NBER WORKING PAPER SERIES

MARKET EFFICIENCY IN THE AGE OF BIG DATA

Ian Martin
Stefan Nagel

Working Paper 26586
http://www.nber.org/papers/w26586

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2019

We are grateful for the comments of Svetlana Bryzgalova, John Campbell, Gene Fama, Cam Harvey, Ralph Koijen, Sendhil Mullainathan, Lubos Pastor, Andrew Patton, Andrei Shleifer, Allan Timmerman, Laura Veldkamp and seminar participants at the University of Chicago. 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. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2019 by Ian Martin and Stefan Nagel. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Market Efficiency in the Age of Big Data
Ian Martin and Stefan Nagel
NBER Working Paper No. 26586
December 2019
JEL No. C11,G12,G14

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 model economy, which resembles a typical machine learning setting, N assets have cash flows that are a linear function of J firm characteristics, but with uncertain coefficients. Risk-neutral Bayesian investors impose shrinkage (ridge regression) or sparsity (Lasso) when they estimate the J coefficients of the model and use them to price assets. When J is comparable in size to N, returns appear cross-sectionally predictable using firm characteristics to an econometrician who analyzes data from the economy ex post. A factor zoo emerges even without p-hacking and data-mining. Standard in-sample tests of market efficiency reject the no-predictability null with high probability, despite the fact that investors optimally use the information available to them in real time. In contrast, out-of-sample tests retain their economic meaning.

Ian Martin
Department of Finance
London School of Economics
Houghton Street
London WC2A 2AE
United Kingdom
i.w.martin@lse.ac.uk

Stefan Nagel
University of Chicago
Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
stefan.nagel@chicagobooth.edu

I. 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 an outside observer, but also by economic decision-makers 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. For this purpose, it is useful to model the economic actors as machine learners.

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, risk-neutral, and price stocks based on the predictive distribution 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 relationship 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 relationship.

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 zero.

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 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.

That returns are not predictable out-of-sample does not imply absence of in-sample predictability, however. An econometrician conducting an in-sample predictability test uses data that was not 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 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 we also obtain rejection probabilities close to one for finite N when J is comparable in size to N. This overwhelming rejection of the no-predictability null occurs despite the fact that investors optimally use the information available to them in real time.

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 asset such that asset returns are not predictable under their predictive distribution, an econometrician restricted to constructing return forecasts using only data that was available to investors in real time is not be 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 (i.e., the typical notion of RE in asset pricing and macroeconomics)? Or does "fully reflect" mean that investors employ Bayesian updating when they learn from data about the parameters of the cash-flow prediction model? 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 it 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 world in which investors have to take into account the valuation implications of a large number of forecasting variables. From the perspective of our model, it is not surprising that the technology-driven explosion in the number of available predictor variables has coincided with

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.

an explosion in the number of return predictors that are found significant in asset pricing studies (Cochrane 2011; Harvey, Liu, and Zhu 2016). Even without p-hacking, multiple testing, and data mining (Lo and MacKinlay 1990; Harvey, Liu, and Zhu 2016), 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). In contrast, if investors face a learning problem, there is no fixed return-generating process, and 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 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 ridge 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 the most recent 15 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.

That there was some out-of-sample predictability in the earlier parts of the sample may indicate the presence of ex-ante risk premia or mispricing at the time. It is also possible, however, that investors several decades ago were not able to process the information in each stock's price history as effectively as investors are able to do 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) and Gabaix (2014). 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. With regard to empirical studies, this means that out-of-sample predictability in early parts of the sample may reflect the fact that some of the variables that researchers can retroactively construct today were not readily available to investors at the time, or could not feasibly be included in their forecasting models.

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 or data on investor expectations), the case for a risk premium or a persistent behavioral bias is much stronger.

The insight that learning can induce in-sample return predictability in our setting relates to an earlier literature that studies time-series, rather than cross-sectional, return predictability (e.g., Timmermann 1993; Lewellen and Shanken 2002; Collin-Dufresne, Johannes, and

Lochstoer 2016). This literature examines low-dimensional settings with few return predictors, but with samples that are sufficiently short in time so that learning effects matter. In contrast, in our model, we consider the case where a large cross-section would be sufficient to learn parameters precisely in a low-dimensional setting, but learning effects persist when the number of predictors is large.

Our approach has antecedents elsewhere in the literature. Aragones, Gilboa, Postlewaite, and Schmeidler (2005) and Al-Najjar (2009) treat decision makers as statisticians that 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, Calzolari, Denicolò, and Pastorello (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.

II. Bayesian pricing in a high-dimensional setting

Consider an economy in discrete time, $t \in \{1, 2, ...\}$, 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.

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

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. And 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 if one sampled at higher frequencies over many shorter sub-periods. 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,

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

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

In a rational expectations model, for example, investors know $\boldsymbol{g}$, so that $\widetilde{\mathbb{E}}_t \left(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1} \right) = \boldsymbol{X}\boldsymbol{g}$. 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 don't 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 $\widetilde{\mathbb{E}}_t \left(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1} \right) = \boldsymbol{X}\tilde{\boldsymbol{g}}_t$, 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 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 $\widetilde{\mathbb{E}}_t \left(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1} \right) = \Delta \boldsymbol{y}_t$ 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 $\operatorname{var}\left(\boldsymbol{e}_{t+1} - \boldsymbol{e}_t \right)$, 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 $\operatorname{var}\left(\boldsymbol{X}\boldsymbol{g} + \boldsymbol{e}_{t+1} \right)$. If a relatively small component of cash-flow growth is predictable—that is, if $\operatorname{var}\left(\boldsymbol{X}\boldsymbol{g} \right) \ll \operatorname{var} \boldsymbol{e}_{t+1}$—then the random walk forecast MSE may be substantially lower than the OLS forecast MSE.

II.A. 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 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*

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

That prior beliefs are centered around zero means that investors a priori don't know which characteristics predict cash-flow growth 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.

Combined with the data-generating process (1), this setup maps into the Bayesian linear model of Lindley and Smith (1972). 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

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

The posterior mean $\tilde{\boldsymbol{g}}_1$ takes the form of a Tikhonov-regularized regression estimator, where the inverse of $\boldsymbol{D}_1^{-1}$ 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 (see, for example, Shalev-Shwartz and Ben-David (2014)).[2]

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

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

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 $\tilde{\boldsymbol{g}}_{GLS,1}$.

III. 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.

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 uncorrelated.

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 $\boldsymbol{Xg}$ is $\boldsymbol{X}\boldsymbol{\Sigma}_g\boldsymbol{X}'$, so the cross-sectional average prior

3. See, e.g., 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.

variance of the predictable component of cash-flow growth satisfies

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

using Assumptions 1 and 4.

We view the matrix of characteristics, $\boldsymbol{X}$, as observable and form the eigendecomposition

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

where $\boldsymbol{\Lambda}$ is diagonal with the eigenvalues λ_j of $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$ on its diagonal and $\boldsymbol{Q}$ is an orthogonal matrix (i.e., $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I}$) whose columns are the corresponding eigenvectors of $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$. This is possible because $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$ is symmetric. As it is also positive definite, the eigenvalues satisfy $\lambda_j > 0$ for all j. Lastly, the normalization $\operatorname{tr}\boldsymbol{X}'\boldsymbol{X} = NJ$ (Assumption 1) implies that the average eigenvalue equals one:

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

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

If $\boldsymbol{X}'\boldsymbol{X}$ has eigenvalues that are very close to zero, then the columns of $\boldsymbol{X}$ are roughly collinear. For, 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—call it $\lambda_{\min}$—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. Our next assumption can therefore be thought of as ensuring that we really are in a Big Data environment.

Assumption 5. *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 $N \to \infty$.*

In fact all we need for our main results to go through is that λ_j does not tend to zero

at a rate faster than $1/\sqrt{j}$. Assumption 5 is therefore stronger than we need, but it will be satisfied in our applications below.

Combining Assumption 4 with equation (2), we obtain

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

where

$$\boldsymbol{\Gamma}_t = \boldsymbol{Q}\left(\boldsymbol{I} + \frac{J}{N\theta t}\boldsymbol{\Lambda}^{-1}\right)^{-1}\boldsymbol{Q}' \tag{5}$$

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 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.

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

$$\boldsymbol{r}_{t+1} = \boldsymbol{y}_{t+1} - \boldsymbol{p}_t = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{X}\boldsymbol{\Gamma}_t(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t + \boldsymbol{e}_{t+1} \tag{6}$$

where $\bar{\boldsymbol{e}}_t = \frac{1}{t}\sum_{s=1}^{t}\boldsymbol{e}_s$.

Realized returns thus have three components. The first term on the right-hand side of (6) 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 that an econometrician would study, it is not zero.

The second term represents the effect of noise on investors' posterior mean. To the extent 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 components correlated with the columns of $\boldsymbol{X}$. As we show now, these components may induce certain forms of return predictability.

III.A. 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 $\boldsymbol{X}$, using OLS.[4] The econometrician obtains a vector of cross-sectional regression coefficients

$$\begin{aligned}\boldsymbol{h}_{t+1} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{r}_{t+1} \\ &= (\boldsymbol{I}-\boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{\Gamma}_t\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_t + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}_{t+1}. \end{aligned} \tag{7}$$

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 don't show formal results for this case, but we have explored the issue in simulations.

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. Under this hypothesis, the first two terms in (7) would be zero. Given that the elements of $\boldsymbol{e}_{t+1}$ are distributed $N(0,1)$, it would follow, under this null, that

$$\sqrt{N}\boldsymbol{h}_{t+1} \sim N\left(0, N(\boldsymbol{X}'\boldsymbol{X})^{-1}\right), \tag{8}$$

using the usual OLS asymptotic variance formulas.[5] In deriving these properties, the econometrician conditions on the observed predictors $\boldsymbol{X}$. From (8) it would follow—again, under the econometrician's rational expectations null—that

$$\boldsymbol{h}_{t+1}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{h}_{t+1} \sim \chi^2_J. \tag{9}$$

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 if we let the econometrician consider a scaled version of this test statistic:

$$T_{re} \equiv \frac{\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} - J}{\sqrt{2J}}. \tag{10}$$

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

$$T_{re} \xrightarrow{d} N(0,1) \quad \text{as } N, J \to \infty,\ J/N \to \psi > 0. \tag{11}$$

But the actual asymptotic distribution of T_{re} is influenced by the terms involving $\bar{\boldsymbol{e}}_t$ and $\boldsymbol{g}$ in (7). These alter the asymptotic distribution and may lead the rejection probabilities of a test using $N(0,1)$ critical values based on (11), or χ^2 critical values based on (9), to differ

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 asymptotics.

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 equation (6). 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 deviated 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 (6), the rational expectations econometrician will use inappropriately small standard errors, in the sense that $\boldsymbol{\Sigma}_{b} - \boldsymbol{\Sigma}_{re}$, $\boldsymbol{\Sigma}_{b}\boldsymbol{\Sigma}_{re}^{-1} - \boldsymbol{I}$, and $\boldsymbol{\Sigma}_{re}^{-1} - \boldsymbol{\Sigma}_{b}^{-1}$ are all positive definite.[6]

Our first result shows that the first two cross-sectional moments of the eigenvalues of $\boldsymbol{\Sigma}_{b}\boldsymbol{\Sigma}_{re}^{-1}$ characterize the asymptotic behavior of T_{re}. These eigenvalues (ζ_j) can be written explicitly in terms of the eigenvalues (λ_j) of $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X}$ as

$$\zeta_j = 1 + \frac{\lambda_j}{t\lambda_j + \frac{\psi}{\theta}}\,. \tag{12}$$

As $\lambda_j > 0$ and $\psi/\theta > 0$, we have $1 < \zeta_j < 2$ for all $t \geq 1$. We write the limiting mean and

6. For example, in the notation of the proof of Proposition 1, we can write $\boldsymbol{\Sigma}_{b}\boldsymbol{\Sigma}_{re}^{-1} - \boldsymbol{I} = N\,\mathbb{E}\left[\boldsymbol{h}_{t+1}\boldsymbol{h}_{t+1}'\right]\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{I} = \boldsymbol{Q}\boldsymbol{\Omega}\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}' - \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{Q}\left(\boldsymbol{\Omega}\boldsymbol{\Lambda} - \boldsymbol{I}\right)\boldsymbol{Q}'$. Thus $\boldsymbol{\Sigma}_{b}\boldsymbol{\Sigma}_{re}^{-1} - \boldsymbol{I}$ is symmetric and its eigenvalues are all positive (as the diagonal matrix $\boldsymbol{\Omega}\boldsymbol{\Lambda}$ has diagonal entries that are greater than one, as shown in the proof of Proposition 1). It is therefore positive definite.

variance of the eigenvalues as

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

By the "Big Data" Assumption 5, we have $\mu \in (1,2)$ and $\sqrt{\mu^2+\sigma^2} \in (1,2)$ for all $t \geq 1$. Without it, we could have $\mu = 1$ and $\sigma = 0$ if $\lambda_j \to 0$.

Proposition 1. *If returns are generated according to* (6) *then in the large* N, J *limit*

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

Thus the rational expectations case (11) *is analogous to assuming* $\zeta_j = 1$ *for all* j*.*

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 $\sqrt{J}$:

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

(For comparison, the rational expectations econometrician thinks T_{re} is asymptotically standard Normal, as in equation (11).) Thus the rejection probability tends rapidly to one as N and J tend to infinity.

Proposition 2. *In a test of return predictability based on the rational expectations null* (11)*, we would have, for any critical value* c_α *and at any time* t*,*

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

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:

$$\lim_{N\to\infty} -\frac{1}{N}\log\mathbb{P}\left(T_{re} < c_\alpha\right) = \frac{(\mu-1)^2\psi}{4\left(\mu^2+\sigma^2\right)}, \tag{13}$$

for any fixed critical value c_α.

Thus in-sample predictability tests lose their economic meaning when J is not small relative to N, in the sense that the usual interpretation of in-sample return predictability evidence is not warranted.[7] 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.

Another way to examine in-sample predictability is to consider a trading strategy with weights proportional to in-sample predicted returns, $\boldsymbol{X}\boldsymbol{h}_{t+1}$, according to the econometrician's regression coefficient estimates:

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

We scale the predicted returns with $1/N$ so that if the distribution of predicted returns is kept fixed (normal, centered at zero, and fixed standard deviation), the expected sum of absolute portfolio weights remains constant as $N \to \infty$. This in turn means that, in expectation, the size of the combined long and short positions is constant as $N \to \infty$. To the extent that the expected portfolio return changes when $N, J \to \infty$, this would have to come from a change

7. In the conventional low-dimensional asymptotics case, with J fixed as $N \to \infty$, the test rejects more often that it would under rational expectations, albeit to a small degree. In this case $\boldsymbol{\Gamma}_t \to 1$ so that the first term vanishes, but the second term in (7) is still relevant. This term shrinks with N at the same speed as the third term (which would be the only one under the rational expectations null), and leads to excessive rejection rates. Predictability vanishes in economic terms, however, as the return of the trading strategy that we will consider in Proposition 3 converges to zero if J is held fixed as $N \to \infty$.

in the distribution of predicted returns as J grows rather than a mechanical effect due to greater N.

As we now show, however, the expected return converges to a constant asymptotically. More precisely, the objective expected return that an econometrician estimates by sampling repeatedly from this economy converges to a constant. In contrast, the Sharpe ratio explodes.

Proposition 3. *If returns are generated according to* (6)*, then an in-sample trading strategy with weights* $\boldsymbol{w}_{IS,t} = \frac{1}{N}\boldsymbol{X}\boldsymbol{h}_{t+1}$ *and returns* $r_{IS,t+1} = \boldsymbol{r}'_{t+1}\boldsymbol{w}_{IS,t}$ *has expected returns that converge to a constant,*

$$\lim_{N,J\to\infty,J/N\to\psi} \mathbb{E}\, r_{IS,t+1} = \psi\mu > 0\,,$$

and a Sharpe ratio $SR_{IS} = \mathbb{E}\, r_{IS,t+1}/\operatorname{var}(r_{IS,t+1})^{1/2}$ *that grows at rate* $\sqrt{N}$*,*

$$\lim_{N,J\to\infty,J/N\to\psi} \frac{SR_{IS}}{\sqrt{N}} = \frac{\mu\sqrt{\psi}}{\sqrt{2\left(\mu^2+\sigma^2\right)}}\,.$$

The expected return of the in-sample trading strategy also represents the amount of in-sample predictable return variation in the cross-section because $\mathbb{E}[\boldsymbol{r}'_{t+1}\boldsymbol{w}_{IS,t}] = \frac{1}{N}\,\mathbb{E}[\boldsymbol{h}'_{t+1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1}]$, i.e., the average portfolio return is equal to the average squared predicted returns. Our assumptions in (4) ensure—as is economically reasonable—that the amount of predictable variation in dividend growth stays bounded as N and J tend to infinity. As a result, the amount of predictable return variation approaches a finite limit. But the variance of the portfolio return shrinks with N and J, so that in the high-dimensional case in-sample moments seem to suggest incredible investment opportunities in mean-variance terms. These in-sample results do not represent returns of trading strategies that are implementable in real time, however.

III.B. (Absence of) out-of-sample return predictability

The situation looks very different with regards to out-of-sample predictability. Here we consider a trading strategy with weights

$$\boldsymbol{w}_{OOS,t} = \frac{1}{N}\boldsymbol{X}\boldsymbol{h}_t \tag{14}$$

that is implementable at the end of period t based on the econometrician's regression coefficient estimates at that time. Proposition 4 shows that the econometrician, on average, does not find any out-of-sample predictability.

Proposition 4. *If returns are generated according to* (6)*, then an out-of-sample trading strategy that uses coefficients estimated at time* $s+1$*, and hence has portfolio weights* $\boldsymbol{w}_{OOS,s+1} = \frac{1}{N}\boldsymbol{X}\boldsymbol{h}_{s+1}$*, has expected return* $\mathbb{E}\, r_{OOS,t+1} = 0$ *whenever* $t \neq s$*.*

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 is available at the time of making the prediction.

That the result also applies backwards in time, with $t < s$, is more surprising. This case does not represent a tradable strategy, but it is interesting from an econometrician's perspective. 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 that 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). It is an interesting question for future research to what extent one can generalize the backwards result.

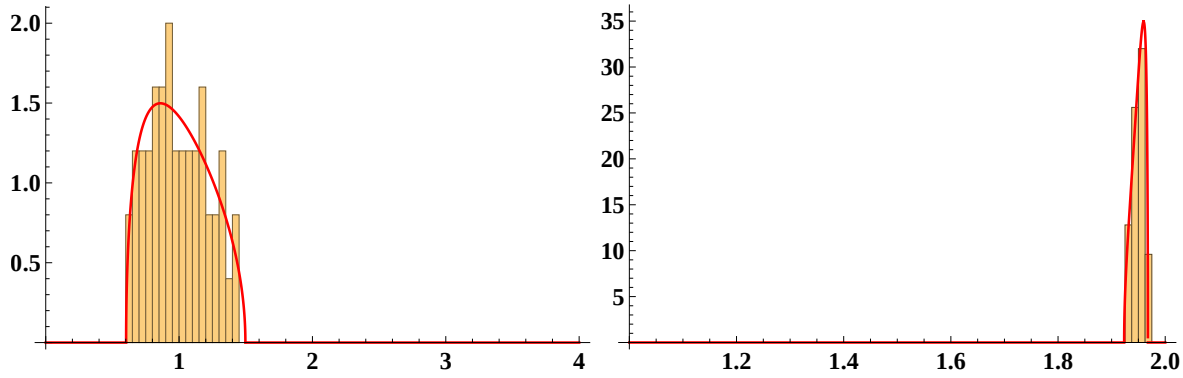
III.C. An example

To see explicitly how these results play out in a concrete example, suppose that characteristics are determined at random, so that the matrix of firm characteristics $\boldsymbol{X}$ has IID entries x_{ij} with mean zero, unit variance, and finite fourth moment. 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 re-running the economy. 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.

Moreover, 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, Bai, and Krishnaiah (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.

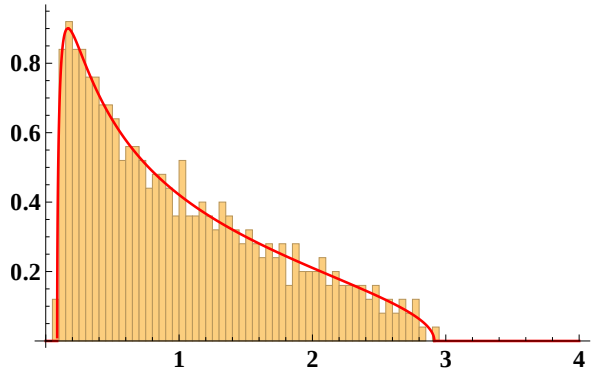
Figure I shows histograms of the eigenvalue distributions in examples with $\boldsymbol{X}$ drawn randomly with $N(0,1)$ entries, setting $N = 1000$ and $J = 10$, 100, 500, and 900. Solid red lines in the figures illustrate the limiting Marchenko–Pastur distribution for the eigenvalues λ_j in each case; we also calculate the corresponding asymptotic distribution of the eigenvalues ζ_j by change of variable, using equation (21) in the appendix. When $\psi = J/N$ is close to one there is considerable mass near zero, implying that there are many approximately collinear relationships 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 ζ_j.

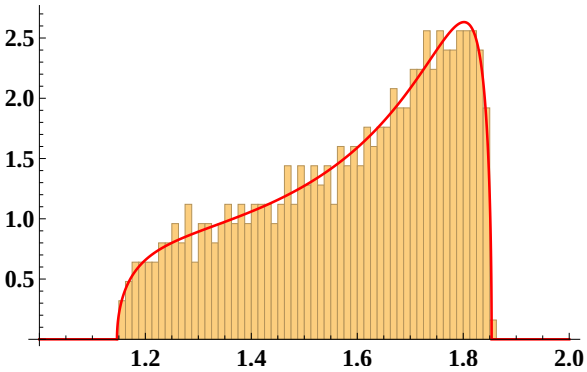


(A) Histogram of λ_j, $J/N = 0.05$.

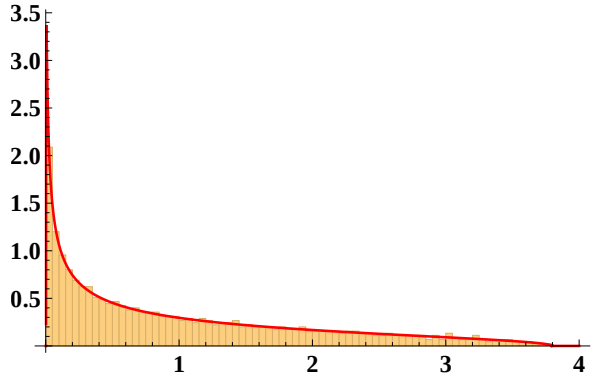
(B) Histogram of ζ_j, $J/N = 0.05$.



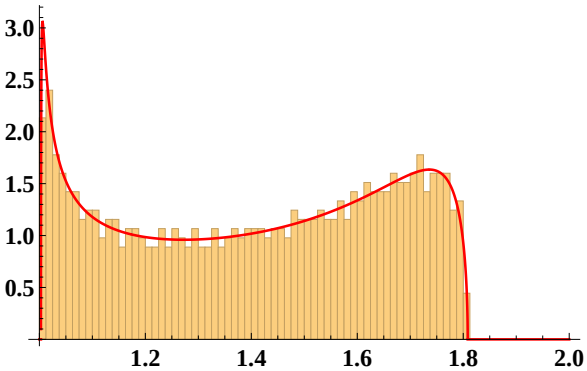
(C) Histogram of λ_j, $J/N = 0.5$.



(D) Histogram of ζ_j, $J/N = 0.5$.



(E) Histogram of λ_j, $J/N = 0.9$.



(F) Histogram of ζ_j, $J/N = 0.9$.

FIGURE I

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.

Proposition 5. *The cross-sectional moments of ζ_j satisfy*

$$\mu = 1 + \frac{\psi + \theta t(\psi+1) - \sqrt{\left[\psi + \theta t(\psi+1)\right]^2 - 4\theta^2 t^2 \psi}}{2\theta t^2 \psi} \tag{15}$$

and

$$\sigma^2 = \frac{\theta^2 t^2 \psi - (\theta t + \psi)^2}{2\theta^2 t^4 \psi^2} + \frac{\theta t \psi \left(\theta^2 t^2 (\psi - 2) - \theta t \psi + \psi^2\right) + (\theta t + \psi)^3}{2\theta^2 t^4 \psi^2 \sqrt{\left[\psi + \theta t(\psi+1)\right]^2 - 4\theta^2 t^2 \psi}}\,. \tag{16}$$

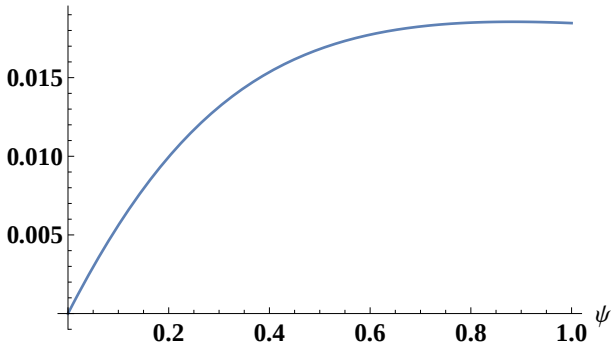
This result allows explicit calculation of the limit moments that appear in Propositions 1, 2, and 3. For example, Proposition 2 shows that the probability of rejecting the null of no predictability declines exponentially fast as N increases, and derives the rate of the exponential decay. Similarly, Proposition 3 shows that the expected return on a portfolio that uses in-sample information to construct portfolio weights is $\mathbb{E}\, r_{IS,t+1} = \psi\mu$. In both cases, Proposition 5 can be used to derive exact analytical expressions in terms of the model primitives θ, ψ, and t.

Figure II shows how the resulting expressions depend on ψ for $\theta = 1$ and $t = 1$. Panel (a) plots the rate function (13). 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, $N \geq 300$. Panel (b) plots the expected in-sample portfolio return, $\mathbb{E}\, r_{IS,t+1} = \psi\mu$, from Proposition 3. As we noted earlier, $r_{IS,t+1}$ also represents the average squared in-sample predicted return. For comparison, the average squared in-sample unpredictable return is less than $\frac{1}{N}\boldsymbol{e}'_{t+1}\boldsymbol{e}_{t+1} \approx 1.0$. Therefore, the two panels show that for large ψ there is substantial in-sample predictability in, respectively, statistical and economic terms.

IV. Finite-sample analysis: Simulations

In this section, we report the results of finite-sample simulations. The high-dimensional asymptotics, with $N, J \to \infty$ and $J/N \to \psi$, are intended to provide an approximation for

(A) The rate function given in equation (13)



(B) Expected return of in-sample trading strategy

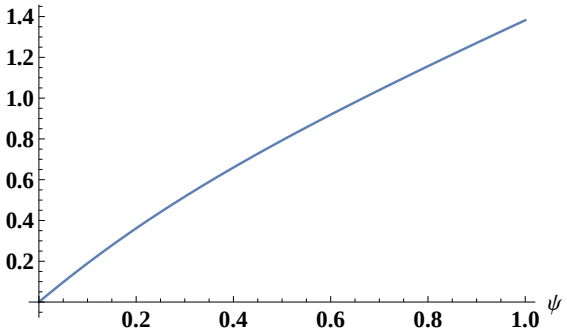


FIGURE II
In-sample return predictability in the asymptotic limit with $\theta = 1$ and $t = 1$, for different values of ψ.

the properties of return predictability tests in the case where J is not small relative to N (just as conventional large N-fixed J asymptotics provide an approximation for the small J/N case). Finite-sample simulations provide some insight to what extent 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 this 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 RHS) is constant w.r.t. T, while the variance of the residual component (second term on the RHS) 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

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

and 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

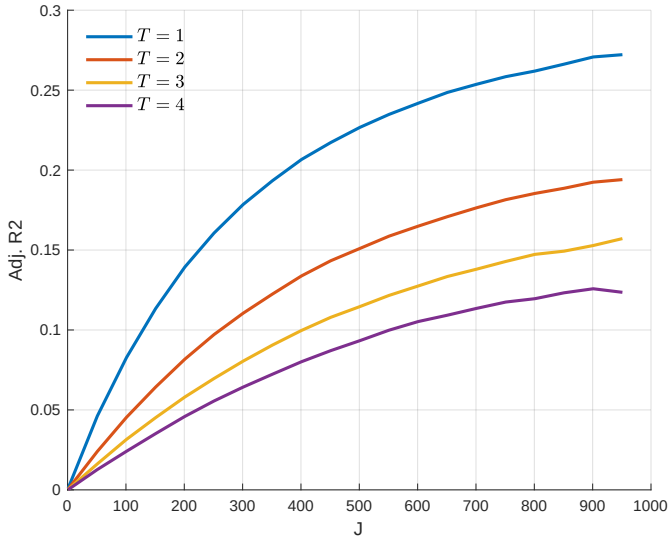
for various revenue and profit measures in Chan, Karceski, and Lakonishok (2003). When stocks are sorted based on IBES analysts forecasts (their Table IX), Chan et al. find an inter-decile spread of slightly around 10 percentage points (pp) for annualized growth rates over the next one, three, and five years. Extrapolating to 10 years, we would have 10 pp also at a 10-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 inter-decile 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 $\theta = 1$.

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 that 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 IIIa presents the (in-sample) adjusted R^2 from these regressions. As J increases towards N, the adjusted R^2 also increases, and hence returns become more predictable in sample. If investors have learned for more periods, return predicability gets weaker.

Figure IIIb 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 (8). The dotted line shows the nominal size of 5% that the test would have,

8. For this analysis, they don't report percentiles at sufficient detail to calculate the inter-decile spread, but they report means for quintile bins, and the spread between top and bottom quintile bin mean should correspond, approximately to the inter-decile range.

(A) Adjusted R^2



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

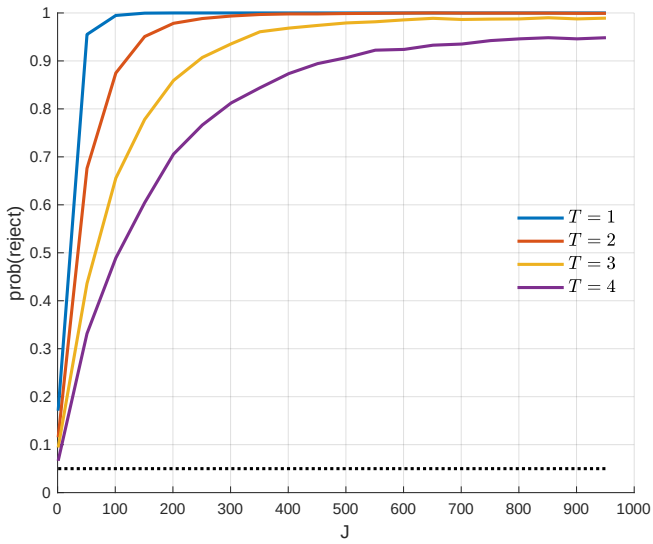


FIGURE III
In-sample return predictability tests

Based on cross-sectional regression 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. 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 $\boldsymbol{g}$). Actual size shows the rejection probability when this test is applied in setting where Bayesian investors with an objectively correct informative prior estimate $\boldsymbol{g}$.

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 increase 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.

V. 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, (i) investors' prior is that the elements of $\boldsymbol{g}$ are drawn from a Laplace distribution; and (ii) investors price assets based on the mode rather than the mean of the posterior distribution (i.e., a maximum-a-posteriori, or MAP 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 IV shows that the results in the Laplace prior case are extremely similar to those in Figure III 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.

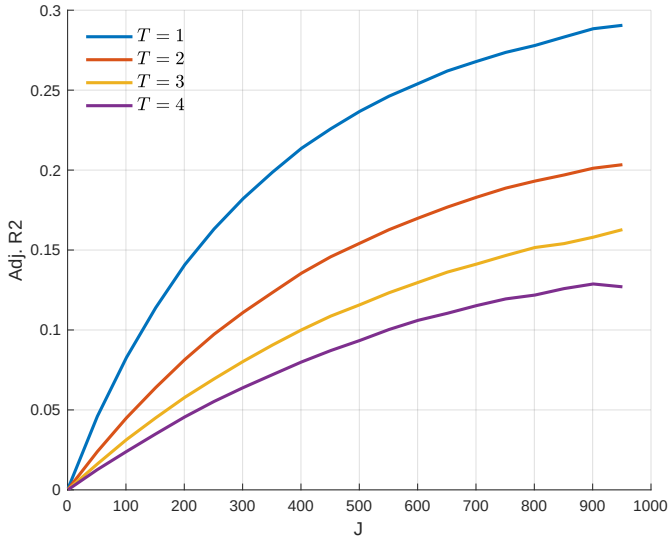
VI. Excess shrinkage or sparsity

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 V 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 is generated, as before, with $\theta = 1$. However, investors' prior beliefs are now based on a value of θ that is different. In the excessive shrinkage and sparsity cases, we let investors form beliefs based on

(A) Adjusted R^2



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

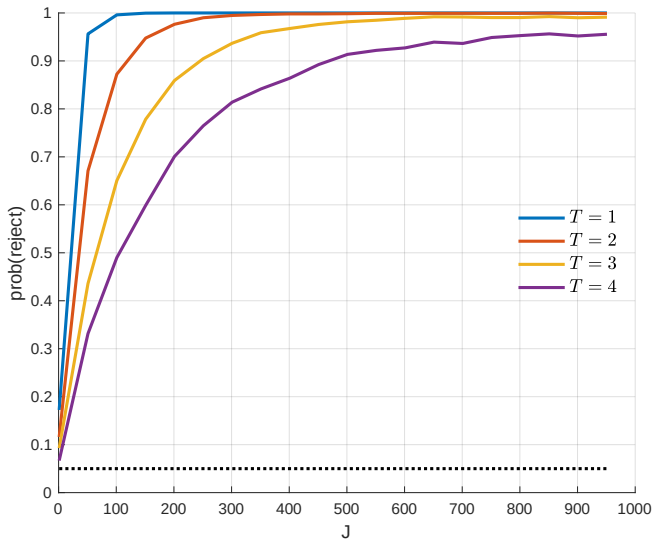
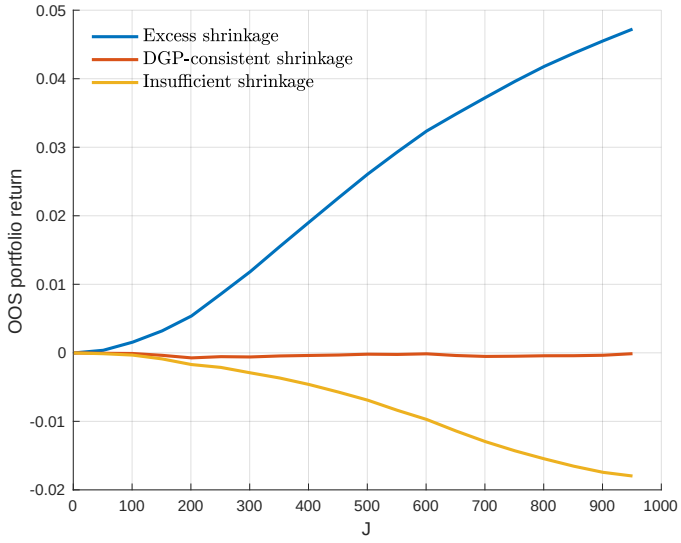


FIGURE IV
Lasso: In-sample return predictability tests

Based on cross-sectional regression 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. 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 $\boldsymbol{g}$). Actual size shows the rejection probability when this test is applied in setting where investors estimate $\boldsymbol{g}$ (which is drawn from a Laplace distribution) with Lasso.

(A) Ridge



(B) Lasso

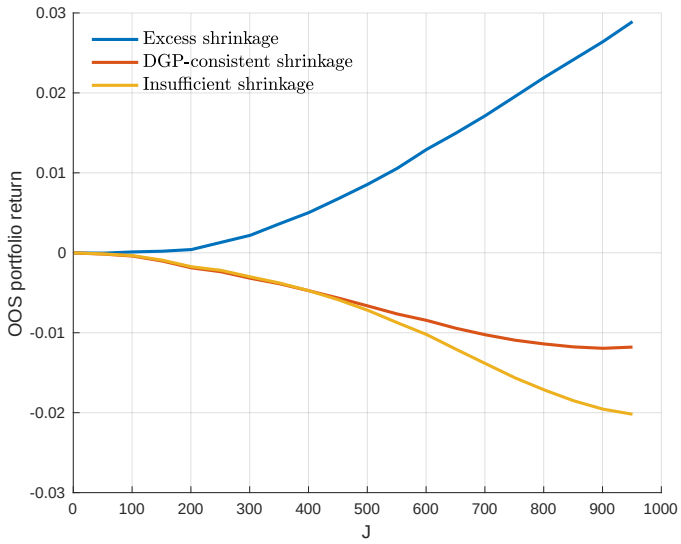


FIGURE V
Out-of-sample portfolio returns when investors apply excess shrinkage or sparsity

Based on cross-sectional regression 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 $\theta = 1$.

$\theta = 0.5$. This 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 (14). Panel (a) shows the results in the normal prior/ridge regression case. As expected from Proposition 4, the average OOS return in the DGP-consistent case is zero. In contrast, when shrinkage is excessive, investors end up downweighting too much pieces of information in $\boldsymbol{X}$ that predict cash flows. As a result, an econometrician sampling returns from this economy is able to forecast returns OOS. 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 OOS average return that differs 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 (6) plays a bigger role in 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 (7) 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 OOS 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, by assumption, 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 take-away from all this is that OOS 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 OOS-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 OOS predictability.[9]

This analysis also provides a perspective on the likely effects of technological progress in data construction and data analysis on return predictability observed in empirical analyses. Many studies of the cross-section of stock returns use data that goes back to time periods in which data availability and analysis was much more constrained than it is 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 at the time when they priced stocks several decades ago may have been excessively sparse relative to the model 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-) OOS tests in which the researcher re-constructs investors' learning process, but without taking into account the additional model complexity constraints that investors faced in real time.

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

VII. Empirical Application: Predicting Stock Returns with Past Returns

To illustrate how our model provides an interpretation of in-sample and out-of-sample stock return predictability evidence, 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 IS and OOS predictability. In the model, OOS predictability is zero, but the model abstracted from risk premia and (except for our analysis of excess shrinkage) from behavioral biases and bounded rationality that could induce OOS predictability. For this reason, in our empirical analysis we focus on the wedge between IS and OOS predictability, and its evolution over time, rather than on the absolute level of predictability.

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, all the way back to the start of the CRSP database that we will use for our analysis. Many predictors based on accounting variables do not satisfy this criterion because they become 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 search and multiple testing, we do not want a set of predictors that may already be the product of past 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 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 next month's stock return.

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 at all 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 OOS return predictability in parts of the sample where technological constraints may have prevented them from doing so—consistent with our excess shrinkage results in the previous section.

Focusing on price-history-based predictors, without pre-selecting particular subsets of those based on earlier evidence of predictability, allows us to side-step, 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 don't 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 IS and OOS predictability.

We use all U.S. stocks in the CRSP database except small stocks that have market capitalization below the 20th NYSE percentile and 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

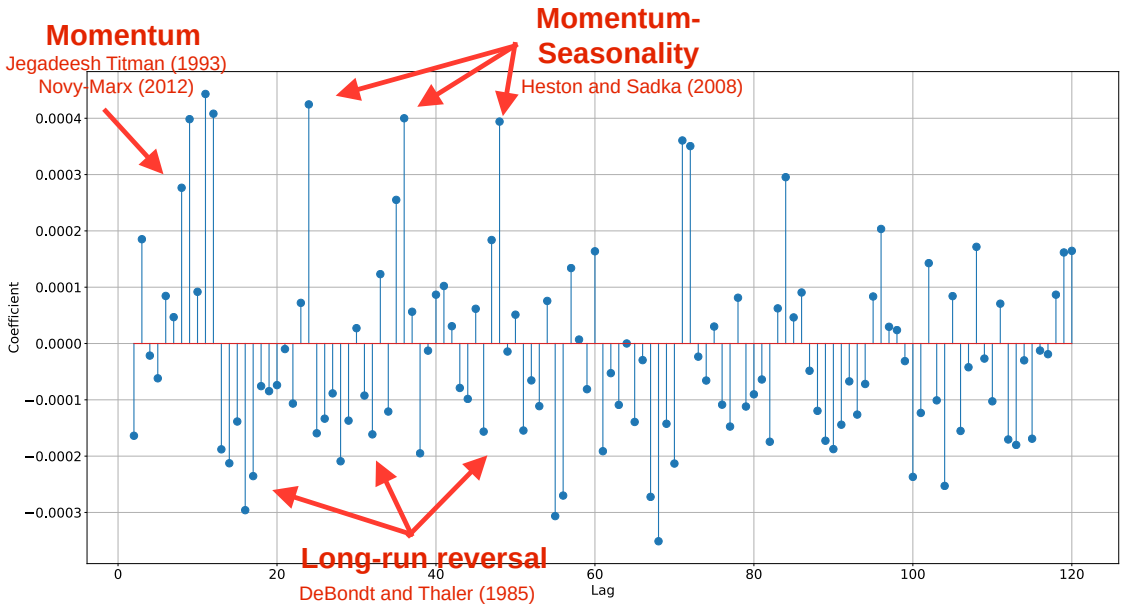
in month t in a panel regression. We demean the dependent variable and all explanatory variables month by month to focus purely on cross-sectional variation. In addition, we cross-sectionally standardize all predictor variables to unit standard deviation each month. We weight the observations each month 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 Figure VI 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 7 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. The monthly in-sample R^2 in this ridge regression is 0.23%. With market-adjusted annualized volatility of

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 and 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.

(A) Coefficients for past returns



(B) Coefficients for past squared returns

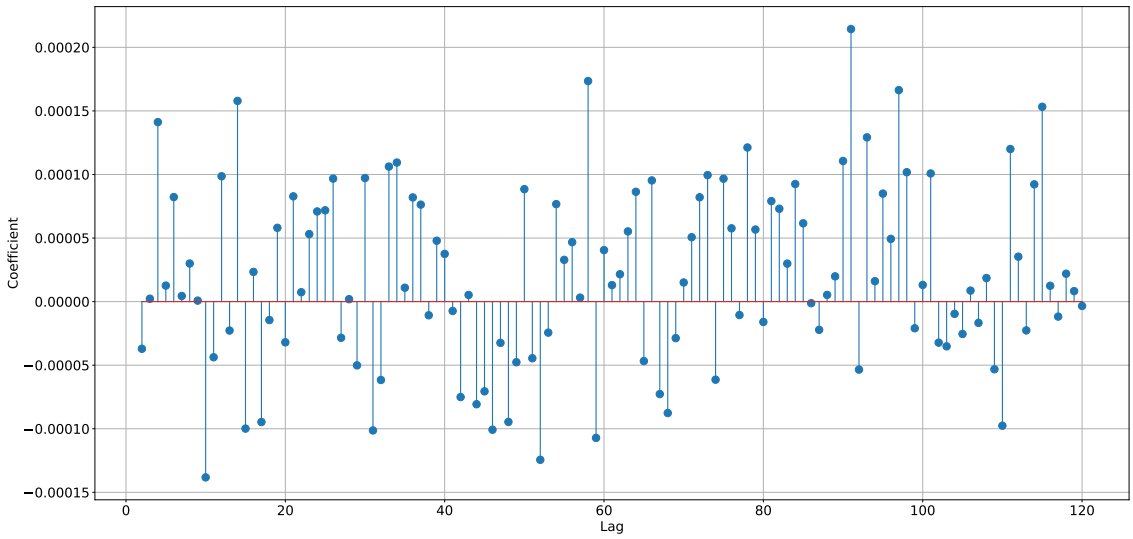


FIGURE VI
Ridge Regression Coefficient Estimates

roughly 50% for the dependent variable, this R^2 implies a cross-sectional annualized standard deviation of fitted in-sample expected returns of approximately 8%, which is quite big.

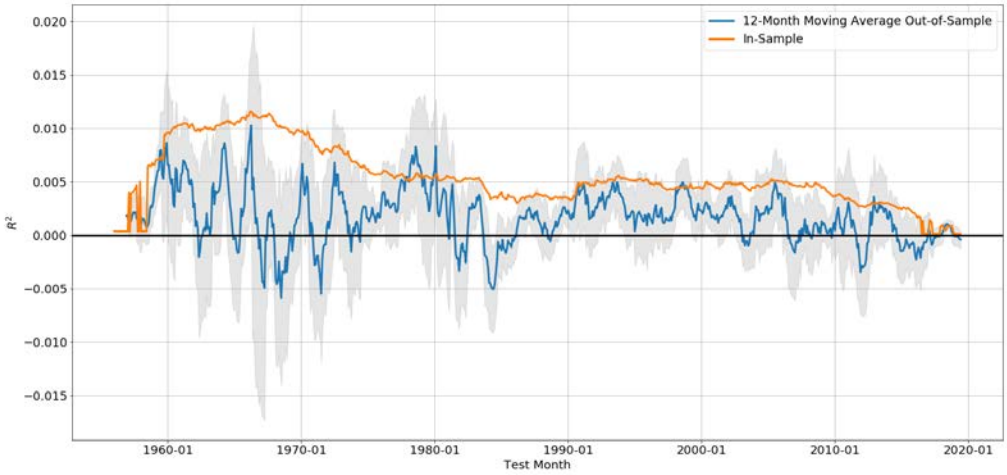
However, as Figure VII shows, in terms of OOS predictability, the picture looks very different. This figure is based on rolling regressions with 20-year estimation windows. We use the entire CRSP data set starting in 1926, subject to the same NYSE size cutoff and minimum lagged price requirement as in Figure VI. We choose the penalty parameter using leave-one-year-out cross-validation in each 20-year window. Based on the coefficient estimates from each window, we then forecast returns in the first month after the estimation window ends. 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 record the OOS R^2 in this month and then move the 20-year window forward by one month to repeat the process.

In panel (a) of Figure VII, we show the time-series of these OOS R^2 in the form of a 12-month moving average. For comparison, we also show, for each month t, the in-sample R^2 from the estimation window ending in month t. As the figure shows, the OOS R^2 is almost everywhere smaller than the IS R^2.

Panel (b) looks at the IS and OOS returns, as 12-month moving average, of a portfolio strategy that weights individual stocks with their predicted returns based on the rolling ridge regression estimates, akin to the portfolio returns we analyzed in Propositions 3 and 4, respectively. This is a zero-investment long-short portfolio that goes long in stocks with positive predicted returns and short in stocks with negative predicted returns. One difference between this portfolio return and the R^2 is that the portfolio return depends, approximately, only on the cross-sectional covariance between predicted returns and realized returns. Idiosyncratic prediction noise that is uncorrelated with the realized returns diversifies away in the portfolio. In this sense, the portfolio return does not penalize as much for estimation error as the R^2 measures do.[11] The in-sample portfolio return is very stable across time and

11. For example, if the predicted return was pure noise, the R^2 would be negative, and more so the greater the variance of the noise, while the portfolio return would be zero.

(A) In-Sample and Out-of-Sample R^2



(B) In-Sample and Out-of-Sample Portfolio Return

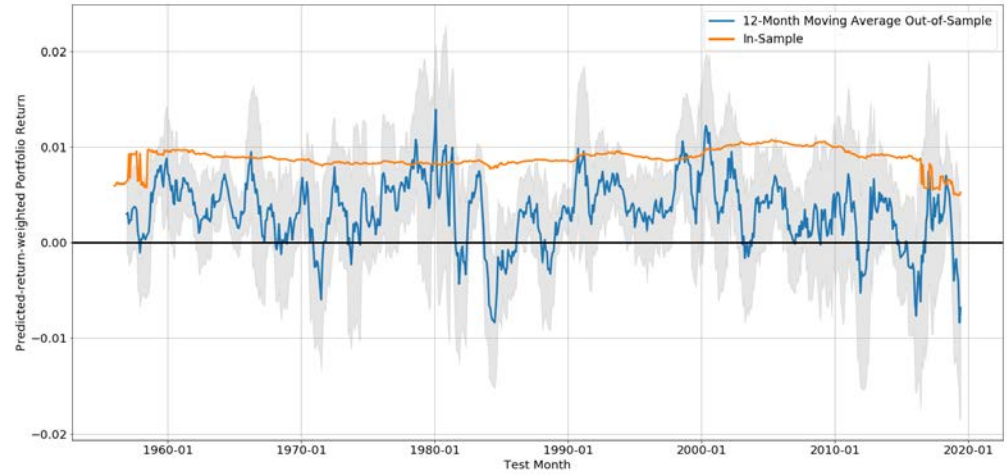


FIGURE VII
In-Sample and Out-of-Sample Return Predictability

Panel (a) shows the 12-month moving averages of in- and out-of-sample R^2 from 20-year rolling window ridge regressions of individual stock returns on past returns and past squared returns, with penalty chosen, for each window, with leave-one-year-out cross-validation. The out-of-sample prediction is for the first month after the regression window ends. Panel (b) shows the in-sample and out-of-sample returns of a portfolio strategy that weights stocks by their predicted returns. Stocks with market capitalization below the 20th NYSE percentile and price lower than one dollar at the end of the prior month are excluded everywhere.

TABLE I

IN-SAMPLE AND OUT-OF-SAMPLE R^2: SUMMARY STATISTICS

Columns (i) and (ii) in this table reports summary statistics for the time series of IS and OOS ridge regression R^2 and portfolio returns shown in Figure VII. For comparison, we also show the statistics for IS and OOS R^2 of OLS regressions. Column (iii) reports backwards OOS test where the coefficients estimated in the estimation window are applied backwards to predict the return in the month prior to the start of the estimation window. The numbers are reported in percent per month.

		In-Sample (i)	Forward Out-of-Sample (ii)	Backward Out-of-Sample (iii)
		Panel A: R^2		
Ridge	Mean	0.54	0.17	0.25
	S.D.	0.28	0.78	0.96
OLS	Mean	0.90	-0.25	-0.17
	S.D.	0.27	1.74	2.01
		Panel B: Portfolio return		
Ridge	Mean	0.88	0.32	0.31
	S.D.	0.10	1.14	1.43
OLS	Mean	0.94	0.32	0.33
	S.D.	0.08	1.02	1.20

reliably above zero. In contrast, the out-of-sample return is much lower and frequently close to zero or below.

Table I reports the summary statistics for these time series. In addition to the statistics for the forward OOS series from Figure VII shown in the second column, column (iii) also presents statistics for a backward OOS series. This backwards OOS analysis is motivated by the prediction in Proposition 4 that absence of OOS predictability should not only hold forward in time, but also backwards. We compute this backward OOS series in the same way as the forward OOS series, with the only difference being that we start at the end of the sample with a 20-year estimation sample from July 1999 to June 2019 and then we use the coefficient estimates from this estimation window, applied to returns and squared returns from June 1989 to March 1999, to predict returns in May 1999. Then we move the estimation window backwards by one month and repeat the process.

Panel A shows that while the mean IS R^2 is 0.54%, the forward OOS R^2 is less than a third of this magnitude. The second set of rows in Panel A shows that if one runs OLS regressions rather than ridge regressions, the difference between IS and forward OOS R^2 is even bigger: 0.90% IS vs. -0.25% OOS.

Panel B shows the summary statistics of the return on the predicted-return-weighted portfolio. Here, too, there is a strong degradation from IS to forward OOS performance.[12] Unlike for the R^2, the results are very similar for ridge regression and OLS. This reflects the fact that estimation-error-induced noise in predicted returns lowers the R^2 (which is why the R^2 can be negative), but it diversifies away in the portfolio return.[13]

The third column in Table I shows statistics for the backwards OOS analysis. They are fairly close to the forward OOS numbers in column (ii). Like in our model, there doesn't seem to be much difference between forward and backwards OOS predictability.

Overall, the results illustrate in a relevant empirical cross-sectional asset pricing setting that there can be a big wedge between IS and OOS predictability. This underscores the message from our model that IS cross-sectional return predictability evidence is not a good motivation for seeking risk-based or behavioral economic explanations. Much of the IS predictability here does not carry over into OOS predictability and hence does not reflect risk premia demanded by investors ex-ante or persistent belief distortions.

There are a number of concerns one might have about this interpretation. First, wouldn't it 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-) OOS 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

12. Market betas of these portfolio returns are very close to zero and hence the degradation in terms of abnormal returns relative to the CAPM is essentially the same as for the mean returns shown in the table.

13. The fact that the greater noise in the OLS estimates diversifies away in the portfolio return also underscores that the assumption in our model that the econometrician uses OLS to run cross-sectional predictability regressions (rather than ridge or lasso) is innocuous.

example, consider that the literature on momentum has shifted from emphasizing momentum based on returns in months $t-2$ to $t-12$, as in Jegadeesh and Titman (1993), towards highlighting returns in months $t-7$ to $t-12$ as those that really matter for predicting returns (Novy-Marx 2012). It's 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, OOS tests may have low power to detect return predictability. This point has been made by Inoue and Kilian (2005), Campbell and Thompson (2008), and Cochrane (2008) in a time-series setting. However, their arguments are based on a rational expectations setting in which investors know the parameters of the data-generating process and only the econometrician faces the problem of recovering the parameters of a fixed model from observed data. In this case, IS and OOS methods test the same economic hypothesis. Under the alternative hypothesis that there is predictability, such a fixed model implies that predictability should be there IS and OOS—and since IS methods are more powerful by virtue of using the full sample of available data, it seems natural to prefer IS methods.

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 underlying model that an econometrician could recover. Instead, the properties of the data are evolving over time as investors learn. As a consequence, IS and OOS methods test different economic hypotheses. IS predictability tests basically lose their economic meaning because they cannot discriminate between predictability induced by learning and predictability induced by risk premia or behavioral biases. Only OOS tests can do so. For this reason, even if OOS tests have low power, IS tests are simply not a viable alternative method because they test a different, largely uninteresting hypothesis without clear economic interpretation.

Turning back to Figure VI, another interesting fact is that the OOS R^2 has come down

over time. In the last 15 years of the sample, it was basically zero on average. This further underscores the point that it may not make much sense to regard the apparent anomalies that show up so nicely in the in-sample evidence in Figure VI as facts about the cross-section of expected stock returns that require a risk-based or behavioral economic explanation.

That the OOS R^2 tended to be above zero in earlier decades could be an indication that investors' ability to simultaneously digest information from large numbers of predictor variables was more constrained than it is today. As we discussed in Section VI, such constraints can lead to out-of-sample predictability of returns.

VIII. Conclusion

Our analysis provides a new perspective on markets in which decision-makers face high-dimensional prediction problems. Learning how to translate observed pricing-relevant 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, but 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 can be relevant for forecasting.

In the cross-sectional asset pricing setting, in-sample tests of return predictability largely lose their economic meaning when investors are faced with a large number of potential predictors of asset cash flows. 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.

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. As

the number of predictor variables that are available to researchers and investors has grown enormously, it is to be expected, even with Bayesian investors, that many of these variables show up as in-sample statistically significant cross-sectional return predictors. For this reason, it not clear that one should look for risk-based explanations or behavioral explanations for their in-sample predictive power.

In contrast to in-sample tests, out-of-sample tests retain their economic meaning in the high-dimensional case. Researchers should focus on explaining out-of-sample predictable variation in returns with economic models of priced risk or behavioral biases.

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, where investors observe a large number of exogenous and endogenous signals from which they can extract information not only about asset fundamentals, but also about the trading behavior of other investors.

References

Al-Najjar, Nabil I, 2009, "Decision Makers as Statisticians: Diversity, Ambiguity, and Learning," *Econometrica* 77, 1371–1401.

Anatolyev, Stanislav, 2012, "Inference in Regression Models with Many Regressors," *Journal of Econometrics* 170, 368–382.

Aragones, Enriqueta, Itzhak Gilboa, Andrew Postlewaite, and David Schmeidler, 2005, "Fact-Free Learning," *American Economic Review* 95, 1355–1368.

Bai, Z. D., and Y. Q. Yin, 1993, "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix," *Annals of Probability* 21, 1275–1294.

Calvano, Emilio, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello, 2018, "Artificial intelligence, Algorithmic Pricing and Collusion," Working paper, CEPR.

Campbell, John Y., and Samuel B. Thompson, 2008, "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?," *Review of Financial Studies* 21, 1509–1531.

Chan, Louis KC, Jason Karceski, and Josef Lakonishok, 2003, "The Level and Persistence of Growth Rates," *Journal of Finance* 58, 643–684.

Cochrane, John H., 2008, "The Dog That Did Not Bark: A Defense of Return Predictability," *Review of Financial Studies* 21, 1533–1575.

Cochrane, John H., 2011, "Presidential Address: Discount Rates," *Journal of Finance* 66, 1047–1108.

Collin-Dufresne, Pierre, Michael Johannes, and Lars A Lochstoer, 2016, "Parameter Learning in General Equilibrium: The Asset Pricing Implications," *American Economic Review* 106, 664–698.

DeBondt, Werner F.M., and Richard Thaler, 1985, "Does the Stock Market Overreact?," *Journal of Finance* 40, 793–805.

Dobriban, Edgar, and Stefan Wager, 2018, "High-Dimensional Asymptotics of Prediction: Ridge Regression and Classification," *Annals of Statistics* 46, 247–279.

Fama, Eugene, 1970, "Efficient Capital Markets: A Review of Theory and Empirical Work," *Journal of Finance* 25, 383–417.

Gabaix, Xavier, 2014, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics* 129, 1661–1710.

Harvey, Campbell R, Yan Liu, and Heqing Zhu, 2016, "... and the Cross-Section of Expected Returns," *Review of Financial Studies* 29, 5–68.

Heston, Steven L, and Ronnie Sadka, 2008, "Seasonality in the Cross-Section of Stock Returns," *Journal of Financial Economics* 87, 418–445.

Inoue, Atsushi, and Lutz Kilian, 2005, "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," *Econometric Reviews* 23, 371–402.

Jegadeesh, Narasimhan, and Sheridan Titman, 1993, "Returns to Buying Winners and Selling Losers: Implications for Market Efficiency," *Journal of Finance* 48, 65–91.

Klein, Timo, 2019, "Autonomous Algorithmic Collusion: Q-Learning Under Sequential Pricing," Working paper 2018-15, University of Amsterdam.

Lewellen, Jonathan, and Jay Shanken, 2002, "Learning, Asset-Pricing Tests and Market Efficiency," *Journal of Finance* 57, 1113–1145.

Lindley, Dennis V, and Adrian FM Smith, 1972, "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society: Series B (Methodological)* 34, 1–18.

Linnainmaa, Juhani T, and Michael R Roberts, 2018, "The History of the Cross-Section of Stock Returns," *The Review of Financial Studies* 31, 2606–2649.

Lo, Andrew W., and A. Craig MacKinlay, 1990, "Data-Snooping Biases in Tests of Financial Asset Pricing Models," *Review of Financial Studies* 3, 431–467.

McLean, David R., and Jeffrey Pontiff, 2016, "Does Academic Research Destroy Stock Return Predictability?," *Journal of Finance* 71, 5–32.

Novy-Marx, Robert, 2012, "Is Momentum Really Momentum?," *Journal of Financial Economics* 103, 429–453.

Shalev-Shwartz, Shai, and Shai Ben-David. *Understanding machine learning: From theory to algorithms* (Cambridge University Press, New York, NY, 2014).

Sims, Christopher A., 2003, "Implications of Rational Inattention," *Journal of Monetary Economics* 50, 665 – 690.

Tibshirani, Robert, 1996, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B (Methodological)* 58, 267–288.

Timmermann, Allan G, 1993, "How Learning in Financial Markets Generates Excess Volatility and Predictability in Stock Prices," *Quarterly Journal of Economics*, 1135–1145.

Yin, Y. Q., Z. D. Bai, and P. R. Krishnaiah, 1988, "On the Limit of the Largest Eigenvalue of the Large Dimensional Sample Covariance Matrix," *Probability Theory and Related Fields* 78, 509–521.

A. Proofs

Proof of Proposition 1. We start by showing that $\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1}$ is asymptotically Normal after appropriate standardization. (Recall that we view $\boldsymbol{X}$ as fixed and given.) Note first that as $\boldsymbol{h}_{t+1}$ is a linear combination of normally distributed mean-zero random variables, it is itself normally distributed with $\mathbb{E}\,\boldsymbol{h}_{t+1} = 0$. From equation (7), we have

$$N\,\mathbb{E}[\boldsymbol{h}_{t+1}\boldsymbol{h}_{t+1}'] = \frac{N\theta}{J}(\boldsymbol{I} - \boldsymbol{\Gamma}_t)^2 + \frac{N}{t}\boldsymbol{\Gamma}_t\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{\Gamma}_t + N\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{17}$$

after imposing $\boldsymbol{\Sigma}_g = \frac{\theta}{J}\boldsymbol{I}$ and $\boldsymbol{\Sigma}_e = \boldsymbol{I}$ (Assumption 4). It follows from the eigendecomposition (4) $\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$ that the terms on the right-hand side can be expressed in the form

$\boldsymbol{Q}\boldsymbol{\Xi}\boldsymbol{Q}'$ for appropriately chosen diagonal matrices $\boldsymbol{\Xi}$:

$$\boldsymbol{\Gamma}_t = \boldsymbol{Q}\left(\boldsymbol{I} + \frac{J}{\theta N t}\boldsymbol{\Lambda}^{-1}\right)^{-1}\boldsymbol{Q}' \tag{18}$$

$$\boldsymbol{I} - \boldsymbol{\Gamma}_t = \boldsymbol{Q}\left(\boldsymbol{I} + \frac{\theta N t}{J}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{Q}'. \tag{19}$$

We will repeatedly exploit the fact that $\boldsymbol{Q}$ is orthogonal, that is, $\boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$.

Adding the three terms on the right-hand side of equation (17), we find

$$N\,\mathbb{E}[\boldsymbol{h}_{t+1}\boldsymbol{h}_{t+1}'] = \boldsymbol{Q}\boldsymbol{\Omega}\boldsymbol{Q}'\,,$$

where $\boldsymbol{\Omega}$ is a diagonal matrix of eigenvalues with positive diagonal elements

$$\omega_j = \frac{1}{t\lambda_j + \frac{J}{\theta N}} + \frac{1}{\lambda_j}\,.$$

Now define $\widetilde{\boldsymbol{u}}_{t+1} = \sqrt{N}\boldsymbol{Q}\boldsymbol{\Omega}^{-1/2}\boldsymbol{Q}'\boldsymbol{h}_{t+1}$, so that $\sqrt{N}\boldsymbol{h}_{t+1} = \boldsymbol{Q}\boldsymbol{\Omega}^{1/2}\boldsymbol{Q}'\widetilde{\boldsymbol{u}}_{t+1}$ and $\widetilde{\boldsymbol{u}}_{t+1}$ is standard Normal: $\mathbb{E}\,\widetilde{\boldsymbol{u}}_{t+1}\widetilde{\boldsymbol{u}}_{t+1}' = N\boldsymbol{Q}\boldsymbol{\Omega}^{-1/2}\boldsymbol{Q}'\frac{1}{N}\boldsymbol{Q}\boldsymbol{\Omega}\boldsymbol{Q}'\boldsymbol{Q}\boldsymbol{\Omega}^{-1/2}\boldsymbol{Q}' = \boldsymbol{I}$. We can then write

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

The penultimate line exploits the fact that $\boldsymbol{\Omega}^{1/2}$ and $\boldsymbol{\Lambda}$ commute, as they are diagonal. The last line carries out a final simplification by defining $\boldsymbol{u}_{t+1} = \boldsymbol{Q}'\widetilde{\boldsymbol{u}}_{t+1}$. As $\boldsymbol{Q}$ is orthogonal, $\boldsymbol{u}_{t+1}$ is also a standard Normal random vector: $\mathbb{E}\,\boldsymbol{u}_{t+1}\boldsymbol{u}_{t+1}' = \mathbb{E}\,\boldsymbol{Q}'\widetilde{\boldsymbol{u}}_{t+1}\widetilde{\boldsymbol{u}}_{t+1}'\boldsymbol{Q} = \boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$.

Thus we can write

$$\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} = \sum_{j=1}^{J}\zeta_j u_j^2, \tag{20}$$

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

$$\zeta_j = \omega_j\lambda_j = \frac{\lambda_j}{t\lambda_j + \frac{J}{\theta N}} + 1. \tag{21}$$

(For comparison, under the rational expectations null we would have $\boldsymbol{\Omega} = \boldsymbol{\Lambda}^{-1}$ and hence $\zeta_j = 1$ for all j.) As we have $\lambda_j > 0$ by positive definiteness of $\boldsymbol{X}'\boldsymbol{X}$, it follows that for $t \geq 1$, $\zeta_j \in (1,2)$. Moreover, as $\lim_{J,N\to\infty}\frac{J}{N} = \psi > 0$ and (by Assumption 5) $\lambda_j > \varepsilon$, ζ_j 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, which here requires that for some $\delta > 0$

$$\lim_{N,J\to\infty} \frac{1}{s_J^{2+\delta}} \sum_{j=1}^{J} \zeta_j^{2+\delta}\, \mathbb{E}\left[(u_j^2-1)^{2+\delta}\right] = 0 \quad \text{where} \quad s_J^2 = 2\sum_{j=1}^{J} \zeta_j^2.$$

It is enough to show that this holds for $\delta = 1$. But as $\mathbb{E}\left[(u_j^2-1)^3\right] = 8$ and $\zeta_j \in (1,2)$,

$$\lim_{N,J\to\infty} \frac{8\sum_{j=1}^{J}\zeta_j^3}{\left(2\sum_{j=1}^{J}\zeta_j^2\right)^{3/2}} \leq \lim_{N,J\to\infty} \frac{64J}{2^{3/2}J^{3/2}} = 0\,,$$

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

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

The remaining results follow immediately. □

Proof of Proposition 2. The first statement follows from the second. To prove the second, note that Proposition 1 implies that

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

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 3. By decomposing returns into in-sample predicted component and residual, we can write the portfolio return as

$$\begin{aligned} r_{IS,t+1} &= \frac{1}{N}\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} + \frac{1}{N}\boldsymbol{h}_{t+1}'\boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{e}_{t+1} \\ &= \frac{1}{N}\boldsymbol{h}_{t+1}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} \\ &= \frac{1}{N}\sum_{j=1}^{J}\zeta_j u_j^2\,, \end{aligned}$$

where the last line comes from equation (20) and the u_j are IID standard Normal. Thus

$$\mathbb{E}\, r_{IS,t+1} = \frac{\sum_{j=1}^{J} \zeta_j}{N} \quad \text{and} \quad \operatorname{var} r_{IS,t+1} = \frac{2\sum_{j=1}^{J} \zeta_j^2}{N^2}\,,$$

so that $\mathbb{E}\, r_{IS,t+1} \to \psi\mu$ and $N \operatorname{var} r_{IS,t+1} \to 2\psi\left(\mu^2 + \sigma^2\right)$ as $N, J \to \infty$ with $J/N \to \psi$. This gives the result. □

Proof of Proposition 4. We want to calculate

$$\begin{aligned}\frac{1}{N}\,\mathbb{E}\left[\boldsymbol{r}_{t+1}'\boldsymbol{X}\boldsymbol{h}_{s+1}\right] = \frac{1}{N}\,\mathbb{E}\Big[&\boldsymbol{g}'\left(\boldsymbol{I}-\boldsymbol{\Gamma}_t\right)\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{I}-\boldsymbol{\Gamma}_s\right)\boldsymbol{g} + \bar{\boldsymbol{e}}_t'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{\Gamma}_t\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma}_s\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_s \\ &- \bar{\boldsymbol{e}}_t'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{\Gamma}_t\boldsymbol{X}'\boldsymbol{e}_{s+1} - \boldsymbol{e}_{t+1}'\boldsymbol{X}\boldsymbol{\Gamma}_s\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\bar{\boldsymbol{e}}_s \\ &+ \boldsymbol{e}_{t+1}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}_{s+1}\Big]\,. \end{aligned} \tag{22}$$

The right-hand side of this equation is unaffected if s and t are interchanged. Thus we may assume that $s < t$ without loss of generality. But then—using the fact that $\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{I}-\boldsymbol{\Gamma}_t\right) = \frac{J}{\theta t}\boldsymbol{\Gamma}_t$, which follows from equations (4), (18), and (19), together with the cyclic property of the trace, and Assumption 4—equation (22) reduces to

$$\frac{1}{N}\,\mathbb{E}\left[\boldsymbol{r}_{t+1}'\boldsymbol{X}\boldsymbol{h}_{s+1}\right] = \frac{1}{N}\operatorname{tr}\left[\frac{1}{t}\boldsymbol{\Gamma}_t\left(\boldsymbol{I}-\boldsymbol{\Gamma}_s\right) + \frac{1}{t}\boldsymbol{\Gamma}_t\boldsymbol{\Gamma}_s - \frac{\boldsymbol{\Gamma}_t}{t}\right] = 0\,. \qquad \square$$

Proof of Proposition 5. When $t > 0$, the cross-sectional moments of ζ_j can be computed using equation (12) 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 equations (15) and (16); and we can calculate the probability density function of ζ_j in the asymptotic limit by change of variable using the relationship between ζ_j and λ_j given in equation (21). □