Asset Pricing and Machine Learning

Princeton Lectures in Finance
Lecture 2

Stefan Nagel[1]

[1]University of Chicago, NBER, CEPR, and CESIfo

May 2, 2019

Outline

1. More on ML techniques relevant for asset pricing
2. ML used by econometrician outside the market: SDF extraction in high-dimensional setting
3. ML used by investors inside the market: Rethinking market efficiency in the age of Big Data
   - Based on work-in-progress with Ian Martin
4. Conclusion: Agenda for further research

ML and financial market equilibrium

- **Now**: ML and the prediction problem of investors inside a financial market
- Real-world investors have to make predictions based on a huge set of potential predictor variables.
- Useful to think of investors as machine-learners?
- Implications for financial market equilibrium?
  - asset price dynamics
  - econometric testing of asset pricing models
  - search for anomalies, factors

Investor beliefs and econometric analysis

- Empirical asset pricing hypotheses typically involve orthogonality conditions

$$\mathbb{E}[(r_{t+1} - r_{b,t+1})x_t] = 0$$

  with some risk-appropriate benchmark return $r_{b,t+1}$ and time-t observable conditioning variables x_t.
  - e.g., market efficiency tests
- Let's abstract from risk pricing here: Assume r_{t+1} already adjusted for an appropriate benchmark return so that

$$\mathbb{E}[r_{t+1}x_t] = 0$$

- AP theory implies that the orthogonality conditions hold under investor expectations $\tilde{\mathbb{E}}[.]$

Investor beliefs: Learning

- How do investor expectations relate to estimates of expected values, $\mathbb{E}[.]$, by the econometrician studying data ex post?
- Much of literature: Rational expectations (RE) (here: investors know model & param. of DGP) so that $\tilde{\mathbb{E}}[.] = \mathbb{E}[.]$
  - LLN $\frac{1}{T}\sum_{t=1}^{T}[.] \to \mathbb{E}[.]$ allows econometrician to recover $\tilde{\mathbb{E}}[.] = \mathbb{E}[.]$ in empirical applications and test AP model
- But if investors learn about parameters/model from data: $\tilde{\mathbb{E}}[.]$ of investors $\neq \mathbb{E}[.]$ of econometrician
- Even in low-dimensional case, this changes how we should interpret asset price data
  - e.g., there will be in-sample return predictability, $\mathbb{E}[r_{t+1}x_t] \neq 0$, even if $\tilde{\mathbb{E}}[r_{t+1}x_t] = 0$ (e.g., Lewellen and Shanken 2002)
- Does high-dimensionality make this problem "worse"?

Investor learning in the age of Big Data

- What do investors learn about? Realistically, enormous (and expanding!) set of potentially relevant variables for pricing of stocks. High-dimensional!
  - Existing learning models look at very low-dimensional learning problem
- Key lesson from lecture 1: In high-dimensional setting shrinkage/variable selection crucial to obtain good forecasts

$\Rightarrow$ Investors must use prior knowledge about models/parameters to shrink/select variables

- What are the consequences of learning from high-dimensional data with shrinkage/selection for observed asset prices?

Example: Learning about stock fundamentals

- Simple example before laying out more general framework
- Cross-section of N assets with payoffs (dividends)

$$\boldsymbol{y}_t = b_1 \boldsymbol{x}_1 + b_2 \boldsymbol{x}_2 + \boldsymbol{e}_t, \qquad \boldsymbol{e}_t \sim IID$$

  with two firm characteristics $\boldsymbol{x}_1, \boldsymbol{x}_2$, where $\boldsymbol{x}_1' \boldsymbol{x}_2 = 0$.
- Risk-neutral investors learn from $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_t\}$ about $\boldsymbol{b} = (b_1, b_2)'$ and use to forecast $\boldsymbol{y}_{t+1}$
- Prices of claims at t to single next period dividends in $t+1$ ("dividend strips")

$$\boldsymbol{p}_t = \hat{b}_1 \boldsymbol{x}_1 + \hat{b}_2 \boldsymbol{x}_2$$

Example: Learning about stock fundamentals

- For now just suppose that investors use variable selection method that yields $\hat{b}_2 = 0$, $\hat{b}_1 \neq 0$.
  - Unlikely optimal with just two explanatory variables, but in more realistic high-dimensional case e.g. half of all coefficients may be set to zero
- Price

$$\boldsymbol{p}_t = \hat{b}_1 \boldsymbol{x}_1$$

- Subsequent realized return

$$\begin{aligned} \boldsymbol{r}_{t+1} &= \boldsymbol{y}_{t+1} - \boldsymbol{p}_t \\ &= (b_1 - \hat{b}_1)\boldsymbol{x}_1 + b_2 \boldsymbol{x}_2 + \boldsymbol{e}_{t+1} \end{aligned}$$

Example: Learning about stock fundamentals

- Consider an econometrician observing $\boldsymbol{r}_{t+1}$ ex-post and looking for in-sample predictability using $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ as predictors.
- Two sources of in-sample return predictability in

$$\boldsymbol{r}_{t+1} = (b_1 - \hat{b}_1)\boldsymbol{x}_1 + b_2\boldsymbol{x}_2 + \boldsymbol{e}_{t+1}$$

  1. Variable selection induces presence of $b_2\boldsymbol{x}_2$
  2. but $|b_1 - \hat{b}_1|$ should be smaller than it would be without variable selection

- How does this work out when investors use optimal shrinkage/variable selection?

Generalizing the framework

- Homogeneous risk-neutral Bayesian investors
- High-dimensional setting with 1000s of variables
- We explore different priors that induce shrinkage and variable selection
- Study properties of typical asset pricing tests (return predictability): in-sample (IS) and out-of-sample (OOS)

Generalizing the framework

- N risky assets. Risk-free rate normalized to zero.
- $N \times J$ matrix of firm characteristics $\boldsymbol{X}_t$, $J \leq N$
- Each period, assets pay dividends, $\boldsymbol{y}_t$, where

$$\begin{aligned}\Delta \boldsymbol{y}_t &= \boldsymbol{y}_t - \boldsymbol{y}_{t-1} \\ &= \boldsymbol{X}_{t-1}\boldsymbol{g} + \boldsymbol{e}_t, \quad \boldsymbol{e}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_e), \quad \boldsymbol{\Sigma}_e = \boldsymbol{I}\end{aligned}$$

- We focus on pricing of one-period dividend strips: time-t claims to single-period dividends $\boldsymbol{y}_{t+1}$
- Think of one period here as roughly the typical duration of a stock's cash flow (e.g., perhaps a decade).

RE benchmark case

- Rational expectations (RE) equilibrium in which investors know $\boldsymbol{g}$:

$$\boldsymbol{p}_t = \boldsymbol{y}_t + \boldsymbol{X}_t\boldsymbol{g}, \qquad \boldsymbol{r}_{t+1} = \boldsymbol{y}_{t+1} - \boldsymbol{p}_t = \boldsymbol{e}_{t+1}$$

- RE implies orthogonality conditions

$$\mathbb{E}\left[\boldsymbol{r}_{t+1} \otimes \boldsymbol{X}_t\right] = 0$$

- We focus on realistic case where investors don't know $\boldsymbol{g}$: they learn about it from joint history $\{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_t\}$ and $\{\boldsymbol{X}_0, \boldsymbol{X}_1, ..., \boldsymbol{X}_{t-1}\}$.

Investors' prior beliefs

- We assume investors know $\boldsymbol{\Sigma}_e = \boldsymbol{I}$
- Before seeing data, investors hold prior beliefs

$$\boldsymbol{g} \sim N(0, \sigma_g^2 \boldsymbol{I})$$

- Number of variables J potentially very large $\Rightarrow$ OLS (i.e., diffuse prior) would not yield useful forecasts
  - More serious problem in more recent years: Technological change has increased number of potential predictors enormously
- Diffuse prior would not be an economically plausible assumption anyway: Predictable variation in dividends should be limited, i.e., extreme values of $\boldsymbol{g}$ unlikely

Investors' posterior beliefs

- Given observations $\Delta \boldsymbol{y}_1$ realized in $t = 1$, the posterior of $\boldsymbol{g}$ is

$$\boldsymbol{g} | \Delta \boldsymbol{y}_1 \sim N(\boldsymbol{D}_1 \boldsymbol{d}_1, \boldsymbol{D}_1)$$

where

$$\boldsymbol{D}_1 = \left(\sigma_g^{-2} \boldsymbol{I} + \boldsymbol{X}_0' \boldsymbol{X}_0\right)^{-1}$$
$$\boldsymbol{d}_1 = \boldsymbol{X}_0' \Delta \boldsymbol{y}_1$$

- Posterior mean is ridge regression estimator

$$\hat{\boldsymbol{g}}_1 = \boldsymbol{D}_1 \boldsymbol{d}_1 = \left(\sigma_g^{-2} \boldsymbol{I} + \boldsymbol{X}_0' \boldsymbol{X}_0\right)^{-1} \boldsymbol{X}_0' \Delta \boldsymbol{y}_1$$

where $\sigma_g^{-2} \boldsymbol{I}$ dominates for small prior variances and disappears with diffuse prior.

Specializing the setup

- We assume predictors are already orthogonalized

$$\boldsymbol{X}_t = \boldsymbol{U}\boldsymbol{S}_t, \quad \text{where } \boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I} \text{ and } \boldsymbol{S}_t \text{ diagonal}$$

- Then,

$$\boldsymbol{X}_t'\boldsymbol{X}_t = \boldsymbol{S}_t^2$$

  is diagonal with s_j^2, the eigenvalues of $\boldsymbol{X}_t'\boldsymbol{X}_t$, on the diagonal.
- For orthogonalized predictors, the distribution of these eigenvalues captures the predictors' empirical properties
  - affects fragility of estimation and prediction
  - determines effects of ridge regression shrinkage
- Example: two variables almost collinear before orthogonalizing $\Rightarrow$ one very low eigenvalue component after orthogonalizing

Diagonal elements of $\boldsymbol{S}_t^2$

- We pick s_j^2 from a geometrically declining sequence,

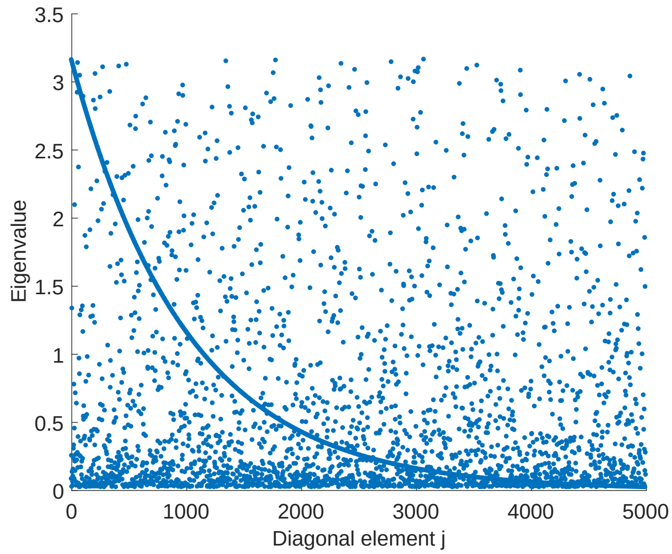
$$\lambda_j = \sqrt{\delta^j \frac{1-\delta}{\delta} N}$$

  partly in order $s_j^2 = \lambda_j^2$, partly randomly permuted each period

$\Rightarrow$ As in typical empirical stock characteristics data, when picking J variables
  - More likely to capture high-eigenvalue predictors when J small
  - But always some low-eigenvalue predictors sprinkled in (i.e., pre-orthogonalization, some predictors close to collinearity)
  - Some chance that predictor associated with low eigenvalue this period will have higher eigenvalue next period (makes OOS prediction fragile)

Example: Diagonal elements of $\boldsymbol{S}_t^2$

Example for $J = N = 5000$:



Shrinkage

- After observing data for t periods, stacked into

$$\Delta \boldsymbol{y}_{1:t} = (\Delta \boldsymbol{y}_1', \Delta \boldsymbol{y}_2', ..., \Delta \boldsymbol{y}_t')'$$
$$\boldsymbol{X}_{0:t-1} = (\boldsymbol{X}_0', \boldsymbol{X}_1', ..., \boldsymbol{X}_{t-1}')'$$

- We can rewrite the posterior mean

$$\hat{\boldsymbol{g}}_t = \left(\frac{1}{\sigma_g^2} \boldsymbol{I} + \boldsymbol{X}_{0:t-1}' \boldsymbol{X}_{0:t-1} \right)^{-1} \boldsymbol{X}_{0:t-1}' \Delta \boldsymbol{y}_{1:t}$$

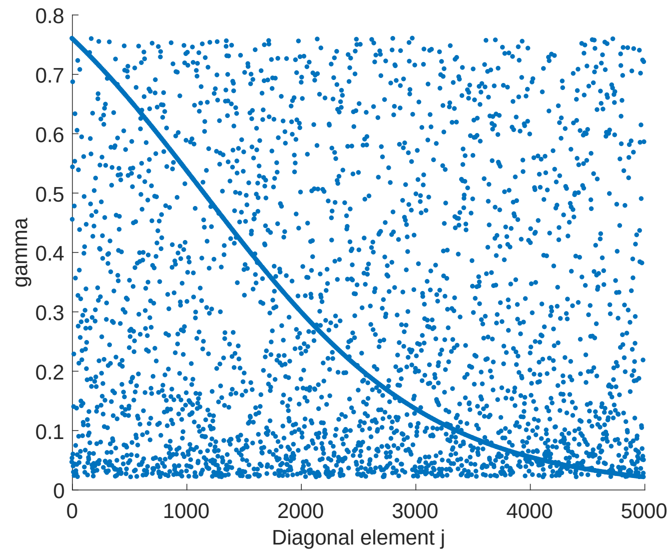
as

$$\hat{\boldsymbol{g}}_t = \boldsymbol{\Gamma}_t \boldsymbol{g} + \boldsymbol{\Gamma}_t (\boldsymbol{X}_{0:t-1}' \boldsymbol{X}_{0:t-1})^{-1} \boldsymbol{X}_{0:t-1}' \boldsymbol{e}_{1:t}$$

- The shrinkage matrix $\boldsymbol{\Gamma}_t$ is diagonal with elements $0 < \gamma_j < 1$
  - introduces estimation error related to $\boldsymbol{g}$
  - in order to reduce the estimation error resulting from $\boldsymbol{e}_{1:t}$.

Diagonal elements of shrinkage matrix Γ_t

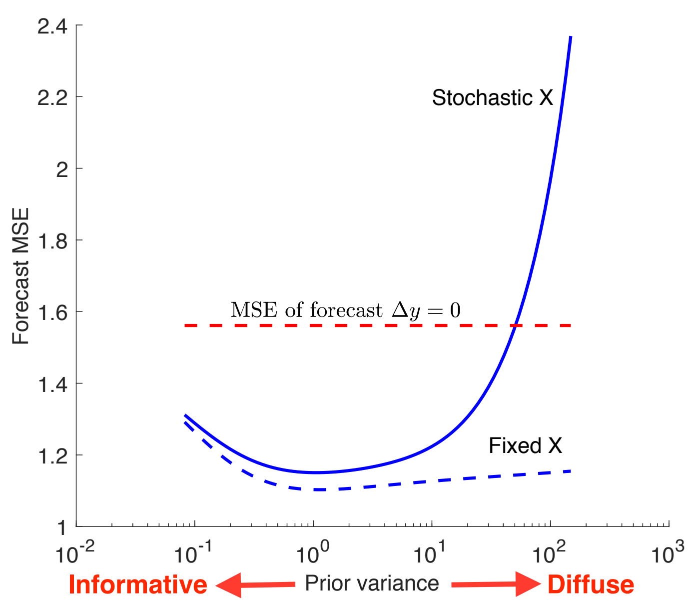
Example for $J = N = 5000$:



$\Rightarrow$ shrinkage strong (γ_j low) for low-eigenvalue components of $\boldsymbol{X}$ (s_j^2 low)

Shrinkage important for forecast performance

MSE in forecasting $\Delta \boldsymbol{y}_{t+1}$ with $\boldsymbol{X}_t \hat{\boldsymbol{g}}_t$ as function of prior variance (data generated with fixed $\sigma_g^2 = 1$):



$\Rightarrow$ Diffuse prior (OLS) yields worse MSE than forecast $\Delta y_{t+1} = 0$.

Prices and returns

- Investors price the assets based on their posterior mean

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

- Realized returns of single-period dividend strips then follow as

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

- Evaluating, we obtain

$$\begin{aligned} \boldsymbol{r}_{t+1} = {} & \boldsymbol{X}_t(\boldsymbol{I} - \boldsymbol{\Gamma}_t)\boldsymbol{g} \quad \text{Shrinkage effect} \\ & - \boldsymbol{X}_t \boldsymbol{\Gamma}_t (\boldsymbol{X}'_{0:t-1}\boldsymbol{X}_{0:t-1})^{-1}\boldsymbol{X}'_{0:t-1}\boldsymbol{e}_{1:t} \quad \text{Learning effect} \\ & + \boldsymbol{e}_{t+1} \quad \text{Unforecastable error} \end{aligned}$$

- Under Bayesian investors' posterior beliefs, all three terms have expected value zero.

In-sample return predictability tests

- Econometrician analyzes sample of returns to test RE null hypothesis of no return predictability

$$H_0 : \boldsymbol{p}_t = \boldsymbol{y}_t + \boldsymbol{X}_t \boldsymbol{g} \quad \Rightarrow \quad \boldsymbol{r}_{t+1} = \boldsymbol{e}_{t+1}$$

- Cross-sectional regression of $\boldsymbol{r}_{t+1}$ on $\boldsymbol{X}_{K,t}$, the first $K \leq J$ columns of $\boldsymbol{X}_t$, yields coefficients

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

- Under H_0,

$$\sqrt{N}\boldsymbol{h}_{t+1} \sim N(0, N\boldsymbol{\Omega}) \qquad \text{where} \quad \boldsymbol{\Omega} = (\boldsymbol{X}'_{K,t}\boldsymbol{X}_{K,t})^{-1}$$

  and

$$\boldsymbol{h}'_{t+1}\boldsymbol{\Omega}^{-1}\boldsymbol{h}_{t+1} \sim \chi^2_K$$

- Is econometrician going to find predictive regression coefficients in $\boldsymbol{h}_{t+1}$ jointly/individually "significant"?

In-sample return predictability tests

- Evaluating $\boldsymbol{h}_{t+1}$, we obtain

$$\boldsymbol{h}_{t+1} = (\boldsymbol{I} - \boldsymbol{\Gamma}_{K,t})\boldsymbol{g}_K \quad \text{Shrinkage effect}$$
$$- \boldsymbol{H}_{K,t}\boldsymbol{\Gamma}_{K,t}\left(\boldsymbol{X}'_{K,0:t-1}\boldsymbol{X}_{K,0:t-1}\right)^{-1}\boldsymbol{X}'_{K,0:t-1}\boldsymbol{e}_{1:t} \quad \text{Learning effect}$$
$$+ \left(\boldsymbol{X}'_{K,t}\boldsymbol{X}_{K,t}\right)^{-1}\boldsymbol{X}'_{K,t}\boldsymbol{e}_{t+1} \quad \text{Estimation error}$$

  where $\boldsymbol{H}_{K,t} = \boldsymbol{I} + \boldsymbol{S}_{K,t-1}\boldsymbol{S}_{K,t}^{-1} + ... + \boldsymbol{S}_{K,0}\boldsymbol{S}_{K,t}^{-1}$.
- Under the RE hypothesis H_0 the first two terms are exactly zero, leading to the standard OLS variance
- But with learning, the first two components are not zero $\Rightarrow$ with $\boldsymbol{g}$ drawn from the prior distribution,

$$\boldsymbol{h}'_{t+1}\boldsymbol{\Omega}^{-1}\boldsymbol{h}_{t+1} \sim \text{a (complicated) weighted sum of } \chi_1^2 \text{ r.v}$$

  i.e., not χ_K^2!

Simulations

- We simulate

$$\Delta\boldsymbol{y}_{t+1} = \boldsymbol{X}_t\boldsymbol{g} + \boldsymbol{e}_{t+1}$$

- Parameters
  - Eigenvalues of $\boldsymbol{X}'_t\boldsymbol{X}_t$: as in earlier plot, fraction $q = \frac{1}{2}$ of columns randomly permuted
  - Number of stocks: $N = 5000$
  - Number of predictor variables: $J = 1$ to N
  - Number of predictors available to econometrician: $K = J$
  - Prior variance: $\sigma_g^2 = 1$
- Prior variance assumption implies ratio of maximum forecastable to residual variance of $\Delta\boldsymbol{y}_{t+1}$ of

$$\frac{\frac{1}{N}\operatorname{tr}(\boldsymbol{X}'_{0:t-1}\boldsymbol{X}_{0:t-1})}{1} \approx 1$$

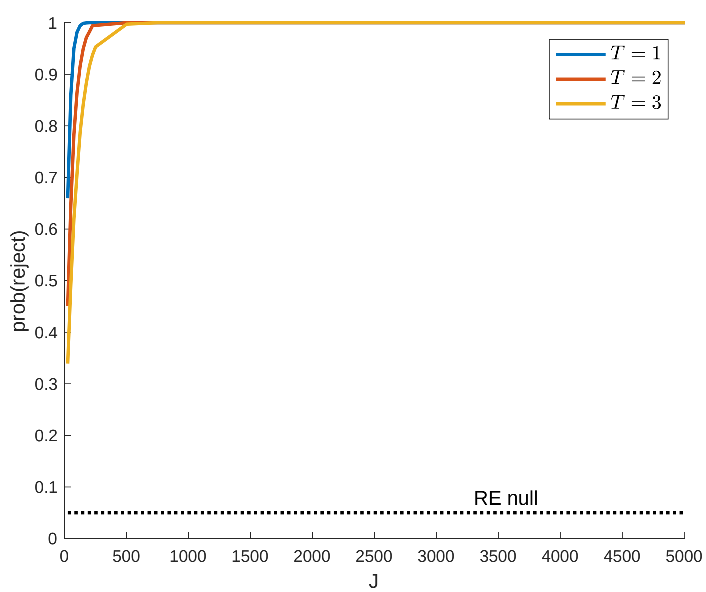
  which is important for mapping length of one time period, learning speed, to empirical data.

Simulations: Interpretation of time period length

- With persistent forecastable component and IID residual, the ratio of maximum forecastable to residual variance falls over time
- Evidence in Chan, Karceski, and Lakonishok (2003) based on a number of firm revenue and profit growth measures: horizon $>$ 10 years required for this ratio to exceed unity.
  - Predictable growth based on IBES analyst forecasts as predictors as lower bound for maximum forecastable variance

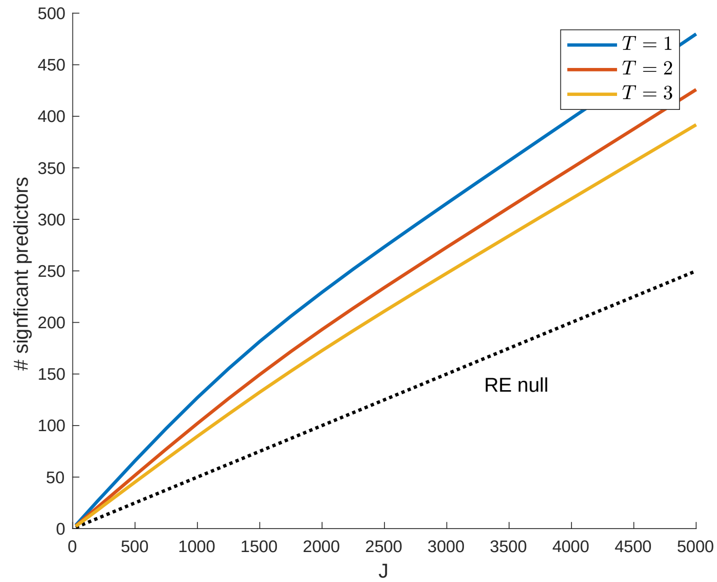
$\Rightarrow$ Think of one period in the model as approximately a decade

In-sample predictability: Joint test ($p < 0.05$)



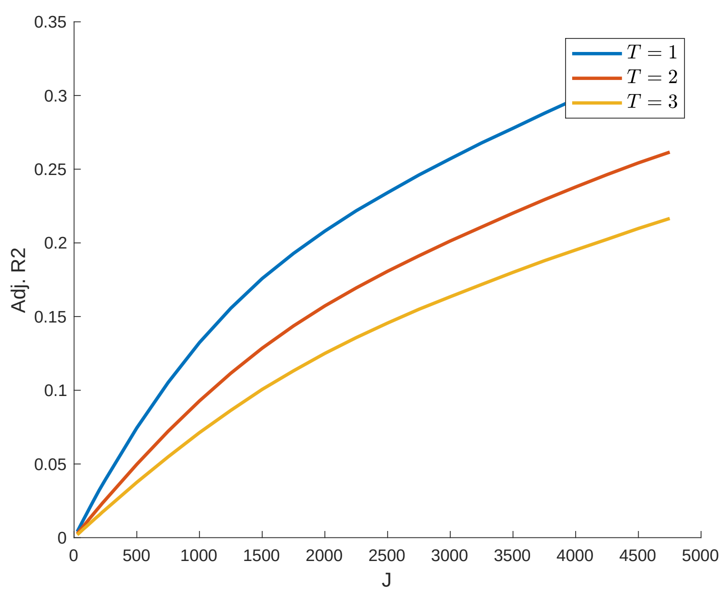
$\Rightarrow$ Almost certain to reject H_0 as soon as J moderately high

In-sample predictability: Number of individually "significant" factors ($p < 0.05$)



$\Rightarrow$ % of characteristics for which H_0 rejected is much higher than nominal test size

In-sample predictability: Joint adj. R^2



Out-of-sample predictability

- How could we test the learning-with-shrinkage hypothesis? Out-of-sample tests?
- Econometrician's return forecast based on on period t regression coefficient: $\boldsymbol{X}_{K,t}\boldsymbol{h}_t$
- OOS investment strategy with weights based on this return forecast

$$\boldsymbol{w}_t = \frac{1}{\sqrt{K}}\boldsymbol{X}_{K,t}\boldsymbol{h}_t$$

- Realized return in OOS period

$$\begin{aligned}\boldsymbol{r}'_{t+1}\boldsymbol{w}_t = &\ \boldsymbol{g}'(\boldsymbol{I}-\boldsymbol{\Gamma}_t)\boldsymbol{X}'_t\boldsymbol{w}_t \\ &- \boldsymbol{e}'_{1:t}\boldsymbol{X}_{0:t-1}(\boldsymbol{X}'_{0:t-1}\boldsymbol{X}_{0:t-1})^{-1}\boldsymbol{\Gamma}_t\boldsymbol{X}'_{0:t-1}\boldsymbol{w}_t \\ &+ \boldsymbol{e}'_{t+1}\boldsymbol{w}_t\end{aligned}$$

Out-of-sample predictability

- With $\boldsymbol{g}$ drawn from the prior distribution, one can show

$$\mathbb{E}[\boldsymbol{r}'_{t+1}\boldsymbol{w}_t] = 0$$

  because Bayesian shrinkage exactly balances the effects on OOS predictability of first and second terms in $\boldsymbol{r}'_{t+1}\boldsymbol{w}_t$ expression.
- But still, there is a catch: while the two terms cancel out in expectation, they don't cancel in a given sample for a given draw of $\boldsymbol{g}$ and $\boldsymbol{e}_{1:t}$ $\Rightarrow$ effect on sampling variance of OOS return

Out-of-sample predictability

- Recall that one time period here is meant to be long ($\approx$ a decade), so think of the OOS evaluation period $t+1$ as one decade
- Standard way to assess statistical significance would be to use intra-period, e.g., $m = 120$ monthly returns.
- For intra-period returns we have

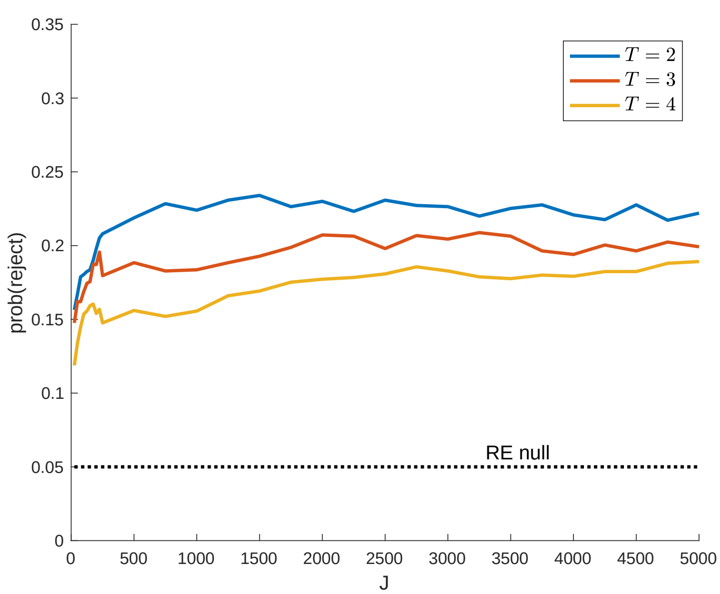
$$\boldsymbol{r}'_{t+\tau}\boldsymbol{w}_t \approx \boldsymbol{e}'_{t+\tau}\boldsymbol{w}_t$$

  with $\{\tau\} = \{1/m, 2/m, ..., 1\}$, because intra-period, $\hat{\boldsymbol{g}} - \boldsymbol{g}$ (reflecting $\boldsymbol{g}$, $\boldsymbol{\Gamma}_t$, and $\boldsymbol{e}_{1:t}$) is approximately constant.
- Econometrician estimates portfolio return variance

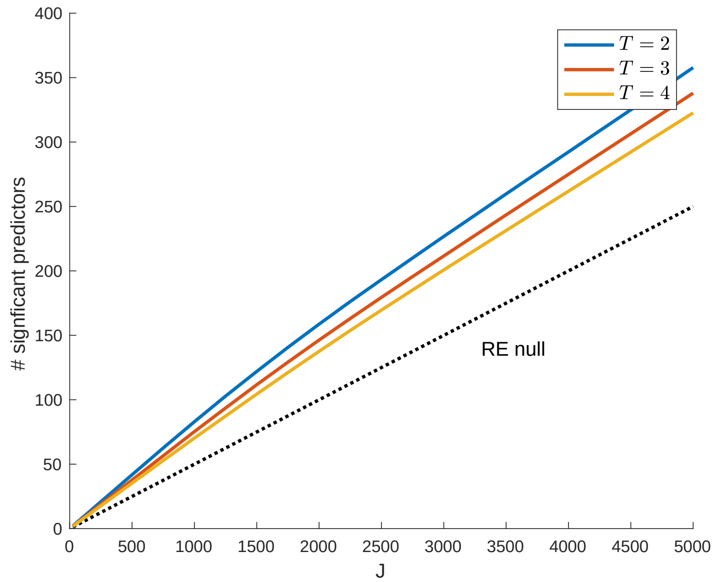
$$\text{var}\left(\boldsymbol{w}'_t\boldsymbol{r}_{t+1}|\boldsymbol{w}_t,\right) \approx \text{var}\left(\boldsymbol{w}'_t\boldsymbol{e}_{t+1}|\boldsymbol{w}_t,\right) = \boldsymbol{h}'_t\boldsymbol{S}^2_K\boldsymbol{h}_t$$

- But actual sampling variance is higher because
  - terms involving $\boldsymbol{g}$, $\boldsymbol{e}_{1:t}$ don't perfectly balance
  - distribution is non-normal (involves cross-products between normal and squared normal r.v.)

Out-of-sample predictability: Joint test ($p < 0.05$)



Out-of-sample predictability: Number of individually "significant" factors ($p < 0.05$)



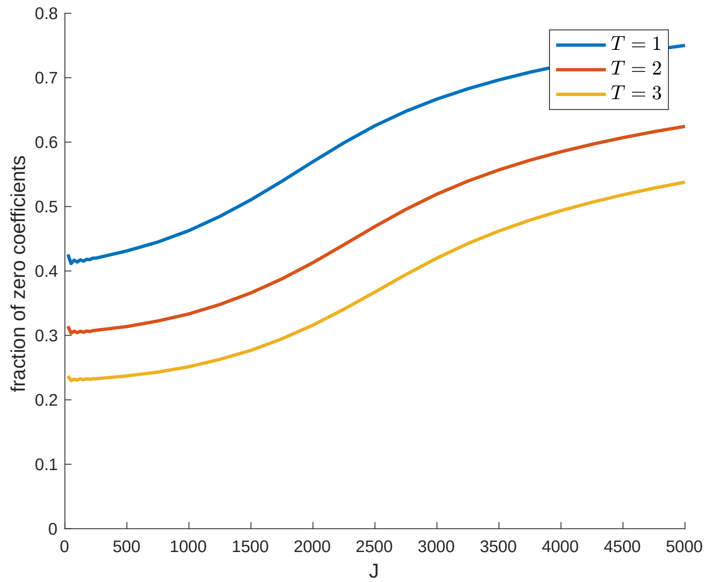
$\Rightarrow$ Lower than in-sample, but only by a bit (about 2/3 of number of in-sample significant factors)

Sparsity

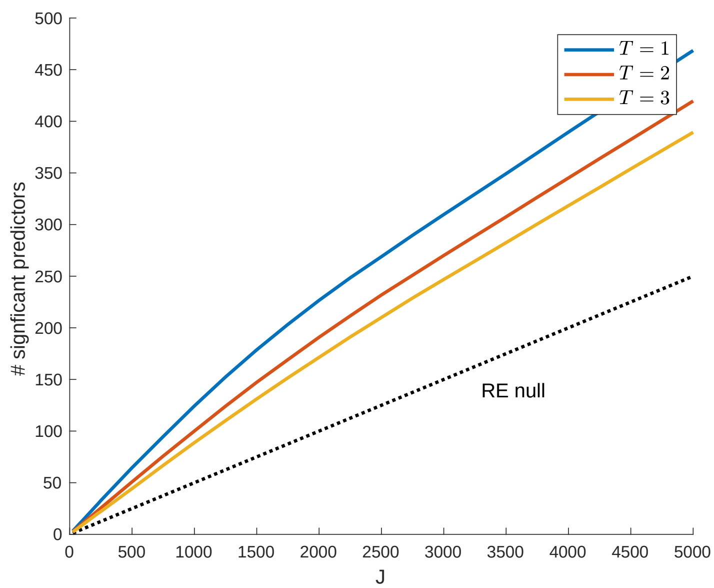
- So far we studied learning with shrinkage. What about sparsity, variable selection?
- Similar effects with priors that induce sparsity: $\boldsymbol{g} \sim$ Laplace

$\Rightarrow$ Investors use Lasso to estimate $\boldsymbol{g}$ in $\Delta \boldsymbol{y}_t = \boldsymbol{X}_{t-1}\boldsymbol{g} + \boldsymbol{e}_t$.

Sparsity: Number of coefficients set to zero by Lasso

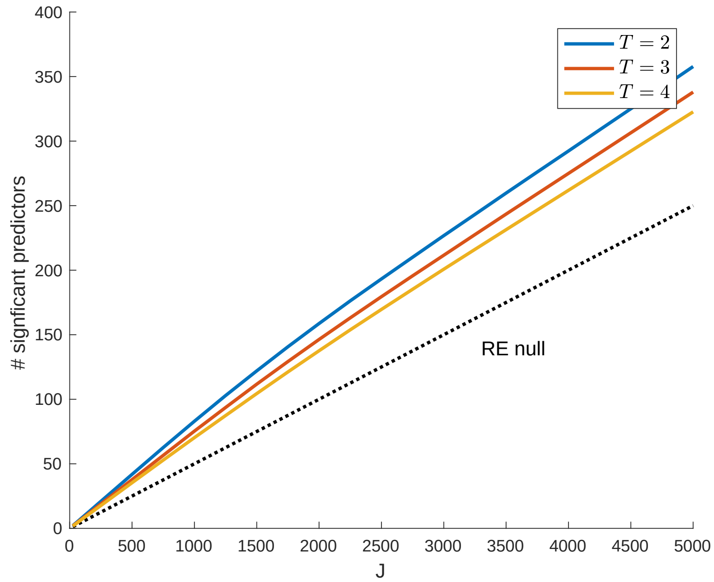


In-sample predictability: Number of individually "significant" factors ($p < 0.05$) with Lasso



$\Rightarrow$ Very similar to normal prior/ridge regression case

Out-of-sample predictability: Number of individually "significant" factors ($p < 0.05$) with Lasso



$\Rightarrow$ Very similar to normal prior/ridge regression case

Market efficiency in the age of Big Data: Summary

- High-dimensionality of fundamentals predictor space magnifies learning effects in the cross-section of stock returns
  - More likely to reject no-predictability null IS and OOS
  - More likely to find factors with "significant" abnormal returns IS and OOS
- Documenting a new "significant" factor, anomaly, becomes "less interesting" in high-dimensional setting—even without data mining, multiple testing problems.
- Analysis of high-dimensional case underscores that market efficiency (ME) is a fuzzy concept:
  - Does ME mean investors have RE with DGP parameters known? (Underlying assumption of most ME tests)
  - Does ME mean investors are Bayesian learners? (We don't have generic testing approaches for this version)
- Open question: Adjustment to test statistics so that we can test the learning hypothesis in a generic way?

Outline

1. More on ML techniques relevant for asset pricing
2. ML used by econometrician outside the market: SDF extraction in high-dimensional setting
3. ML used by investors inside the market: Rethinking market efficiency in the age of Big Data
4. Conclusion: Agenda for further research

Agenda for further research : ML in AP

- ML methods seem well-suited to address to address needs of
  - econometrician studying AP data ex post
  - investors learning from high-dimensional data in real time
- **To do:** Further work on priors
  - Given low signal-to-noise ratio in AP, prior knowledge more important than in other ML applications ⇒ important to fuse ML methods with economic restrictions
  - Example from lecture 1: Prior based on absence of near-arbitrage and concentration of factor premia
  - Other potentially useful avenues:
    - priors on heterogeneity of limits to "arbitrage," short-sale constraints, ...
    - priors tilted towards risk premia implied by structural economic models

Agenda for further research : ML as a tool for the econometrician in AP

- By now clear that using low-dimensional characteristics-sparse factor models (e.g., FF 5-factor) as
  - representation of investment opportunity set
  - benchmark for abnormal return measurement (e.g., for newly proposed anomaly, factor)

  is not appropriate anymore $\Rightarrow$ ML methods should become standard part of toolkit
- **To do:** Allow for drift in parameters, moments, penalties
  - Asset return moments change over time as investors learn, the economy evolves, arbitrageurs trade
  - Potentially promising: fused lasso, fused ridge regression
- **To do:** Connect SDF extracted with ML methods back to economic models of financial markets
  - correlate ML-based SDF with macro, sentiment variables?

Agenda for further research : ML as approximation of investor learning in AP models

- Thinking of investors as forecasting using ML tools seems appropriate, given the arguably high-dimensional problem faced by real-world investors
- **To do:** The setting we have considered makes learning in many ways still too easy. Would be realistic to add
  - Uncertainty about second moments
  - Time-varying parameters
  - Additional costs of model complexity
  - Model robustness concerns
  - Risk premia

Agenda for further research : ML as model of investor learning

- **To do:** Introducing investor heterogeneity in a high-dimensional setting, e.g.,
  - some investors learn from the fundamentals history and forecast fundamentals
  - some investors learn from the return history and forecast returns
  - (plus perhaps some investors that misinterpret data)

  could lead to additional interesting cross-sectional predictions.
- **To do:** Learning from return history (a high-dimensional object) could also be source of interesting dynamics
- **To do:** Dynamic process of anomaly discovery (and elimination by arbitrageurs) in high-dimensional setting