Market efficiency in the age of big data

Journal of Financial Economics July 2022

Ian W.R. Martin, Stefan Nagel

汇报人：王光耀

2022.08.03

IAN MARTIN

- **Professor of Finance, London School of Economics, 2013-present**

- Harvard University, 2003-2008, PhD, Economics
- London School of Economics and Political Science, 2002-2003, MSc, Economics,
- Trinity College, University of Cambridge, 1995-1999, MA, Mathematics

- *"Sentiment and Speculation in a Market with Heterogeneous Beliefs", with Dimitris Papadimitriou, American Economic Review (2022), 112:8:2465-2517*
- *"Volatility, Valuation Ratios, and Bubbles: An Empirical Measure of Market Sentiment", with Can Gao, Journal of Finance (2021), 76:6:3211-3254*
- *"On the Autocorrelation of the Stock Market", Journal of Financial Econometrics (2021), 19:1:39- 52*

Stefan Nagel

- BOOTH SCHOOL OF BUSINESS UNIVERSITY OF CHICAGO , NBER , CEPR , CESIfo
  - 2016 - 2022 Executive Editor, Journal of Finance
  - 1999 - 2003 London Business School (UK), Ph.D. in Finance, 2003
  - 1993 - 1999 University of Trier (Germany), Diplom (M.S. equiv.), 1999, in Business Economics
- Asset pricing, investor behavior, and risk-taking of financial intermediaries.
  - Nagel S, Xu Z. Asset pricing with fading memory[J]. The Review of Financial Studies, 2022, 35(5): 2190-2245.
  - Korteweg A G, Nagel S. Risk-adjusted returns of private equity funds: a new approach[J]. Available at SSRN, 2022.
  - Nagel S. Machine learning in asset pricing[M]. Princeton University Press, 2021.

Abstract

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

1. Introduction

Introduction - Background

- Machine learning methods have proved useful in forecasting problems with huge numbers of predictor variables. High-dimensional prediction problems of this kind are faced not only by data scientists studying data as outside observers, but also by economic decision-makers in the marketplace. Many forward-looking economic decisions require predictions for which large numbers of variables could potentially be relevant, but the exact relationship between predictors and forecast target is unknown and must be learned from observed data. **In this paper, we argue that to understand market outcomes in such settings, it is important to take into account the high-dimensional nature of decision-makers' prediction problem.**
- We demonstrate this in an asset-pricing setting. We show that properties of asset prices are strongly affected by the dimensionality of investors' prediction problem.
- **Conventional notions of how to test market efficiency and how to interpret pricing anomalies break down in the high-dimensional case.**

Introduction -- Model

- Cash-flow growth rates of a cross-section of N firms are a linear function of J firm characteristics that are fixed over time.
- Investors are Bayesian, homogeneous, and risk neutral; they price stocks based on the predictive distribution of cash flows.
- Realized asset returns in this setting are simply equal to investors' forecast errors.
- N (__) J
- Rational Expectations Equilibrium
- A High-Dimensional Learning Problem

Introduction -- Machine Learning Methods

- 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.
- optimizing out-of-sample forecasting performance:
  - 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

Introduction -- Shrinkage

- 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.
  - Second, shrinkage implies underweighting the predictive information in the predictors. Naturally, this second component, too, is correlated with the predictor variables.
- To stack the deck against return predictability, we endow investors with prior beliefs that the coefficients of the cash flow–generating model （linear) are drawn from this prior distribution.
- With this objective prior, the optimal amount of shrinkage exactly balances the two components in such a way that investors' forecast errors and asset returns, are unpredictable out-of-sample.

Introduction -- In-Sample Predictability

- The fact that returns are not predictable out-of-sample, however, does not imply that there is no in-sample predictability.
- RE
- Small J Big N
- Small N Big J
- In a high-dimensional setting, the econometrician's ability to see data realized ex post, after investors' pricing decisions are made, gives her a substantial advantage.

Introduction – Econometrician In-Sample Predictability

- When J is vanishing in size relative to N, there is almost no predictability ?
- $N \to \infty$, $with\ J\ fixed$.
- $N, J \to \infty\ and\ \frac{J}{N} \to \psi$.
- In simulations, we show that these high-dimensional asymptotic results are a good approximation for the case of finite N with J comparable in size to N.
- The high rejection rates correctly reflect the fact that equilibrium asset prices contain in-sample predictable components that are large in a high-dimensional setting.

Introduction – Econometrician OOS Predictability

- The situation is different for out-of-sample tests.
- Intuitively, since Bayesian investors optimally use information available to them and price assets such that returns are not predictable under their predictive distribution, an econometrician who is restricted to constructing return forecasts using only data that had been available to investors in real time is not able to predict returns out-of-sample either.

Introduction -- Reject Market Efficiency

- These results illustrate forcefully that the economic content of the (semi-strong) market efficiency notion that prices “fully reflect” all public information is not clear in this high-dimensional setting, even though we abstract from the joint hypothesis problem by assuming that investors are risk-neutral.
- The null hypothesis in a vast empirical literature in asset pricing, including return predictability regressions, event studies, and asset-pricing model estimation based on orthogonality conditions, is the former version of the market efficiency hypothesis.
- Our results show that testing and rejecting this version has little economic content in a high-dimensional setting. An apparent rejection of market efficiency might simply represent the unsurprising consequence of investors not having precise knowledge of the parameters of a data-generating process that involves thousands of predictor variables.

Introduction -- The importance of OOS tests

- From the perspective of our model, it is not surprising that the technology-driven explosion in the number of predictor variables available to investors has coincided with an explosion in the number of return predictors that are found significant in asset-pricing studies
- out-of-sample tests gain additional importance in the age of Big Data.
- Fixed (RE): in- and out-of-sample methods test the same hypothesis
- high-dimensional learning: fundamentally different hypotheses

Introduction -- Perspectives

- out-of-sample portfolio returns can isolate predictable components of returns that reflect risk premia or behavioral biases. Unlike in-sample estimates, the out-of-sample estimates are not distorted by investors' learning-induced forecast errors.
- if the econometrician applies shrinkage similar to ridge regression in the in-sample return prediction regression, with the penalty hyperparameter estimated via cross-validation, then the portfolio return based on these shrinkage estimates can be equivalent to an out-of-sample portfolio return.
- We illustrate the different perspectives provided by in- and out-of-sample tests with an empirical example.

Introduction -- 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 long-term reversals , and momentum seasonality .
- In other words, there is substantial in-sample predictability. In terms of out-of-sample predictability, the picture looks very different.
- Using rolling regressions over 20-year windows to estimate prediction model coefficients and then using those to predict returns in subsequent periods, we find that predictability is generally much weaker out-of-sample than in-sample. Moreover, there is substantial decay over time. While some out-of-sample predictability exists in the early decades of the sample, it is basically nil in recent years.
- This suggests that there may be little reason to seek risk-based or behavioral explanations of the cross-sectional predictability that shows up in the in-sample analysis.

Introduction -- Explanation

- investors several decades ago were not able to process the information in each stock's price history as effectively as investors today.
- One can think of this as bounded rationality that induces excessive shrinkage or sparsity of investors' forecasting models
- 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.

Introduction -- Conclusion

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

Introduction – Literature Review : Learning

- Learning can induce in-sample return predictability relates to an earlier literature that studies learning-induced return predictability in low-dimensional time-series settings with few return predictors
- These earlier time-series analyses do not address the question of how learning affects asset prices and the properties of in- and out-of-sample return predictability tests in a high-dimensional cross-sectional setting. This is the focus of our paper.
- (e.g., Timmermann, 1993, Lewellen, Shanken, 2002, Collin-Dufresne, Johannes, Lochstoer, 2016).

Introduction – Literature Review : Approach

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

2. Bayesian pricing in a high-dimensional setting

Model Setting – Economy

- Economy: discrete time with N assets.
- Each asset is associated with a vector of J firm characteristics observable to investors. -- N x J matrix X
- The assets pay dividends, collected in the vector Yt whose growth is partly predictable based on X.

Model Setting – Assumption 1

Assumption 1

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

- The set of characteristics:
  - The set of characteristics is potentially very large, but for simplicity we assume J < N.
  - The set of characteristics in X exhausts the set of variables that investors can condition on.
  - Due to technological change, this set could change as previously unavailable predictors become available, so we will be concerned with the behavior of prices for various values of J.
- The assumption (trace) that is a normalization that defines a natural scale for the characteristics.
- We assume that the characteristics associated with a firm are constant over time for simplicity.

Model Setting – Assumption 2

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.

Model Setting -- Investors' Expectations

- We focus on the pricing of one-period dividend strips so that ***pt*** 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.
- The errors in are actually the averages of the errors one would find by sampling at higher frequencies over many shorter subperiods.

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

Model Setting -- Investors' Expectations

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

- RE:

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

$$\boldsymbol{p}_t = \boldsymbol{y}_t + \boldsymbol{X}\boldsymbol{g},$$

$$\begin{aligned} \boldsymbol{r}_{t+1} &= \boldsymbol{y}_{t+1} - \boldsymbol{p}_t \\ &= \Delta \boldsymbol{y}_{t+1} - \boldsymbol{X}\boldsymbol{g} \\ &= \boldsymbol{e}_{t+1} \end{aligned}$$

- Learning:

$\{\Delta \boldsymbol{y}_s\}_1^t$ $\boldsymbol{X}$.

We assume that investors know $\boldsymbol{\Sigma}_e$.

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

OLS in high dimension – J is close to N

- OLS

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

$$\boldsymbol{p}_t=\boldsymbol{y}_t+\Delta\boldsymbol{y}_t$$

$$\boldsymbol{r}_{t+1}=\Delta\boldsymbol{y}_{t+1}-\Delta\boldsymbol{y}_t$$

$$\operatorname{var}\left(\boldsymbol{e}_{t+1}-\boldsymbol{e}_t\right),$$

- Random Walk

$$\mathbb{E}_t\Delta\boldsymbol{y}_{t+1}=\mathbf{0}$$

$$\boldsymbol{p}_t=\boldsymbol{y}_t$$

$$\boldsymbol{r}_{t+1}=\Delta\boldsymbol{y}_{t+1}$$

$$\operatorname{var}\left(\boldsymbol{X}\boldsymbol{g}+\boldsymbol{e}_{t+1}\right).$$

If a relatively small component of cash-flow growth is predictable, that is, if $\operatorname{var}(\boldsymbol{X}\boldsymbol{g}) \ll \operatorname{var}\boldsymbol{e}_{t+1}$, then the random walk forecast MSE may be substantially lower than the OLS forecast MSE.

2.1. Priors and posteriors

- The prior implicit in the least-squares estimator is economically unreasonable.
- The posterior mean equals the GLS estimator if investors' prior for g is diffuse. But a diffuse prior for g is not a plausible assumption.
- Economic reasoning should lead investors to realize that the amount of predictable variation in Δy_{t+1} must be limited. It does not make economic sense for investors to believe that arbitrarily large values for g are just as likely as values that give rise to moderate predictable variation in Δy_{t+1}.
- While they might not have very precise prior knowledge of g, it reasonable to assume that the distribution representing investors' prior beliefs about g is concentrated around moderate values of g.

Assumption 3

Before seeing data, investors hold prior beliefs

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

Proposition 1

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

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

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

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

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

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

Assumption 4

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

Assumption 1

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

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

using Assumptions 1 and 4.

Assumption 1

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

Assumption 4

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

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

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

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

$$\tfrac{1}{J}\textstyle\sum_{i=1}^{J}\lambda_i = \tfrac{1}{J}\operatorname{tr}\tfrac{1}{N}\boldsymbol{X}'\boldsymbol{X} = 1, \tag{5}$$

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

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

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

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

Proposition 2

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

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

True Model

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

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

3. Asymptotic analysis

High-dimension: $N, J \to \infty, \frac{J}{N} \to \psi > 0$

The high-dimensional asymptotics are intended to provide a tractable approximation for the case where J and N are finite and J is not small relative to N.

3.1. In-sample predictability

- The econometrician looks for return predictability by regressing r_{t+1} on the variables in X using OLS.
- As the number of predictor variables increases, $J \to \infty$, we are potentially in the realm of Big Data.

Assumption 5 Big Data

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

$\boldsymbol{Xv}$ $(\boldsymbol{Xv})'(\boldsymbol{Xv}) = \lambda_{\min} \approx 0.$

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

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

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

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

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

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

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

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

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

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

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

We write the limiting mean and variance of the eigenvalues as

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

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

Proposition 3

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

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

It follows that the test statistic T_{re} satisfies

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

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

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

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

Proposition 4

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

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

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:

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

for any critical value c_α.

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

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

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

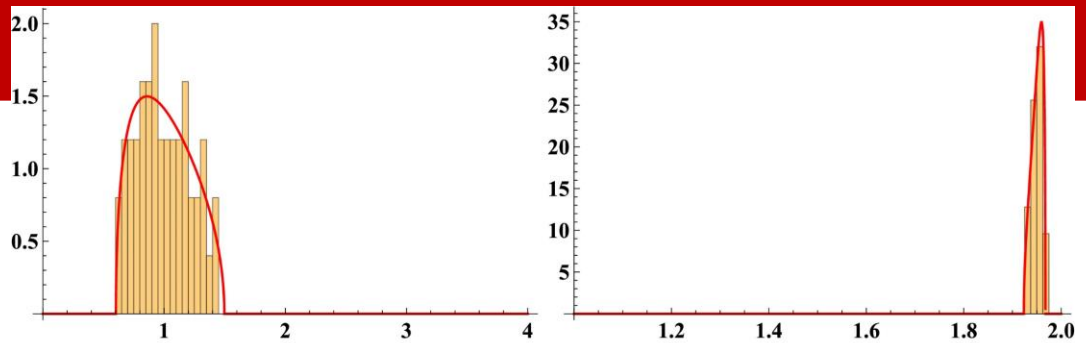
$$\begin{aligned} \boldsymbol{r}_{t+1}' \boldsymbol{w}_{IS,t} &= \frac{1}{N} \boldsymbol{r}_{t+1}' \boldsymbol{X} \boldsymbol{h}_{t+1} \\ &= \frac{1}{N} \boldsymbol{h}_{t+1}' \boldsymbol{X}' \boldsymbol{X} \boldsymbol{h}_{t+1}. \end{aligned}$$

Thus the two approaches are equivalent.

3.2. A benchmark example: X a random matrix

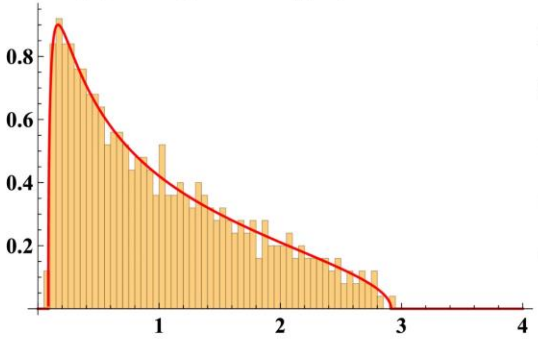
- $X \sim \psi\ , \psi = \frac{J}{N}$
- X has IID entries x_{ij} with mean zero, unit variance, and finite fourth moment.
- Nature generates this matrix once before investors start learning, and it stays fixed thereafter.
- random matrix theory:
  - The eigenvalue distribution converges to the *Marchenko-Pastur distribution*. $N, J \rightarrow \infty, \frac{J}{N} \rightarrow \psi$
  - For ψ close to one, this distribution features substantial probability mass on eigenvalues close to zero, indicating that many of the columns of X are close to being collinear.

$$\lambda_j \in \left[\left(1-\sqrt{\psi}\right)^2, \left(1+\sqrt{\psi}\right)^2\right] \text{ for all } j.$$

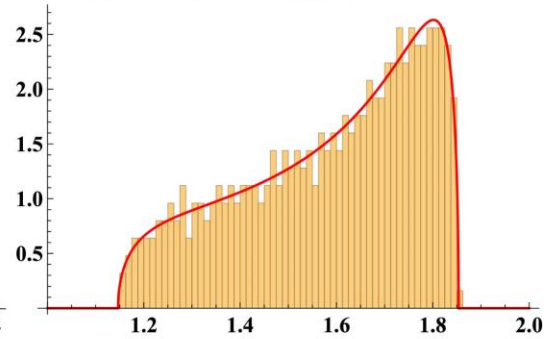


(a) Histogram of λ_j, $J/N = 0.05$.

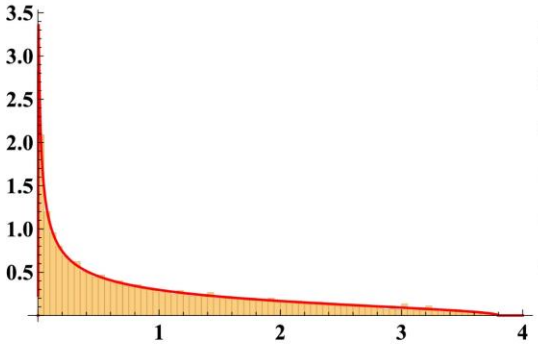
(b) Histogram of $\zeta_{j,t}$, $J/N = 0.05$.



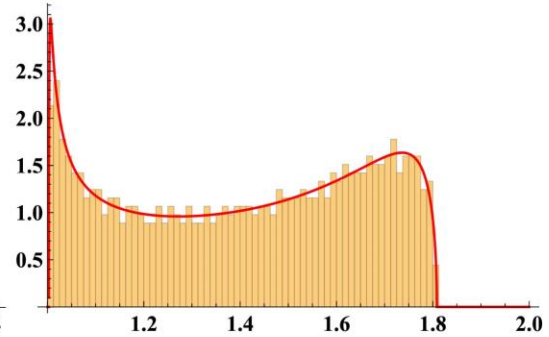
(c) Histogram of λ_j, $J/N = 0.5$.



(d) Histogram of $\zeta_{j,t}$, $J/N = 0.5$.



(e) Histogram of λ_j, $J/N = 0.9$.



(f) Histogram of $\zeta_{j,t}$, $J/N = 0.9$.

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

Proposition 5

The cross-sectional moments of $\zeta_{j,t}$ *satisfy*

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

and

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

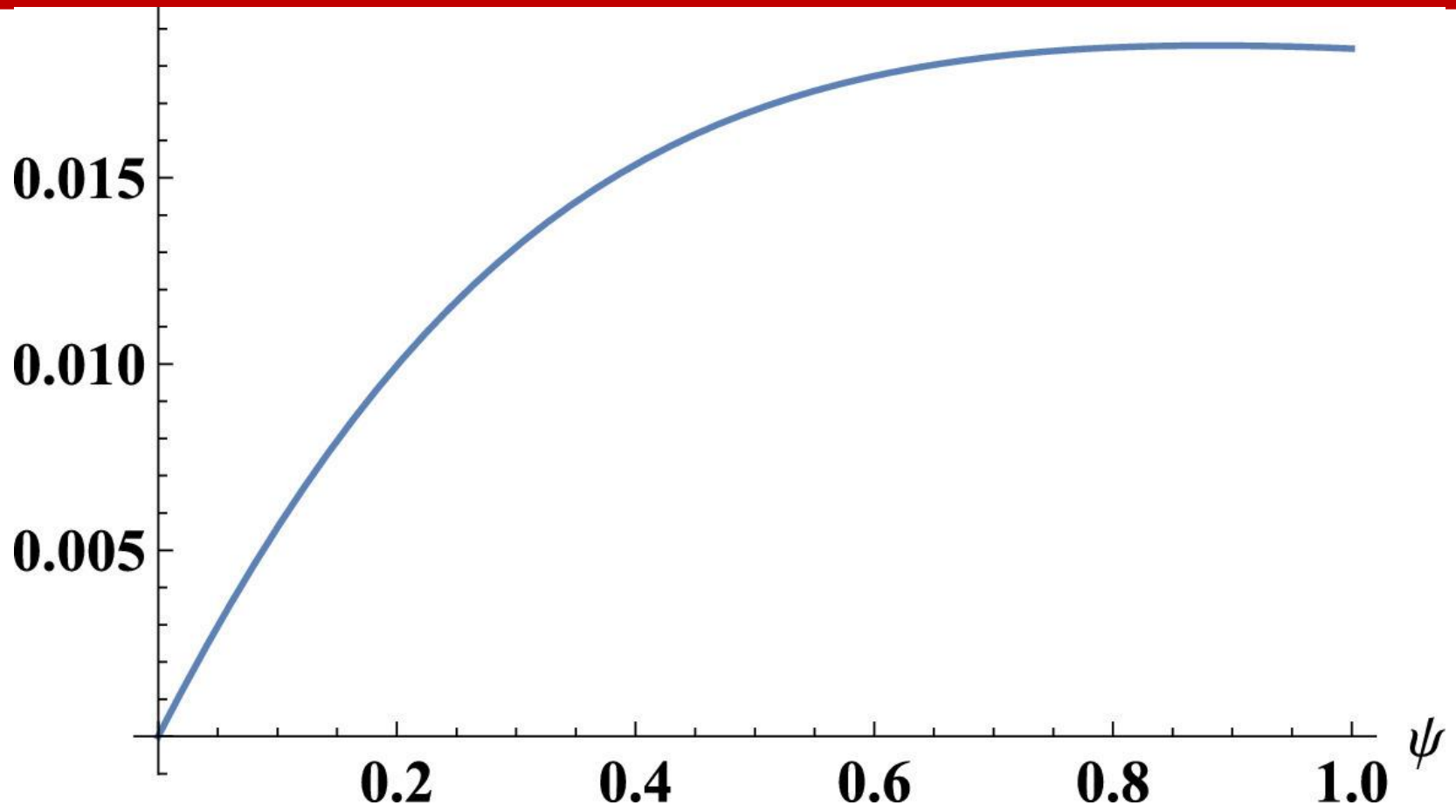


Figure 2 shows how the rate function (16) depends on ψ for $\theta = 1$ and $t = 1$. For $\psi > 0.4$, the rate is higher than 0.015, indicating that the probability of not rejecting the null is on the order of $\exp(-0.015N)$, which is a tiny number even for relatively small cross-sections of, say,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

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

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

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

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

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

Proposition 6

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

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

and, in the large N, J limit,

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

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

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

- Case1: t>s
- Case2: t<s
- cross-validation
- Backwards prediction, forward prediction, and combinations of the two (as in cross-validation) are equivalent.

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

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

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

How can the econometrician estimate the risk premium component $\boldsymbol{X}\boldsymbol{\gamma}$?

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

$$\tfrac{1}{N}\boldsymbol{\gamma}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\gamma} = \mathbb{E}\boldsymbol{r}_{OOS,t+1}. \tag{21}$$

3.5. Comparison with cross-validated penalized regression

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

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

The solution to the problem (22) is

$$\boldsymbol{b}_t = \frac{1}{1+\xi} \boldsymbol{h}_t, \tag{23}$$

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

$$\begin{aligned} \xi^* = \arg\min_{\xi} \mathbb{E} \Bigg[& \left\{ \boldsymbol{r}_{t+1} - \frac{1}{1+\xi} \boldsymbol{X} \boldsymbol{h}_{s+1} \right\}' \\ & \left\{ \boldsymbol{r}_{t+1} - \frac{1}{1+\xi} \boldsymbol{X} \boldsymbol{h}_{s+1} \right\} \Bigg]. \end{aligned} \tag{24}$$

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

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

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

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

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

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

4. Finite-sample analysis: simulations

Parameters Setting

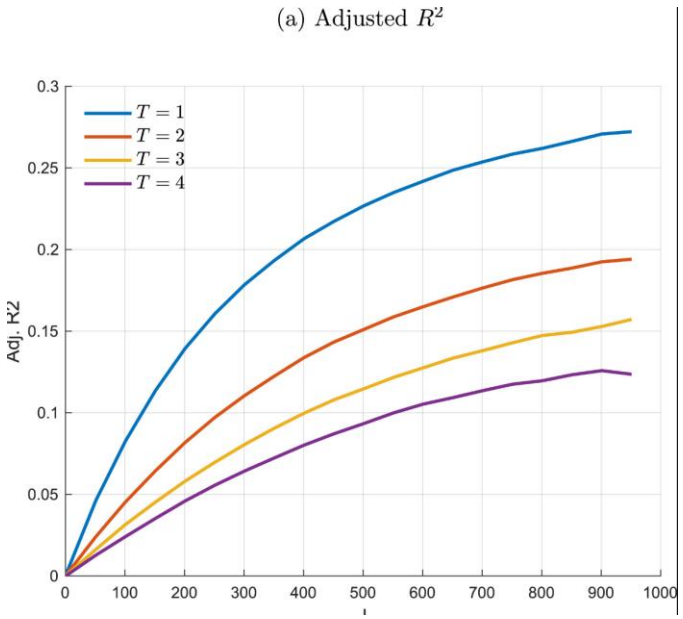
- We set N=1000 and let J vary from 1 to close to 1000. We draw the elements of X from a standard Normal distribution.
- For the purpose of this numerical analysis, we also need to set the parameter θ that pins down the share of predictable variation in cash-flow growth through $\sum_g = \frac{\theta}{J} I.$ Set $\theta = 1$

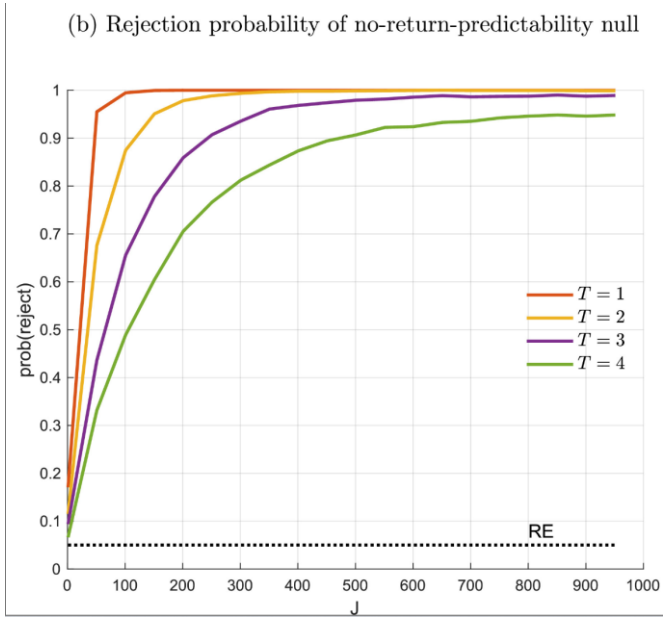
Based on our data-generating process for cash-flow growth in (1), annualized growth rates over a horizon of T periods are

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

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

4.1. Return prediction with many predictors





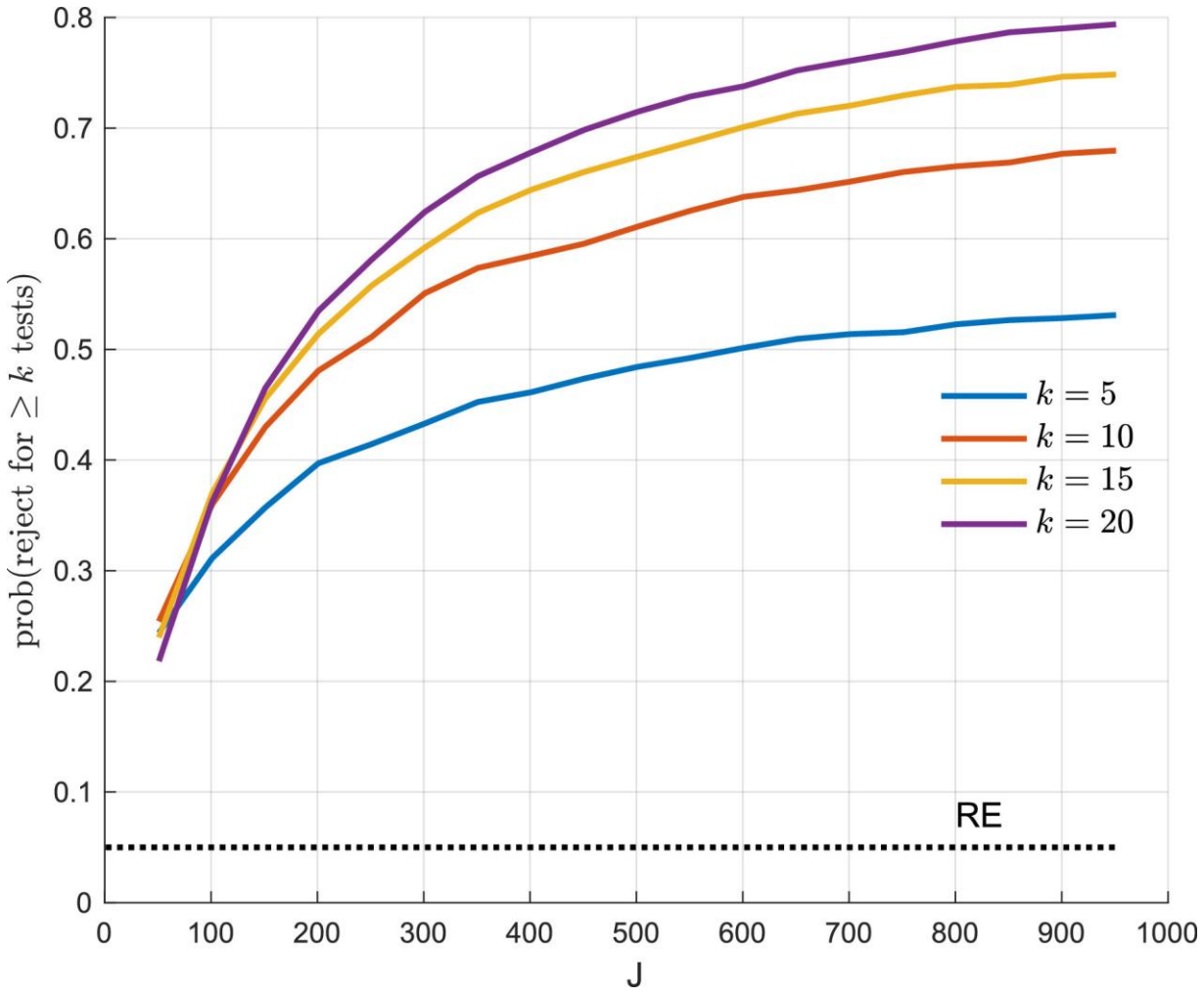
- We now simulate cash flows and, based on investors' Bayesian updating and pricing. N = 1000.
- We then consider an econometrician who samples these returns ex post and runs regressions of on after investors have learned about for periods.
- Fig. 3. In-sample return predictability tests. Based on cross-sectional regressions with N=1000 assets and J predictor variables, predicting the last return in a sample of size T+1 and where investors have learned about g from a sample of size T.

- Fig3a:
- Fig3b:

4.2. Return prediction with single predictors and multiple testing

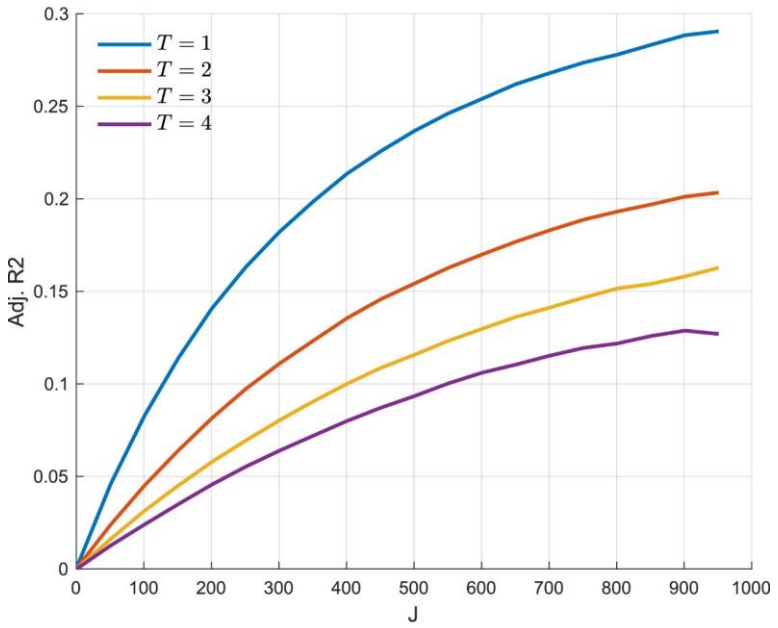
- In our analysis so far, we assumed that the econometrician runs a ***kitchen-sink*** regression using all variables in X as predictors.
- We now imagine that J econometricians, indexed by j=1,2,…,J, run regressions of r_{t+1} on a single characteristic x_j (i.e., column j of X) and each of them tests the hypothesis H_j that x_j does not predict returns. How many of them will reject the no-predictability null? Does high dimensionality of investors' learning problem lead to more rejections?

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

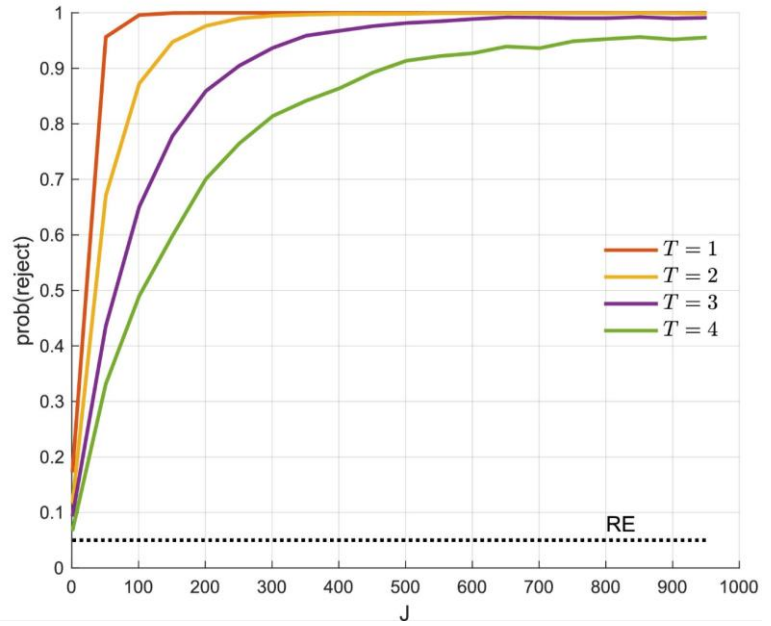


5. Sparsity

(a) Adjusted R^2

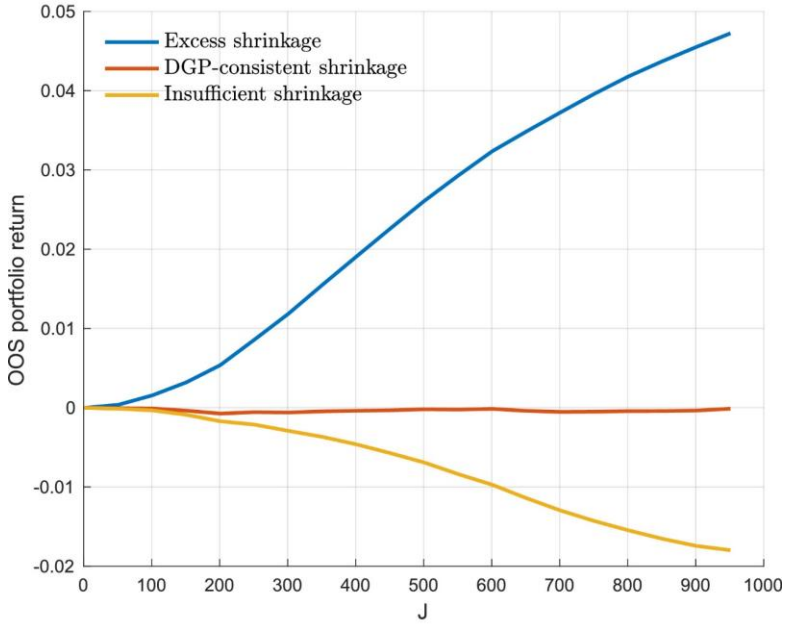


(b) Rejection probability of no-return-predictability null

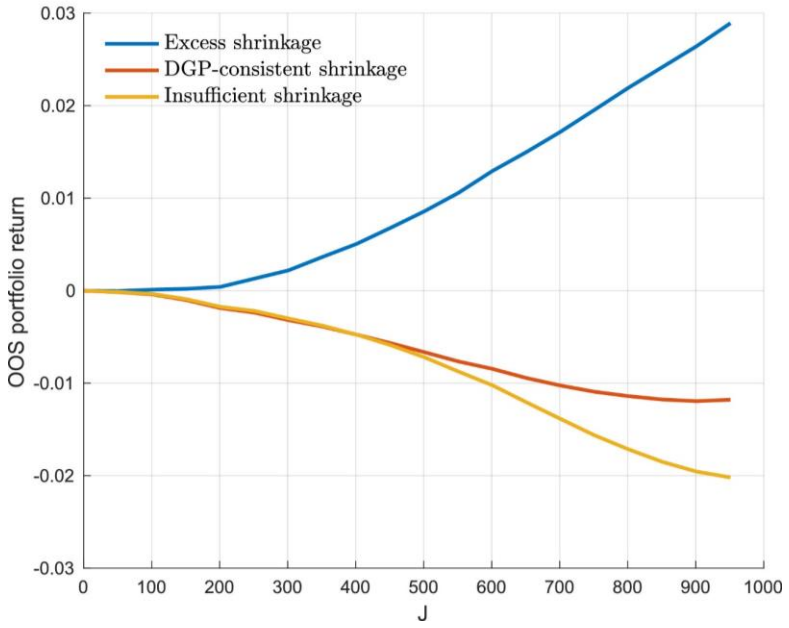


6. Excess shrinkage or sparsity

(a) Ridge

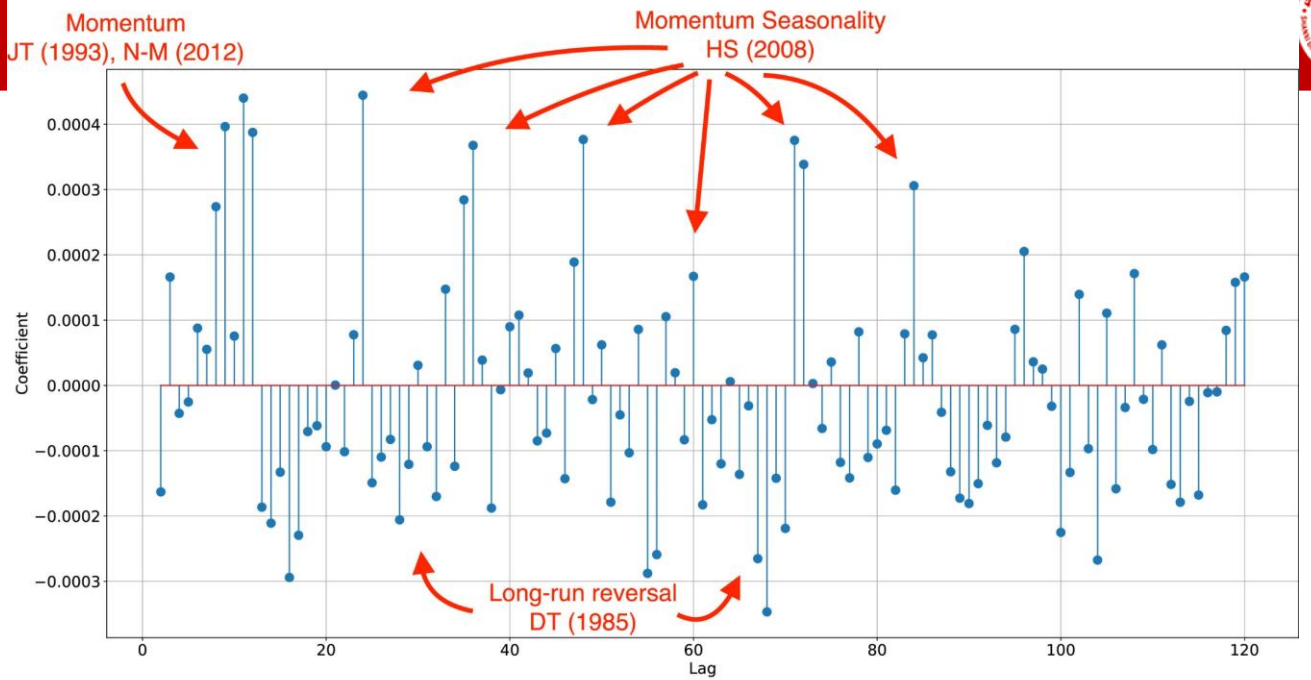


(b) Lasso

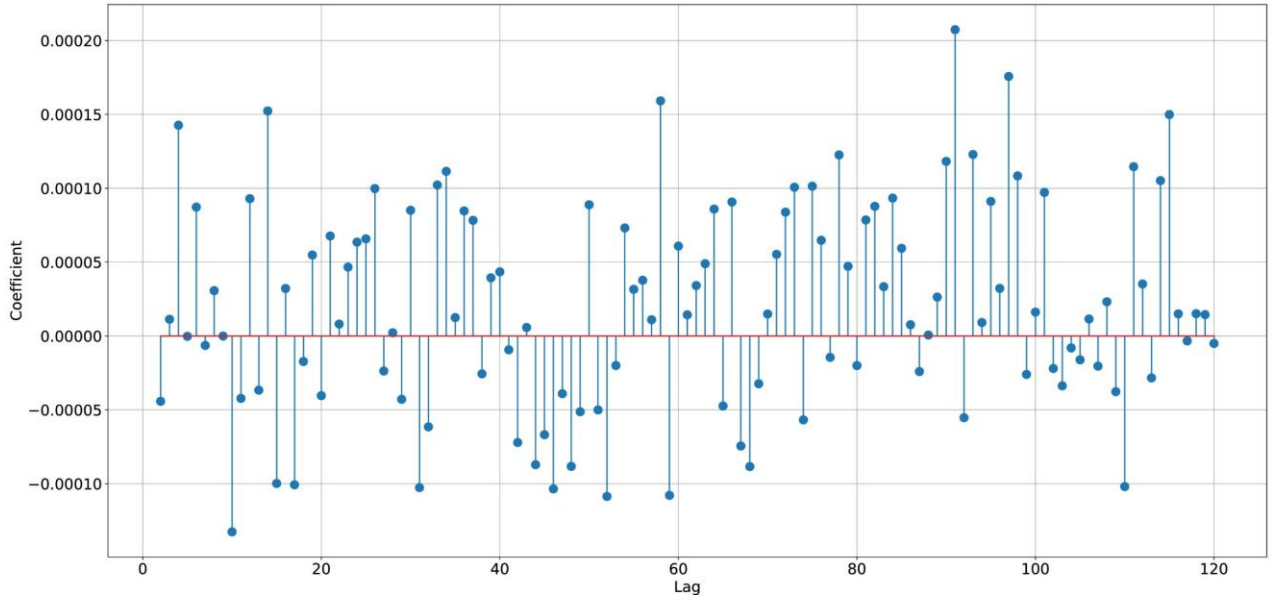


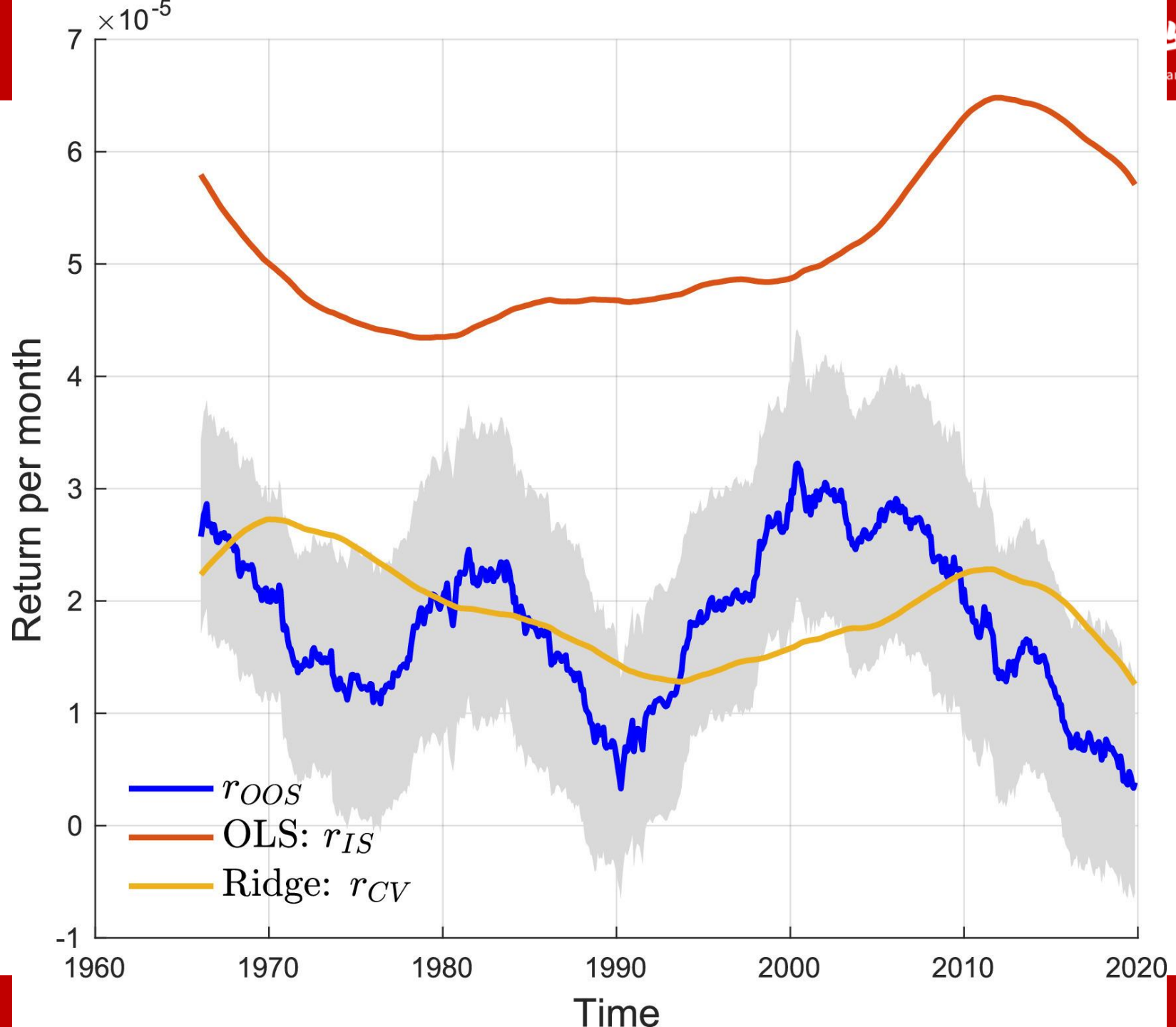
7. Empirical application: predicting stock returns with past returns

(a) Coefficients for past returns



(b) Coefficients for past squared returns




×10^-5
7
6
5
4
3
2
1
0
-1
Return per month
1960
1970
1980
1990
2000
2010
2020
Time
r_{OOS}
OLS: r_{IS}
Ridge: r_{CV}

8. Conclusion

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

- In the cross-sectional asset-pricing setting, in-sample tests of return predictability lose their economic meaning when investors are faced with many possible predictors of asset cash flows.
- The usual economic interpretation that in-sample predictable returns represent priced risks or the effects of investors' behavioral biases does not apply in this case.
- This is not a statistical problem with the sampling properties of the econometrician's predictability tests. Instead, it is a problem with the null hypothesis in these tests.
- As investors' learning problem becomes harder with increasing dimensionality of the set of potential predictors, the true properties of equilibrium prices change. Even in the absence of risk premia and behavioral biases, the usual null hypothesis that returns are unpredictable need not apply.
- Investors' learning of the cash flow–forecasting model parameters leaves in-sample predictable components in returns that reflect investors' real-time estimation error and the shrinkage they optimally apply to reduce it in a high-dimensional setting.

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

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

- 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.