Market efficiency in the age of big data

Ian Martin, Stefan Nagel

Journal of Financial Economics, 2021

Long Zhen

Introduction

- Standard approaches in asset pricing and market efficiency assume rational expectation:
  - No learning problem: investors know $f(\cdot)$ for $E(return)$
  - IS return predictability = risk premium/mispricing
- As technology has improved, the set of available and potentially valuation-relevant predictor variables expanded enormously.
  - →High dimensional prediction problems in finance
- When investors learn about $f(\cdot)$ in big data setting, in-sample predictability takes place in equilibrium.
  - → necessity for OOS testing

To be clear

- In-sample vs. Out-of-sample?
  - In-sample: use samples in the whole period
    - RE: Investors already have perfect knowledge of model parameters. The same would be true in low-dimension.
    - High-dimension: The ability to see data realized ex-post brings substantial advantage
  - Out-of-sample: only use available samples at the time-point.
    - Form portfolios based on stock return predictions
    - RE & High-dimension: not predictable

- When data-generating process is fixed, in- and out-of-sample methods test the same hypothesis, and in-sample tests are more powerful
- When there's no fixed such process, in- and out-of-sample methods test different hypotheses, providing a clear motivation for out-of-sample testing.

Model the economic actors as machine learners:

- ML perform well in high-dimension forecasts.
  - Handle by imposing regularization on the estimation
  - Bayesian interpretation: based on economic considerations, some parameters' posterior beliefs are shrunk towards zero.

Introduction – literature

- Al-Najjar(2009):
  - non-Bayesian high dimension setting, focus on disagreement between agents
- Klein(2019):
  - Strategic interaction of ML pricing in product market

- This paper:
  - A Bayesian linear framework and without strategic interactions.

Research Design – Setup

- Investors are Bayesian, homogeneous, risk-neutral, and price stocks based on the predictive distribution of cash flows
- Rational expectation(RE) equilibrium:
  - β are known or can be estimated ($J \ll N$)
  - → no mispricing, realized asset returns are simply equal to investors forecast errors. → unpredictable (market efficiency)
- Reality: $J \approx N$

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

- Benchmarks:
  - Rational expectations: investors know g
    - This is the null hypothesis indicating market efficiency
  - OLS: regress past cashflow growth on X to estimate g
  - Random walk: give up on forcasting
  - Bayesian learning →

Bayesian pricing in high-dimensional setting

- An economy in discrete time, $t \in \{1,2, \dots\}$, N assets, each associated with J observable characteristics. The matrix X is $N \times J$.
- For simplicity:
  - $J < N$, but large- N and large- J asymptotics
  - The characteristics associated with a firm are constant
- Abstract from risk premia:
  - Investors are risk-neutral and the interest rate is zero
- Dividend growth $\Delta y_{t+1} = \mathrm{Xg} + \mathrm{e_{t+1}}$

- Priors and posteriors
  - Investors' prior beliefs: $g \sim N\left(0, \frac{\theta}{J} I\right)$
    - $J \uparrow, Var(g) \downarrow$, ensuring variance doesn't explode
  - Posterior mean is a ridge regression estimator

$$\tilde{g}_t = \Gamma_t (X'X)^{-1} X' \overline{\Delta y_t}$$

    - i.e., OLS estimator shrunk towards prior mean by the matrix

$$\Gamma_t = Q \left(I + \frac{J}{N\theta t} \Lambda^{-1} \right)^{-1} Q'$$

- Shrinkage strong
  - if t small (short time dimension)
  - if θ small (prior tightly concentrated around zero)
  - if J/N is large (many predictors)
  - along unimportant principal components of X (small eigenvalues)

Equilibrium realized returns

Proposition

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

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

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

- "underreaction" to $\boldsymbol{X}$ due to shrinkage
- "overreaction" to estimation error in $\tilde{\boldsymbol{g}}_t$, dampened by shrinkage
- unpredictable shock (the only term in RE case)

In-sample predictability test:

- Coefficients: $\boldsymbol{h}_{t+1} = (\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} - \boldsymbol{\Gamma}_t \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\bar{\boldsymbol{e}}_t + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{e}_{t+1}$
- Test statistic: $T_{re} \equiv \dfrac{\boldsymbol{h}'_{t+1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{h}_{t+1} - J}{\sqrt{2J}}$
  - Case 1: a few principal components span the data
    - i.e. the eigenvalues of $\frac{X'X}{N} \to 0$
    - Market efficiency test works as usual, $T_{re} \xrightarrow{d} N(0,1)$

- Case 2: “big data”
  - i.e. the eigenvalues of $\frac{X'X}{N} > \epsilon$
  - The entries of X are iid random variables. →

Proposition

In equilibrium, 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 $1 < \sqrt{\mu^2+\sigma^2} < 2$ are determined by eigenvalues

- Therefore,

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

- In a big data world, we are almost certain to reject the RE null

- Simulation: $N = 1000$ assets

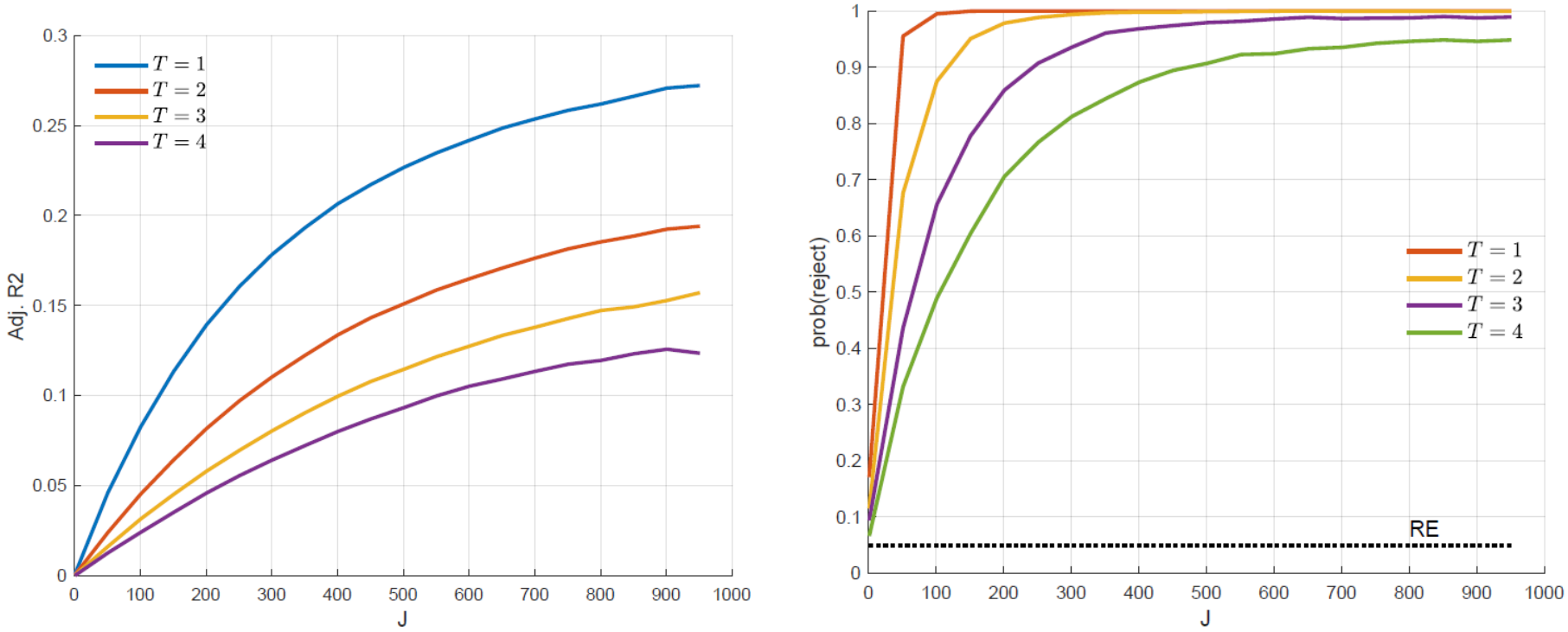


Figure: In-sample return predictability tests

- Finding: some significant anomalies in-sample, but predictability is much weaker out of sample.
  - →Little reason to seek risk-based or behavioral explanations
- In the earlier period out-of-sample predictability indicates ex-ante risk premia or mispricing at that time, or investors were not able to process the information.

Summary so far

- Asset returns are different between big data world and small data or RE world.
- IS return predictability in high dimension need not be consequence of risk premia or behavioral biases
  - IS predictability (the RE null) is simply false in big data

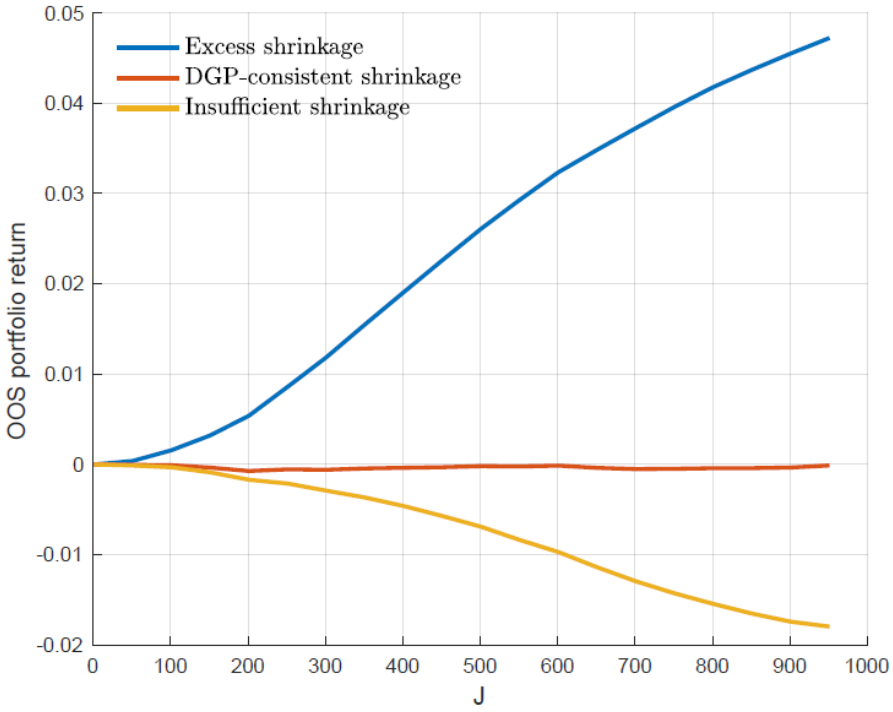
Out-of-sample return predictability

Proposition

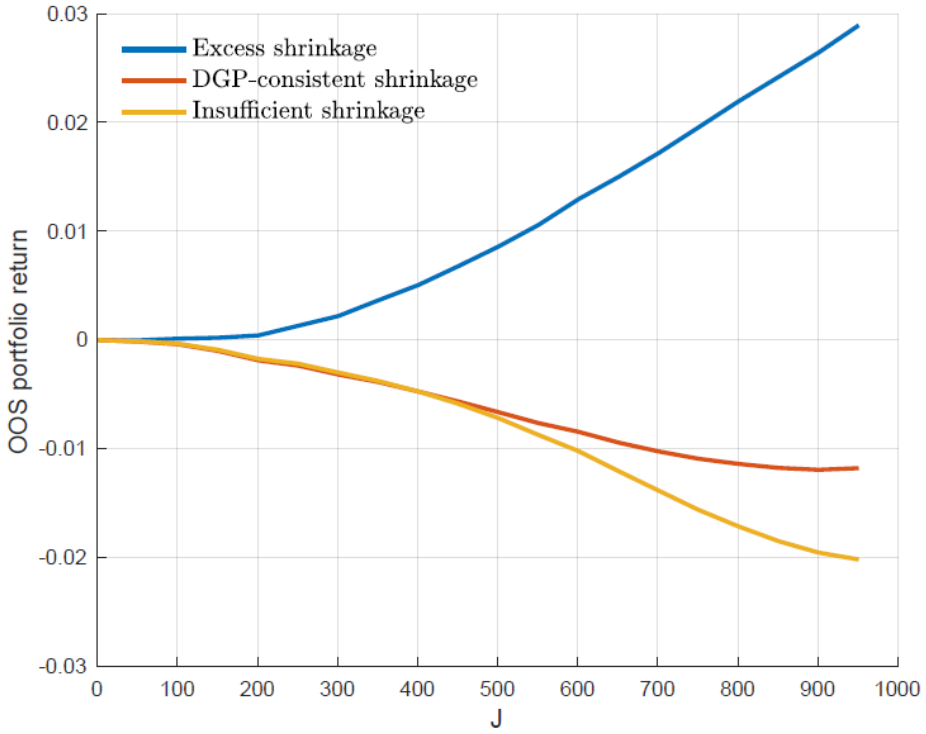
Consider an out-of-sample strategy with predicted returns as portfolio weights, $r_{OOS,t+1} = \boldsymbol{r}'_{t+1}\boldsymbol{X}\boldsymbol{h}_{s+1}$ *where* $t \neq s$*. Then* $\mathbb{E}\, r_{OOS,t+1} = 0$

- Forward case $t > s$ is natural: Investors are Bayesian so the econometrician cannot "beat" investors
- Backward case $t < s$ is more surprising. Not a tradable strategy, but interesting for research
  - Suggests backwards OOS tests (e.g., Linnainmaa and Roberts 2018) and cross-validation (e.g., Kozak, Nagel and Santosh 2020; Bryzgalova, Pelger, and Zhu 2020) could be appropriate for Bayesian learning setting

(A) Ridge



(B) Lasso



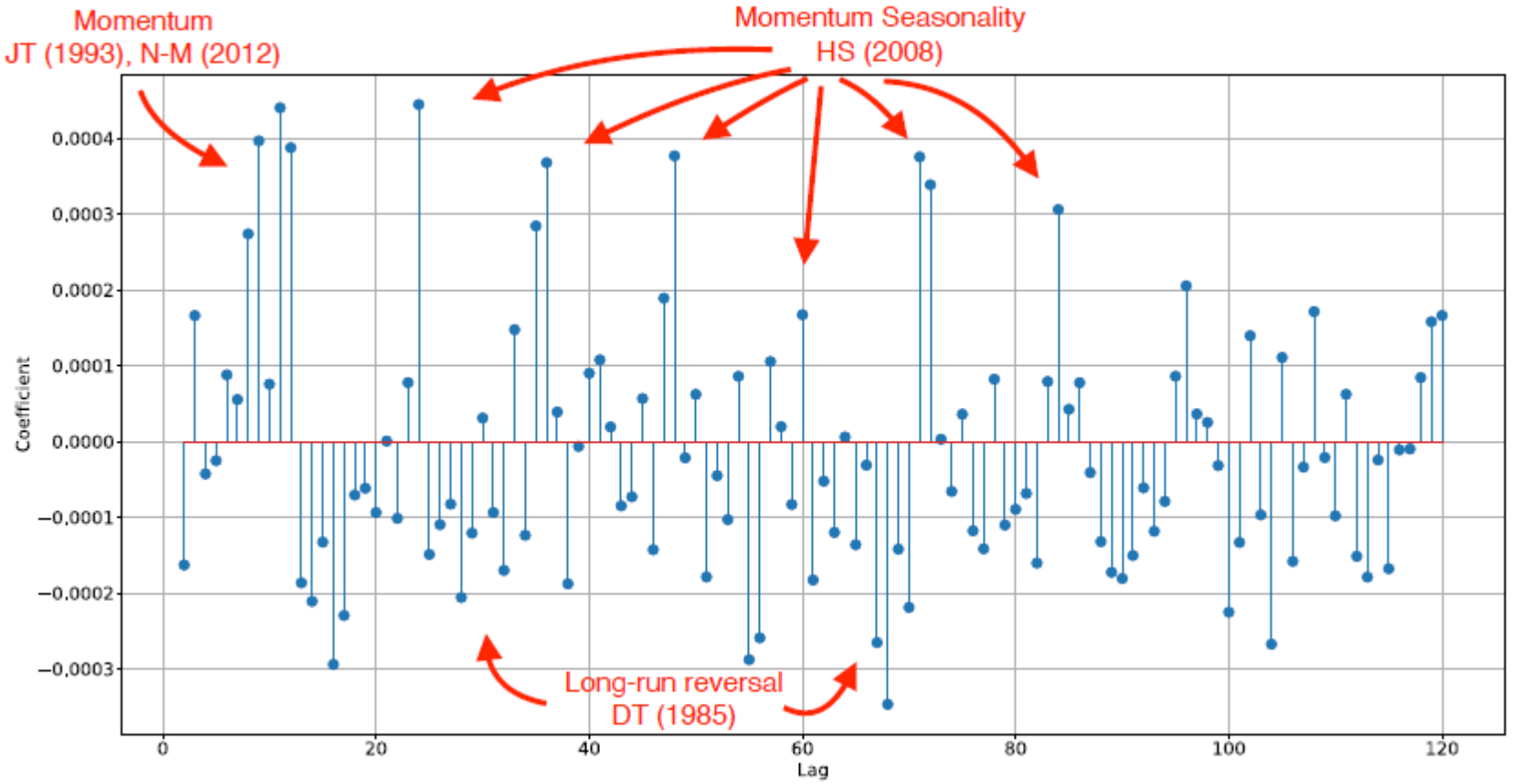
Research Design – simulation

- Finding: sparsity or shrinkage beyond the level called for by objectively correct Bayesian priors leads to positive out-of-sample return predictability
- → out-of-sample predictability may reflect the fact that some of the variables today were not available at the time.

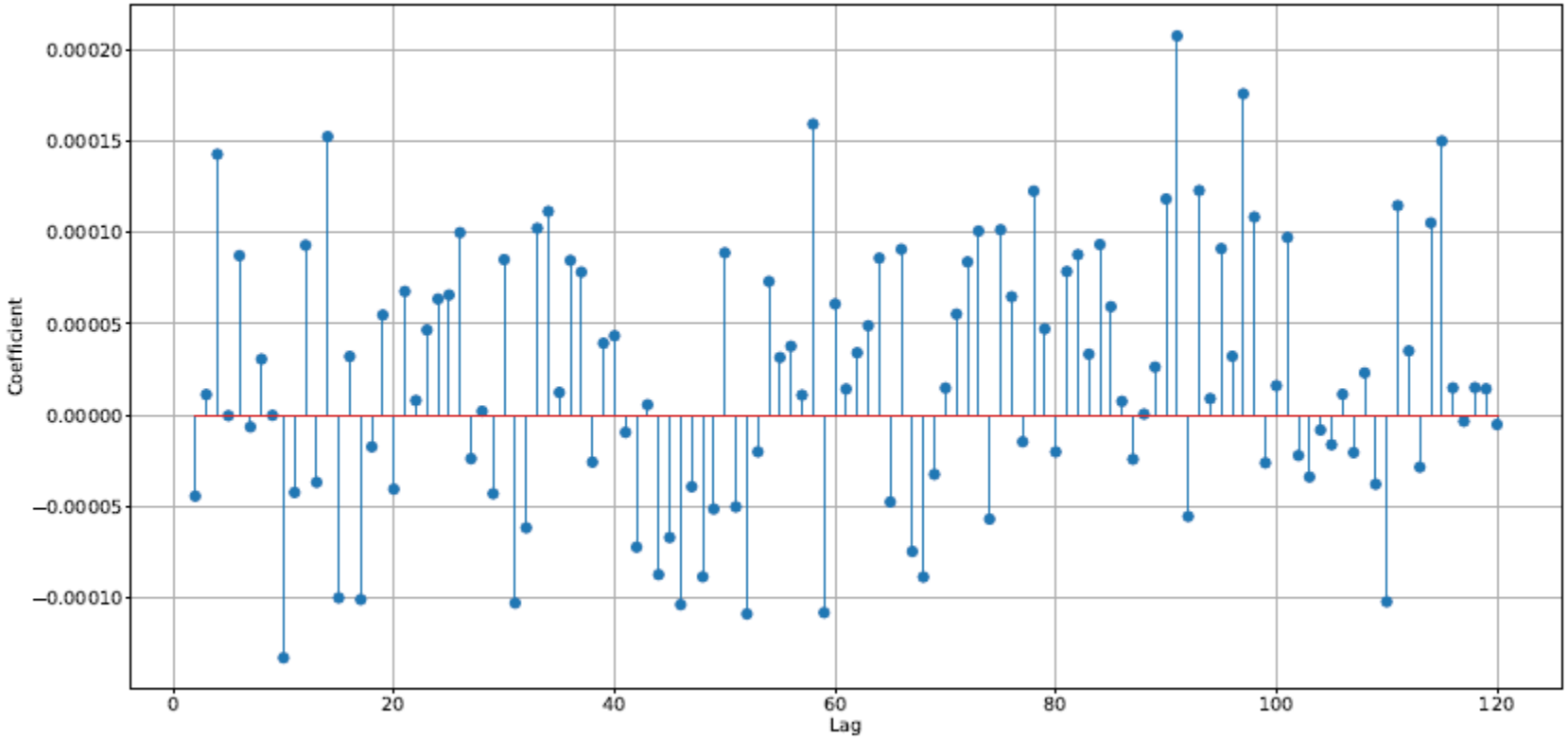
Empirical: IS vs. OOS predictability

- Consider a large set of predictors
  - History of monthly **simple** and **squared returns** over the previous 120 months as a set of return predictors
- The cross-section of US stocks
  - All U.S. stocks on CRSP, except market cap < 20th NYSE percentile or price < $1 at the end of month t-1
- All predictors cross-sectionally demeaned and standardized to unit S.D. each month
- Ridge regression with leave-one-year-out cross-validation to choose penalty parameter value

- In-sample: past return coefficients



- In-sample: past squared return coefficients



IS vs. OOS returns

- RE model: expected IS and OOS portfolio returns both equal $\gamma' X' X \gamma$
- Learning effect:
  - Still true for the OOS portfolio return
  - IS return is distorted by learning-induced components that are not predictable OOS

- →IS predictability doesn't carry over to OOS predictability and hence doesn't reflect risk premia demanded by investors ex ante, or persistent belief distortions

Conclusion

Market Efficiency in the age of big data:

- In big data setting, RE is implausible
- Learning has strong effects on asset prices
- Risk premia and bias theories should focus on explaining OOS, not IS predictability
- Investor learning provides clear motivation for OOS