Market Efficiency in the Age of Big Data

Ian Martin    Stefan Nagel

June 2020
Investors’ Big Data problem

- Investors forecasting cash flows face huge number of potential predictors

- $E[\text{c.f.}] = f(\text{predictors})$ unknown: high-dimensional learning problem

- Consequences for asset pricing? Market efficiency tests? Estimation of risk premia?

![Graph showing data points representing different databases and their associated variables and megabytes.](image.png)
High-dimensional learning in asset pricing

- Standard approaches in asset pricing and market efficiency testing assume rational expectations (RE)
  - Assumes away learning problem: investors know $f(.)$ in $\mathbb{E} \text{ [c.f.] } = f(\text{predictors})$
  - Motivates in-sample (IS) tests of “market efficiency”: IS return predictability = risk premium and/or behavioral bias

- But perfect knowledge of $f(.)$ implausible when #predictors is large

- We show: when investors learn about $f(.)$ in high-dimensional setting, equilibrium asset prices are such that
  - IS return predictability $\neq$ risk premium and/or behavioral bias
  - (Pseudo-)OOS return predictability = risk premium and/or behavioral bias

- Not about econometric problems with predictability tests, but about properties of equilibrium asset prices
Roadmap

Two steps:

1. Investors learn about parameters of cash-flow generating model and price assets accordingly

2. Econometrician analyzes equilibrium prices ex-post using standard return predictability tests
   - Properties of IS tests
   - Properties of OOS tests
1. Asset pricing
Setup

- Economy with $N$ assets, $N \times J$ firm characteristics $X$, $J < N$

- Dividends $y_t$ with growth cross-sectionally predictable based on $X$:
  \[ \Delta y_t = Xg + e_t, \quad e_t \sim N(0, I), \]
  with normalization $\frac{1}{NJ} \text{tr} X'X = 1$.

- Investors are homogeneous, risk-neutral, and the interest rate is zero.

- Dividend strips: $p_t = \text{prices at } t \text{ of claims to } y_{t+1}$
  - Think: one period $\approx$ one decade
Rational expectations: No return predictability

- RE: investors know value of $g$ in $\Delta y_t = Xg$

- Prices
  \[ p_t = \mathbb{E}_t y_{t+1} = y_t + Xg. \]

- Dividend strip returns
  \[ r_{t+1} = y_{t+1} - p_t = e_{t+1} \]
  i.e., unpredictable.

- This is the usual null hypothesis that underlies market efficiency tests, orthogonality conditions, Euler equations.

- But that investors have precise knowledge of $g$ is implausible, especially if $J$ is large $\Rightarrow$ Learning.
Bayesian pricing framework: Prior beliefs

- Before seeing data, investors hold informed prior beliefs

\[ \mathbf{g} \sim N \left( \mathbf{0}, \frac{\theta}{J} \mathbf{I} \right), \quad \theta > 0 \]

- Proportionality of prior covariance matrix to \( \mathbf{I} \): all the predictor variables on an equal footing from the prior perspective

- Variance of the elements of \( \mathbf{g} \) decline with \( J \): ensures that variance of predictable cash flow growth does not explode when \( N, J \to \infty \)

- Investors then learn about \( \mathbf{g} \) by observing \( \mathbf{X} \) and history \( \{ \Delta \mathbf{y}_s \}_{1}^{t} \), summarized by sample average \( \overline{\Delta \mathbf{y}}_t \).
Bayesian pricing framework: Posterior mean

- Posterior mean is a ridge regression estimator

\[ \tilde{g}_t = \Gamma_t (X'X)^{-1} X' \Delta y_t, \]

i.e., OLS estimator shrunk towards prior mean of zeros by matrix

\[ \Gamma_t = Q \left( I + \frac{J}{N\theta_t} \Lambda^{-1} \right)^{-1} Q' \]

where \( Q, \Lambda \) from eigendecomposition \( \frac{1}{N} X'X = Q\Lambda Q' \).

- Shrinkage strong
  - if \( t \) small (short time dimension)
  - if \( \theta \) small (prior tightly concentrated around zero)
  - if \( J/N \) is large (large \# of predictors)
  - along principal components of \( X'X \) with small eigenvalues
Equilibrium realized returns

Proposition

With assets priced based on \( \tilde{g}_t \), realized returns are

\[
    r_{t+1} = y_{t+1} - p_t = X(1 - \Gamma_t)g - X\Gamma_t(X'X)^{-1}X'e_t + e_{t+1}
\]

where \( \bar{e}_t = \frac{1}{t} \sum_{s=1}^{t} e_s \).

- "underreaction" to \( X \) due to shrinkage
- "overreaction" to estimation error in \( \tilde{g}_t \), dampened by shrinkage
- unpredictable shock (the only term in RE case)
2. Properties of return predictability tests
In-sample predictability test

- Econometrician cross-sectionally regresses (OLS)

\[ r_{t+1} = X(I - \Gamma_t)g - X\Gamma_t(X'X)^{-1}X'e_t + e_{t+1} \]

on characteristics matrix \( X \) and obtains coefficients

\[ h_{t+1} = (I - \Gamma_t)g - \Gamma_t(X'X)^{-1}X'e_t + (X'X)^{-1}X'e_{t+1}. \]

- Is econometrician likely to find \( h_{t+1} \) (jointly) significantly different from zero?

- Asymptotic analysis to study large-\( J \) case
  - Change in \( N, J \) alters not only sampling properties of econometric tests, but also investors’ learning problem and equilibrium asset prices
High-dimensional asymptotics

- Conventional low-dimensional asymptotics not suitable:
  \( J \text{ fixed, } N \to \infty \)
  \( \Rightarrow \) Investors could learn \( g \) perfectly in just one period

- Here we consider high-dimensional asymptotics:
  \( N, J \to \infty \), where \( J/N \to \psi > 0 \)
  \( \Rightarrow g \) still uncertain in investors’ minds even with large \( N \)

- Assumption
  The eigenvalues \( \lambda_j \) of \( \frac{1}{N}X'X \) satisfy \( \lambda_j > \varepsilon \) for all \( j \), where \( \varepsilon > 0 \) is a uniform constant as \( N \to \infty \).
  
  - “Big Data assumption” \( \approx \) additional predictors not redundant as \( J \to \infty \)

- \( g \) drawn from the prior distribution
In-sample predictability test: RE null

- Consider the return predictability test statistic

\[ T_{re} \equiv \frac{h_{t+1}X'Xh_{t+1} - J}{\sqrt{2J}}. \]

- Standard approach takes RE as null hypothesis, which implies

\[ h_{t+1} = (X'X)^{-1} X'e_{t+1} \]

and

\[ T_{re} \xrightarrow{d} N(0, 1) \text{ as } N, J \to \infty, J/N \to \psi > 0. \]

- Properties of test based on \( T_{re} \) when econometrician applies it to returns from the Bayesian learning economy where RE null false?
Proposition

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

Therefore,

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

Growth of second term with $\sqrt{J}$: \( \text{prob(rejection)} \to 1 \) rapidly as \( N, J \to \infty \)

\[ \text{prob(no rejection)} \to 0 \] exponentially fast as \( N, J \to \infty \)
Consider a characteristics-based trading strategy with weights

\[ w_{IS,t} = \frac{1}{N} X h_{t+1}, \quad r_{IS,t+1} = w'_{IS,t} r_{t+1} \]

which is IS because \( h_{t+1} \) estimated using returns \( r_{t+1} \).

Proposition

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

and the Sharpe ratio \( SR_{IS} = \mathbb{E} r_{IS,t+1} / \text{var}(r_{IS,t+1})^{1/2} \) grows at rate \( \sqrt{N} \).
Interpretation of in-sample return predictability tests

- IS return predictability tests lose their economic meaning when $J$ is not small relative to $N$: IS return predictability need not be consequence of risk premia or behavioral biases

- Not a statistical problem with the sampling properties of predictability test statistics, but a problem that the RE null hypothesis is false even without risk premia or behavioral biases: Investors’ learning of the cash-flow generating model parameters leaves in-sample predictable components in returns

- That there exists a “factor zoo” based on IS predictability evidence not surprising in high-dimensional setting

- That “factor zoo” emerged at same time when number of available predictors exploded may not be a coincidence
(Absence of) out-of-sample return predictability

- OOS trading strategy that weights $t + 1$ returns based on predictability coefficients estimated at time $s + 1$: 

$$ r_{OOS,t+1} = \mathbf{w}'_{OOS,s+1} \mathbf{r}_{t+1}, \quad \mathbf{w}_{OOS,s+1} = \frac{1}{N} \mathbf{X} \mathbf{h}_{s+1} $$

where $\mathbf{h}_{s+1}$ is estimated by regressing $\mathbf{r}_{s+1}$ on $\mathbf{X}$.

- Forward OOS: $t > s$
- Backward OOS: $t < s$

- Proposition

An out-of-sample trading strategy that uses coefficients estimated at time $s + 1$ has expected return $\mathbb{E} r_{OOS,t+1} = 0$ whenever $t \neq s$. 
(Absence of) out-of-sample return predictability

- **Forward** case is natural: Investors are Bayesian so the econometrician cannot “beat” investors in real-time return prediction.

- **Backwards** case $s < t$ is more surprising. Not a tradable strategy, but interesting for academic 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.

- **Caution**: forward result is likely a general property of Bayesian learning (with objectively correct prior), the backwards result might be somewhat specific to the environment here (e.g., IID cash-flow growth).
  - Generality of this result is an interesting question for future research.
3. Finite-sample analysis: Simulations
Finite-sample analysis: Simulations

- To generate data, we set $\theta = 1$ in
  \[
  \Sigma_g = \frac{\theta}{J} I
  \]

- $\theta =$ ratio of forecastable/residual cash-flow growth variance
  - Based on analyst expectations, Chen, Karceski, and Lakonishok (2003) find forecastable/residual cash-flow growth variance of 0.4 at 10yr horizon

- Simulate cash-flows, prices, returns for $N = 1000$ assets.

- Econometrician regresses $r_{T+1}$ on $X$ after investors have learned about $g$ for $T$ periods.
Adjusted $R^2$
Rejection probability of no-return-predictability null

![Graph showing rejection probability for different values of J and T.]
Sparsity

- Suppose prior about $\mathbf{g}$ not Normal, but Laplace,

$$f(g_j) = \frac{1}{2b} \exp \left(-\frac{|g_j|}{b}\right), \quad 2b^2 = \frac{\theta}{J}$$

and investors use maximum-a-posteriori (MAP) estimator to form $\tilde{\mathbf{g}}$

- Then some elements of $\tilde{\mathbf{g}}$ can be zero: **Sparsity**

- $\tilde{\mathbf{g}}$ are Lasso regression estimates

- Simulation: Again with objectively correct prior, now $\mathbf{g}$ drawn from Laplace distribution
Adjusted $R^2$
Rejection probability of no-return-predictability null
Excess shrinkage or sparsity

- Up to this point, shrinkage or sparsity was purely due to (objectively correct) informative prior beliefs of investors

- Possible reasons for additional sparsity:
  - Predictor variables costly to observe (change over time?)
  - Psychological cost of attention (Sims 2003; Gabaix 2014)

- To incorporate this we now allow prior to be tighter ($\theta = 0.5$) or more dispersed ($\theta = 2$) than distribution that we generate $g$ from ($\theta = 1$)
  - For both normal prior (ridge) and Laplace (lasso)

- Plots: $T = 4$. 
OOS portfolio return: Normal prior (ridge)
OOS portfolio return: Laplace prior (lasso)
4. Empirical illustration of IS vs. OOS predictability
Empirical illustration of wedge between IS and OOS predictability

- Focus on predictor variables that were available (in principle) already many decades ago: Past returns of each stock

- Predict returns in month $t$ with 238 predictors
  - Returns in months $t - 2, ... , t - 120$
  - Squared returns in months $t - 2, ... , t - 120$

- 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

Sample period: 1971-2018
In-sample: Past squared return coefficients

Sample period: 1971-2018
In-Sample and Out-of-Sample $R^2$ of rolling ridge regressions

Rolling regressions with 20-year estimation windows.
In-Sample and Out-of-Sample portfolio return based on rolling ridge regressions

Rolling regressions with 20-year estimation windows.
In-Sample and Out-of-Sample Predictability

<table>
<thead>
<tr>
<th></th>
<th>In-Sample (i)</th>
<th>Forward Out-of-Sample (ii)</th>
<th>Backward Out-of-Sample (iii)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A: ( R^2 )</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.54</td>
<td>0.17</td>
<td>0.25</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.28</td>
<td>0.78</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>Panel B: Portfolio return</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.88</td>
<td>0.32</td>
<td>0.31</td>
</tr>
<tr>
<td>S.D.</td>
<td>0.10</td>
<td>1.14</td>
<td>1.43</td>
</tr>
</tbody>
</table>
Implications: Market Efficiency in the Age of Big Data

- In Big Data setting, it’s implausible that investors know precisely the parameters of the cash-flow generating model.

- “Market efficiency” notion needs clarification: Equilibrium price properties are very different under RE (investors know $g$) and Bayesian learning (about $g$).

- Vast literature on return predictability (“factor zoo”) focuses on IS tests, but IS return predictability does not imply that risk premia/behavioral bias explanations are needed.

- Risk premia & bias theories should focus on explaining OOS return predictability.
  - Investor learning provides clear motivation for (pseudo-)OOS testing which is lacking in RE framework.

- Empirical challenge: characterize cross-section of stock returns OOS.