Machine Learning in Asset Pricing Chapter 3.1-3.3

Presented by: Steven Urry

02/17/2022

Motivation for MLAP

- Problem of high dimensionality
- Most research to date imposes ad hoc sparsity
  - Evaluate individual or small number of variables marginal predictive information relative to a standard set of variables (e.g. FF 5-factor model)
  - Leads to redundancy in the literature
  - No theoretic motivation for ad hoc sparsity
- ML can handle this high dimensionality problem
  - However, ML was not built for finance so we need to tread carefully

Table 3.1
Differences between typical ML and asset pricing applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

TABLE 3.1
Differences between typical ML and asset pricing applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

- We can't observe $\mathbf{E}_t[r_{t+1}]$ rather we observe and train our models on past realized returns (a noisy signal for $\mathbf{E}_t[r_{t+1}]$)
- Additionally, $\mathbf{E}_t[r_{t+1}]$ explains only a small portion of the total cross-sectional variance in returns
- Overall, these cause the SN ratio to be low and the lower the SN ratio the harder it is to tease out the signal and generate consistent predictions

TABLE 3.1
Differences between typical ML and asset pricing applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

- One way to increase the SN ratio is to increase the amount of data we have
- However, there is relatively little data in finance
  - Monthly data back to 1960 would amount to ~2.1 million observations
  - Some natural language processing algorithms use >100s of billions of observations
- Additionally, we are limited to the frequency with which our signals change (e.g. earnings data changes 4 times a year)
- Ideally we get the most granular signals possible, but then we run into microstructure issues

TABLE 3.1
Differences between typical ML and asset pricing applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

- Do optimal estimates for individual stock returns result in optimal portfolios?
  - Open question with no clear answer (Main discussion today)

TABLE 3.1
Differences between typical ML and asset pricing applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

- In asset pricing, prediction error covariances drive the volatility of our portfolio (i.e. covariances matter)
  - This has implications for which ML methods we use, how to regularize it, how to evaluate the algorithms performance, and how to use its output in portfolio construction (next week's discussion)
- Today we will assume 0 or approximately 0 covariance of cross-sectional returns

TABLE 3.1
Differences between typical ML and asset pricing applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

- In many ML applications there are clear prior reasons to exclude certain irrelevant elements
  - Geneticists know which genes are relevant when studying a specific disease
- In Finance it is not clear, a priori, which of the many predictors are irrelevant especially since models typically assume exactly 0 relevance

TABLE 3.1
Differences between typical ML and asset pricing applications

	Typical ML Application	Asset Pricing
Signal-to-noise	High	Very low
Data dimensions	Many predictors, Many observations	Many predictors Few observations
Aggregation level of interest	Individual outcomes	Portfolio outcomes
Prediction error covariances	Statistical nuisance	Important determinant of portfolio risk
Sparsity	Often sparse	Unclear
Structural change	None	Investors learn from data and adapt

- Fortunately, natural laws do not change by our discovering them, in most applications the data generation process is stationary
    - Therefore, the ML literature contains little on how to handle structural changes over time
- However, in finance, what investors know is key to the data generation process (DGP), thus as new data or strategies become available the DGP will change

Simple Example

- The goal of this next exercise is to highlight the issues MLAP faces and potential solutions

- General return prediction problem

$$\mathbb{E}[r_{i,t+1}|x_{i,t}] = f(x_{i,t}) \tag{3.1}$$

Example: Setup

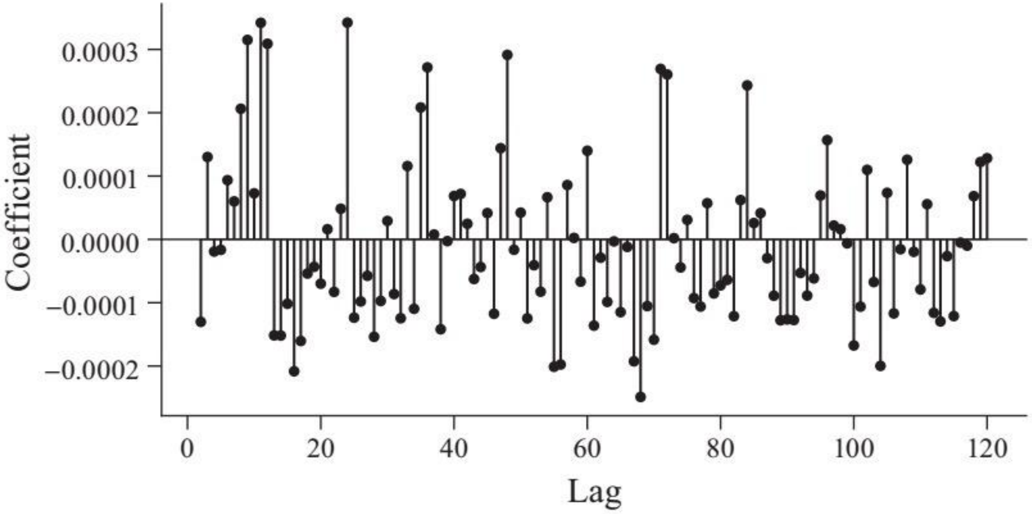
- In this case *x* is a matrix of 120 months of lagged returns for each stock, their square and their cube
- Monthly data from 1970-2019, demean data monthly, **normalize predictors**, and give equal weight to each month in regression
- Regression with 357 predictors:

$$r_{i,t+1} = \sum_{k=1}^{119} b_k r_{i,t-k} + \sum_{k=1}^{119} c_k r_{i,t-k}^2 + \sum_{k=1}^{119} d_k r_{i,t-k}^3 + e_{i,t+1}, \qquad (3.2)$$

Example: Cross Validation

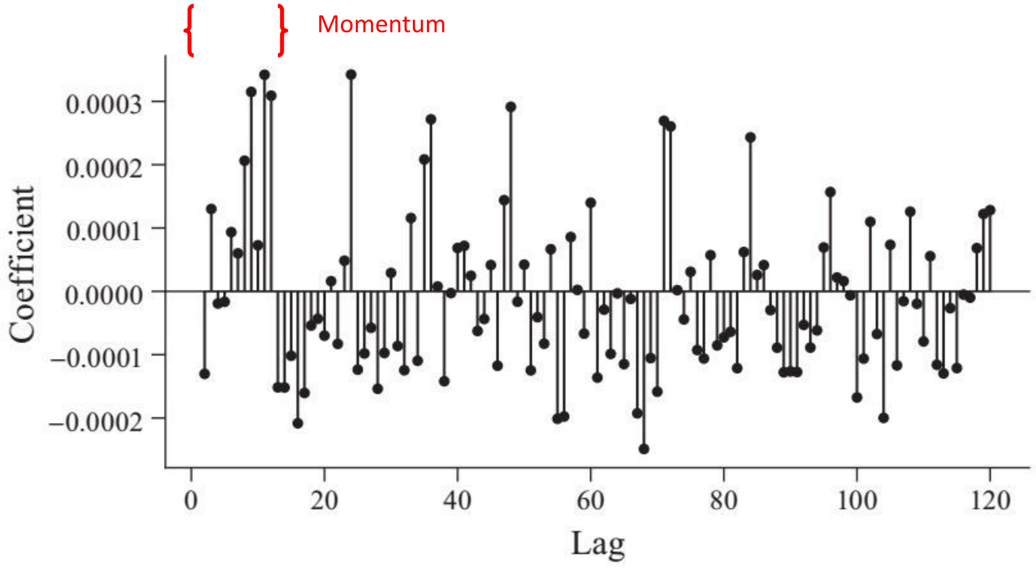
- Leave-one-year-out CV
  - Take one year and set it aside call it Y
  - Run eq. 3.2 for all other years
  - Use these estimates to get expected returns and R^2 for Y
  - Repeat for all years
  - Average R^2 for all set aside years
  - Find ridge penalty value that maximizes the CV R^2

Example: Results



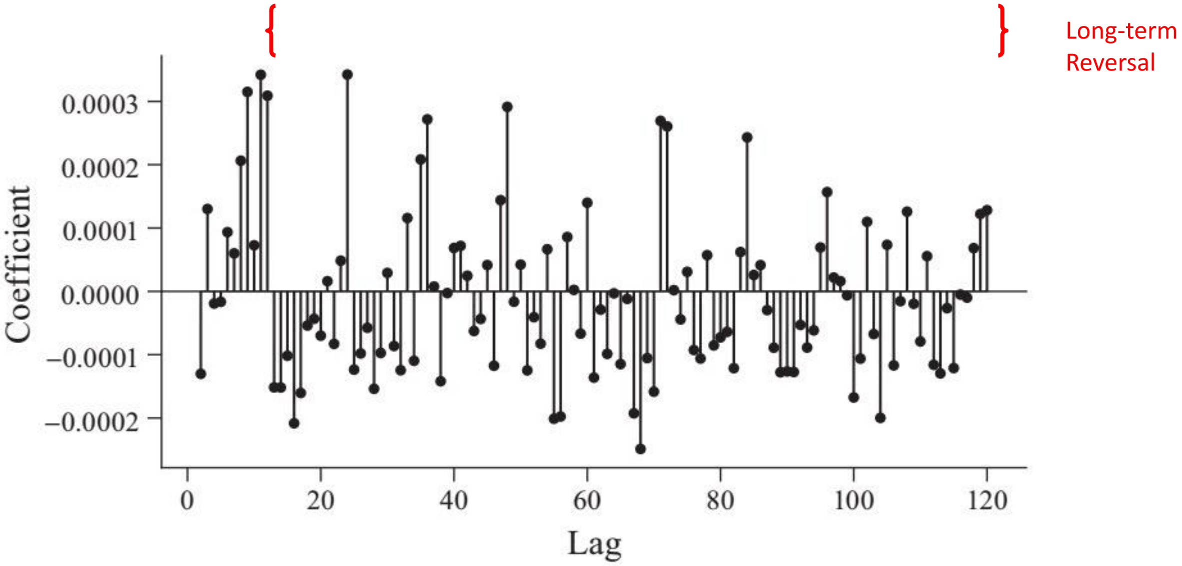
Example: Results

- Several well known anomalies found with a single regression



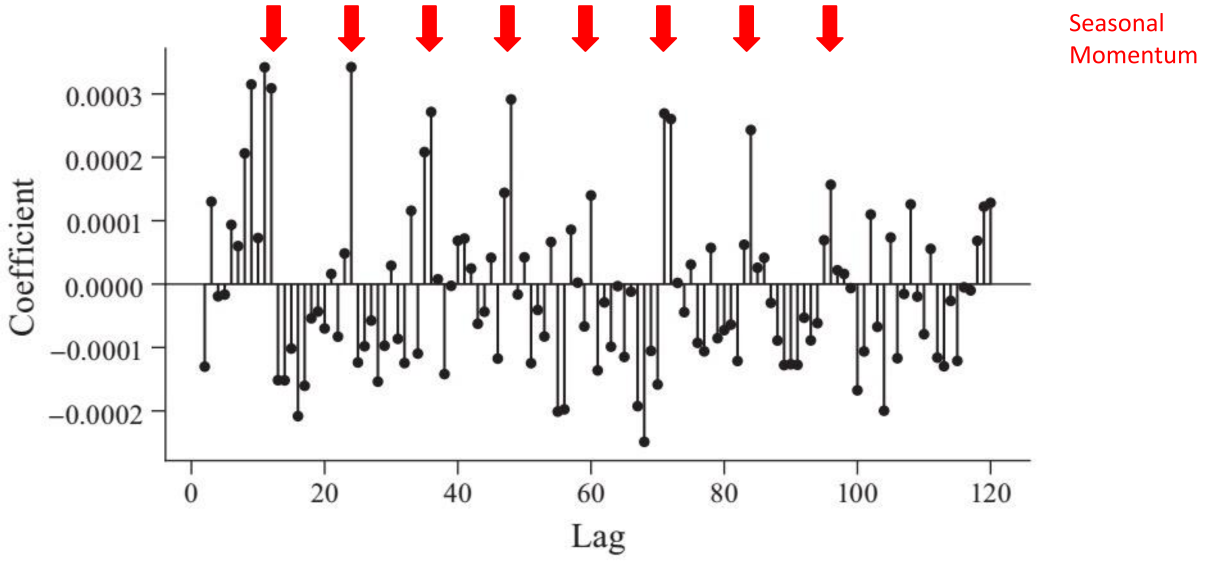
Example: Results

- Several well known anomalies found with a single regression



Example: Results

- Several well known anomalies found with a single regression



Example: Results

TABLE 3.2
Return prediction with a polynomial of lagged returns

						CV portfolio return r_p		
Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30

Example: Results

TABLE 3.2
Return prediction with a polynomial of lagged returns

Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	CV portfolio return r_p: Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30

- OLS has a high IS R^2 but the CV R^2 implies this is due to overfitting
- Clearly the Ridge regression method produces a better CV R^2
  - **i.e. regularization is important to maximizing R^2**

Example: Results

TABLE 3.2
Return prediction with a polynomial of lagged returns

Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	CV portfolio return r_p Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30

- Ridge narrowly produces higher average returns and even produces a lower sharpe ratio
  - Much of the remaining discussion revolves around why this is and what we can do about it

Why negative OOS R^2 in OLS

$$r_t = \mu + \varepsilon_t,$$ Return at time t

$$\mu = Xg.$$ X is constant across time

$$\boxed{\Sigma = I_N \sigma^2.}$$ Covariance matrix

$$\bar{r} = \frac{1}{\tau}\sum_{t=1}^{\tau} r_t$$ IS average return

$$\bar{\varepsilon} = \frac{1}{\tau}\sum_{t=1}^{\tau} \varepsilon_t$$

$$\hat{\mu} = X(X'X)^{-1}X'\bar{r}.$$ IS & OOS predicted return

$$\hat{\mu} = \mu + u, \qquad u = X(X'X)^{-1}X'\bar{\varepsilon}$$

$$\bar{r}_v = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} r_t, \qquad \bar{\varepsilon}_v = \frac{1}{T-\tau}\sum_{t=\tau+1}^{T} \varepsilon_t$$

$$\bar{r}_v = \mu + \bar{\varepsilon}_v$$ OOS average return

$$\bar{r}_v - \hat{\mu} = \mu + \bar{\varepsilon}_v - \hat{\mu} = \bar{\varepsilon}_v - u$$ OOS prediction error

$$R^2_{OOS} = 1 - \frac{(\bar{\varepsilon}_v - u)'(\bar{\varepsilon}_v - u)}{(\bar{\varepsilon}_v + \mu)'(\bar{\varepsilon}_v + \mu)}$$

$$\approx 1 - \frac{\frac{1}{T-\tau}\sigma^2}{\frac{1}{N}\mu'\mu + \frac{1}{T-\tau}\sigma^2} - \frac{\frac{1}{\tau}\sigma^2}{\frac{1}{N}\mu'\mu + \frac{1}{T-\tau}\sigma^2}$$

Why negative OOS R^2 in OLS

$r_t = \mu + \varepsilon_t,$ Return at time t

$\mu = Xg.$ X is constant across time

$\Sigma = I_N \sigma^2.$ Covariance matrix

$\bar{r} = \frac{1}{\tau} \sum_{t=1}^{\tau} r_t.$ IS average return

$\bar{\varepsilon} = \frac{1}{\tau} \sum_{t=1}^{\tau} \varepsilon_t$

$\hat{\mu} = X(X'X)^{-1} X' \bar{r}.$ IS & OOS predicted return

$\hat{\mu} = \mu + u, \qquad u = X(X'X)^{-1} X' \bar{\varepsilon}$

$$\bar{r}_v = \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} r_t, \qquad \bar{\varepsilon}_v = \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} \varepsilon_t$$

$\bar{r}_v = \mu + \bar{\varepsilon}_v$ OOS average return

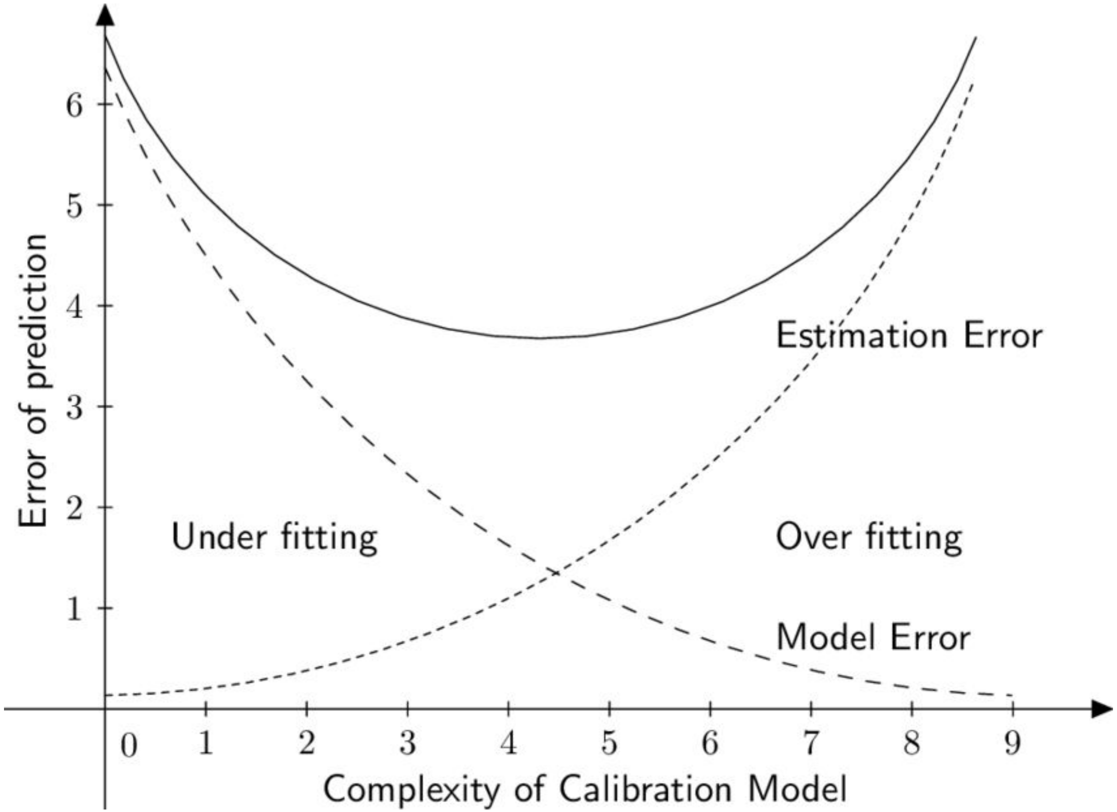
$\bar{r}_v - \hat{\mu} = \mu + \bar{\varepsilon}_v - \hat{\mu} = \bar{\varepsilon}_v - u$ OOS prediction error

$$R^2_{OOS} = 1 - \frac{(\bar{\varepsilon}_v - u)'(\bar{\varepsilon}_v - u)}{(\bar{\varepsilon}_v + \mu)'(\bar{\varepsilon}_v + \mu)}$$

$$\approx 1 - \boxed{\frac{\frac{1}{T-\tau}\sigma^2}{\frac{1}{N}\mu'\mu + \frac{1}{T-\tau}\sigma^2}} - \boxed{\frac{\frac{1}{\tau}\sigma^2}{\frac{1}{N}\mu'\mu + \frac{1}{T-\tau}\sigma^2}}$$

Model error

Estimation error


Error of prediction
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
Complexity of Calibration Model
Estimation Error
Under fitting
Over fitting
Model Error

Effects of Estimation Error on E[r]

$$\hat{\omega} = \frac{1}{\sqrt{\hat{\mu}'\hat{\mu}}}\hat{\mu}.$$

Portfolio Weights

$$\mathbb{E}[\hat{\omega}'\bar{r}_v|\hat{\omega}] \approx \frac{\mu'\mu}{\sqrt{\mu'\mu + \boxed{\frac{N}{\tau}\sigma^2}}}$$

- OOS expected return is also decreasing in estimation error
- Overall, estimation error penalizes R^2 and E[r] however this may change when we no longer assume a diagonal covariance matrix (next class)

OOS R^2 in Ridge Regression

- We now perform the same exercise but using the ridge regression instead of OLS
- Specifically for γ>0:

$$\hat{g} = (X'X + \gamma I_K)^{-1} X'\bar{r}.$$ If $$X'X = I_K$$, then $$\hat{g} = \frac{1}{1+\gamma} X'\bar{r},$$

$$\hat{\mu} = X\hat{g} = \frac{1}{1+\gamma} XX'\bar{r} = \boxed{\frac{1}{1+\gamma}\mu + \frac{1}{1+\gamma}u,}$$

$$u = XX'\bar{\varepsilon}.$$

OOS R^2 in Ridge Regression

$$R^2_{OOS} \approx 1 - \frac{\frac{1}{T-\tau}\sigma^2}{\frac{1}{N}\mu'\mu + \frac{1}{T-\tau}\sigma^2} - \left(\frac{\gamma^2}{(1+\gamma)^2}\right)\frac{\frac{1}{N}\mu'\mu}{\frac{1}{N}\mu'\mu + \frac{1}{T-\tau}\sigma^2} - \left(\frac{1}{(1+\gamma)^2}\right)\frac{\frac{1}{\tau}\sigma^2}{\frac{1}{N}\mu'\mu + \frac{1}{T-\tau}\sigma^2}$$

- As γ increases the 3rd term, driven by estimation error, decreases and R^2 increases
- However, at the same time the 2nd term increases
- The combined effect depends on γ, $1/\tau*\sigma^2$ and $1/N\mu'\mu$
- Overall, as the ratio of $1/\tau*\sigma^2$ (noise variance) and $1/N\mu'\mu$ (signal variance) increases, the γ that maximizes the R^2 increases

How does shrinkage affect E[r] and Sharpe

- We are shrinking $\hat{\boldsymbol{\mu}}$ by the constant factor 1/(1+γ)
- This implies that $\hat{\omega} = \frac{1}{\sqrt{\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}}}\hat{\boldsymbol{\mu}}.$ remains unchanged, thus

$\mathbb{E}[\hat{\omega}'\bar{r}_v|\hat{\omega}] \approx \frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}'\boldsymbol{\mu} + \frac{N}{\tau}\sigma^2}}$, and $\mathrm{var}(\hat{\omega}' r_v|\hat{\omega}) = \frac{1}{T-\tau}\sigma^2$ also unchanged

→ Improvement in OOS R^2 does not guarantee improvement in OOS Portfolio performance

How does shrinkage affect E[r] and Sharpe

TABLE 3.2
Return prediction with a polynomial of lagged returns

						CV portfolio return r_p		
Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30

- **Why does this result appear in the data?**
  - Covariates standardized (i.e. stdev = 1)
  - Low correlation across covariates
  - Thus ***X'X* ~ *I*,** just like in the previous slide, and all coefficients shrink approximately equally
    - This helps the R^2 but not performance metrics
- We are shrinking not only estimation error but also the expected return signal

What if *X'X* != *I*: Setup

- Consider: $\mathbf{X} = \mathbf{Q}_K \mathbf{\Lambda}_K^{\frac{1}{2}}$ with $\boldsymbol{Q}_K$ NxK and $\boldsymbol{\Lambda}_K$ a diagonal matrix with diagonal elements λ_j and $\boldsymbol{Q}_K$ orthogonalized
- Then $\mathbf{Q}_K' \mathbf{Q}_K = \boldsymbol{I}_K$, but $\mathbf{X}'\mathbf{X} = \mathbf{\Lambda}_K$ and the λ_j's determine the cross sectional dispersion of the covariates
- Thus the ridge regression provides predicted returns as:

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \mathbf{X}\left(\mathbf{\Lambda}_K + \gamma \boldsymbol{I}_K\right)^{-1} \mathbf{\Lambda}_K^{\frac{1}{2}} \mathbf{Q}_K' \bar{\boldsymbol{r}} \\ &= \mathbf{X}\left(\boldsymbol{I}_K + \gamma \mathbf{\Lambda}_K^{-1}\right)^{-1} \hat{\boldsymbol{g}}_{OLS}. \end{aligned} \tag{3.17}$$

Shrinkage determined by λ_j's. Low λ_j implies large shrinkage (i.e. small dispersion → hard to spot signal → leads to estimation error → shrink it!)

What if *X'X* != *I*: Performance

- Same weights and diagonal covariance matrix **Σ** as before
- Same covariance matrix means same portfolio variance
- Expected return is different:

$$\mathbb{E}[\hat{w}'\bar{r}_v|\hat{w}] \approx \frac{\sum_{j=1}^{K} \frac{g_j^2 \lambda_j}{1+\gamma\lambda_j^{-1}}}{\sqrt{\sum_{j=1}^{K} \frac{g_j^2 \lambda_j}{(1+\gamma\lambda_j^{-1})^2} + \frac{1}{\tau}\sigma^2 \sum_{j=1}^{K} \frac{1}{(1+\gamma\lambda_j^{-1})^2}}}. \qquad (3.19)$$

$\frac{g_j^2 \lambda_j}{}$

What if ***X'X*** != ***I***: Performance

- Same weights and diagonal covariance matrix as before
- Same covariance matrix means same portfolio variance
- Expected return is different:

$$\mathbb{E}[\hat{w}'\bar{r}_v|\hat{w}] \approx \frac{\boxed{\sum_{j=1}^{K} \frac{g_j^2\lambda_j}{1+\gamma\lambda_j^{-1}}}}{\sqrt{\sum_{j=1}^{K} \frac{g_j^2\lambda_j}{(1+\gamma\lambda_j^{-1})^2} + \boxed{\frac{1}{\tau}\sigma^2 \sum_{j=1}^{K} \frac{1}{(1+\gamma\lambda_j^{-1})^2}}}}. \qquad (3.19)$$

Both terms are shrinking, which shrinks more?

If the covariate *j* contributes little to the predictable return variation (i.e. $g_j^2\lambda_j$ is small) then the cost to shrinking the coefficient on *j* is small E[r] goes up

Summary of Current Points

- Shrinkage does not improve portfolio performance when there is **no cross sectional variation amongst predictors dispersion**
  - This is because shrinkage effects all parameters equally

- **With cross sectional variation amongst predictors dispersion**, assuming cross-sectionally homoskedastic and uncorrelated returns with orthogonalized predictors you can obtain better performance
  - This is because shrinkage effects parameters differently

- More generally, **shrinkage can improve portfolio performance if there is heterogeneity in the covariates' relative contribution to expected return and to estimation error**.
  - For example, shrink the ones that contribute heavily to estimation error but not so much to expected returns (the unimportant covariates)

Why does this matter

- How we scale our predictors matters!
- Typical Lasso and Ridge packages automatically standardize covariates
  - We have seen this costs us most if not all outperformance over OLS

What are we supposed to do then??

- Use our prior knowledge, Bayesian regression
- Rather than letting the Ridge program rescale all variables as if they are drawn from identical distributions, we should rescale certain variables based on our prior beliefs about which variables should affect predictable returns most and least

Back to the Example

- Recall we had 120 months of lagged returns, their squares and cubes
- If we believe that the non-linear relationships are weak we could:
  - First, let the ridge program rescale all covariates
  - Second, divide the squared terms by 2 and the cubed terms by 4 (arbitrary choices)
- Now we have at least some cross-sectional variation in covariate dispersion

Did it work?

Back to the Example

TABLE 3.2
Return prediction with a polynomial of lagged returns

Go to previous Page

Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	CV portfolio return r_p: Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	$\boxed{2.25}$	2.63	0.84	4.20	13.85	0.30
Ridge	Unequal	R^2	$\boxed{1.40}$	2.69	1.18	4.55	12.47	0.37
Ridge	Unequal	$E[r_p]$	3.11	1.75	0.89	4.58	12.94	0.35
Lasso	Unequal	R^2	0.00028	3.55	0.84	4.25	11.79	0.36

Rescaling leads to less shrinkage to maximize R^2

Back to the Example

TABLE 3.2
Return prediction with a polynomial of lagged returns

Go to previous Page

Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	CV portfolio return r_p Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	$\boxed{0.84}$	4.20	13.85	0.30
Ridge	Unequal	R^2	1.40	2.69	$\boxed{1.18}$	4.55	12.47	0.37
Ridge	Unequal	$E[r_p]$	3.11	1.75	0.89	4.58	12.94	0.35
Lasso	Unequal	R^2	0.00028	3.55	0.84	4.25	11.79	0.36

OOS R^2 improves by over 30%

Back to the Example

TABLE 3.2
Return prediction with a polynomial of lagged returns

Go to previous Page

Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	CV portfolio return r_p Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30
Ridge	Unequal	R^2	1.40	2.69	1.18	4.55	12.47	0.37
Ridge	Unequal	$E[r_p]$	3.11	1.75	0.89	4.58	12.94	0.35
Lasso	Unequal	R^2	0.00028	3.55	0.84	4.25	11.79	0.36

Higher average returns and Sharpe Ratio

Back to the Example

TABLE 3.2
Return prediction with a polynomial of lagged returns

Go to previous Page

						CV portfolio return r_p		
Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30
Ridge	Unequal	R^2	1.40	2.69	1.18	4.55	12.47	0.37
Ridge	Unequal	$E[r_p]$	3.11	1.75	0.89	4.58	12.94	0.35
Lasso	Unequal	R^2	0.00028	3.55	0.84	4.25	11.79	0.36

The prior that nonlinearities are less important appears correct

Back to the Example

Go to previous Page

TABLE 3.2
Return prediction with a polynomial of lagged returns

						CV portfolio return r_p		
Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30
Ridge	Unequal	R^2	1.40	2.69	1.18	4.55	12.47	0.37
Ridge	Unequal	$E[r_p]$	3.11	1.75	0.89	4.58	12.94	0.35
Lasso	Unequal	R^2	0.00028	3.55	0.84	4.25	11.79	0.36

Back to the Example

TABLE 3.2
Return prediction with a polynomial of lagged returns

Go to previous Page

Method	Scaling	CV criterion	γ (i)	IS R^2 (ii)	CV R^2 (iii)	CV portfolio return r_p Mean (iv)	S.D. (v)	Sharpe Ratio (vi)
OLS	Equal	n/a	0	5.22	−1.18	4.12	11.60	0.35
Ridge	Equal	R^2	2.25	2.63	0.84	4.20	13.85	0.30
Ridge	Unequal	R^2	1.40	2.69	1.18	4.55	12.47	0.37
Ridge	Unequal	$E[r_p]$	3.11	1.75	0.89	4.58	12.94	0.35
Lasso	Unequal	R^2	0.00028	3.55	0.84	4.25	11.79	0.36

Little to no difference when parameters objective is max E[r]

Conclusions

- Regularization that seeks to maximize R^2 does not necessarily translate to stronger portfolio performance
- Regularization focusing on portfolio performance does not improve OOS portfolio performance relative to R^2 regularization
- The scaling of covariates (which Ridge regression packages tend to standardize) has important implications for regularization
  - We need to bring prior knowledge on which predictors are likely to be most vs least important in estimating E[r]
  - We need to penalize least important predictors most

How to come up with these priors? Next week…