. Econom. 170, 368–382.
Aragones, E., Gilboa, I., Postlewaite, A., Schmeidler, D., 2005. Fact-free learning. Am. Econ. Rev. 95, 1355–1368.
Bai, Z.D., Yin, Y.Q., 1993. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Ann. Probab. 21, 1275–1294.
Balasubramanian, A., Yang, Y., 2020. Statisticians' Equilibrium: Trading with High-Dimensional Data. Working paper. Stanford University.
Bryzgalova, S., Pelger, M., Zhu, J., 2019. Forest Through the Trees: Building Cross-Sections of Stock Returns. Working paper. Stanford University.
Calvano, E., Calzolari, G., Denicolò, V., Pastorello, S., 2018. Artificial Intelligence, Algorithmic Pricing and Collusion. Working paper. CEPR.
Campbell, J.Y., Thompson, S.B., 2008. Predicting excess stock returns out of sample: can anything beat the historical average? Rev. Financ. Stud. 21, 1509–1531.
Chan, L.K.C., Karceski, J., Lakonishok, J., 2003. The level and persistence of growth rates. J. Finance 58, 643–684.
Chinco, A., Neuhierl, A., Weber, M., 2021. Estimating the anomaly base rate. J. Financ. Econ. 140, 101–126.
Chordia, T., Goyal, A., Saretto, A., 2019. Anomalies and false rejections. Rev. Financ. Stud. 33, 2134–2179.
Cochrane, J.H., 2008. The dog that did not bark: a defense of return predictability. Rev. Financ. Stud. 21, 1533–1575.
Cochrane, J.H., 2011. Presidential address: discount rates. J. Finance 66, 1047–1108.
Collin-Dufresne, P., Johannes, M., Lochstoer, L.A., 2016. Parameter learning in general equilibrium: the asset pricing implications. Am. Econ. Rev. 106, 664–698.
DeBondt, W.F.M., Thaler, R., 1985. Does the stock market overreact? J. Finance 40, 793–805.
Dobriban, E., Wager, S., 2018. High-dimensional asymptotics of prediction: ridge regression and classification. Ann. Stat. 46, 247–279.
Fama, E., 1970. Efficient capital markets: a review of theory and empirical work. J. Finance 25, 383–417.
Feng, G., Giglio, S., Xiu, D., 2020. Taming the factor zoo: a test of new factors. J. Finance 75, 1327–1370.
Gabaix, X., 2014. A sparsity-based model of bounded rationality. Q. J. Econ. 129, 1661–1710.
Guo, W., Romano, J., 2007. A generalized Sidak-Holm procedure and control of generalized error rates under independence. Stat. Appl. Genet. Mol. Biol. 6, 1–33.
Hansen, P.R., Timmermann, A., 2015. Equivalence between out-of-sample forecast comparisons and Wald statistics. Econometrica 83, 2485–2505.
Harvey, C.R., Liu, Y., Zhu, H., 2016. ... and the cross-section of expected returns. Rev. Financ. Stud. 29, 5–68.
Heston, S.L., Sadka, R., 2008. Seasonality in the cross-section of stock returns. J. Financ Econ. 87, 418–445.
Inoue, A., Kilian, L., 2005. In-sample or out-of-sample tests of predictability: which one should we use? Econom. Rev. 23, 371–402.
Jegadeesh, N., Titman, S., 1993. Returns to buying winners and selling losers: implications for market efficiency. J. Finance 48, 65–91.
Klein, T., 2019. Autonomous Algorithmic Collusion: Q-Learning under Sequential Pricing. Working paper. University of Amsterdam.
Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. J. Financ. Econ. 135, 271–292.
Lewellen, J., Shanken, J., 2002. Learning, asset-pricing tests and market efficiency. J. Finance 57, 1113–1145.
Lindley, D.V., Smith, A.F.M., 1972. Bayes estimates for the linear model. J. R. Stat. Soc. Ser. B 34, 1–18.
Linnainmaa, J.T., Roberts, M.R., 2018. The history of the cross-section of stock returns. Rev. Financ. Stud. 31, 2606–2649.
Lo, A.W., MacKinlay, A.C., 1990. Data-snooping biases in tests of financial asset pricing models. Rev. Financ. Stud. 3, 431–467.
McLean, D.R., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Finance 71, 5–32.
Molavi, P., Tahbaz-Salehi, A., Vedolin, A., 2020. Asset Pricing with Misspecified Models. Working paper. Boston University.
Novy-Marx, R., 2012. Is momentum really momentum? J. Financ. Econ. 103, 429–453.
Shalev-Shwartz, S., Ben-David, S., 2014. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.
Sims, C.A., 2003. Implications of rational inattention. J. Monet. Econ. 50, 665–690.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Timmermann, A.G., 1993. How learning in financial markets generates excess volatility and predictability in stock prices. Q. J. Econ. 108, 1135–1145.
Yin, Y.Q., Bai, Z.D., Krishnaiah, P.R., 1988. On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Probab. Theory Relat. Fields. 78, 509–521.