# MLAP Chapter 3.4-3.9

Zirui Wang

McCombs School of Business
**The University of Texas at Austin**

Februry 24, 2022

# Table of Contents

# Table of Contents

# Links between Expected Returns and Covariances



- The Bayesian regression framework allows us to tackle the question of **covariate scaling** by showing how economic links between **expected returns** and **covariances** relate to regularization.

# Links between Expected Returns and Covariances



- The Bayesian regression framework allows us to tackle the question of **covariate scaling** by showing how economic links between **expected returns** and **covariances** relate to regularization.
- Cross-sectional regression model $\bar{r} = Xg + \bar{\varepsilon}$

$$\hat{g} = \left( X' \Sigma^{-1} X + \frac{1}{\tau} \Sigma_g^{-1} \right)^{-1} X' \Sigma^{-1} \bar{r}.$$

# Links between Expected Returns and Covariances

- The Bayesian regression framework allows us to tackle the question of **covariate scaling** by showing how economic links between **expected returns** and **covariances** relate to regularization.
- Cross-sectional regression model $\bar{r} = Xg + \bar{\varepsilon}$

$$\hat{g} = \left( X' \Sigma^{-1} X + \frac{1}{\tau} \Sigma_g^{-1} \right)^{-1} X' \Sigma^{-1} \bar{r}.$$

- Within this framework, we now look for a specification of $X$ and prior beliefs about $g$ that are economically sensible. For this purpose, we also need a more realistic specification of the covariance matrix

# Link Intuition



- Te link to the covariance matrix should be such that if a long-short portfolio formed based on the values of a covariate produces a substantial mean return, it should also have substantial return volatility.

# Link Intuition



- Te link to the covariance matrix should be such that if a long-short portfolio formed based on the values of a covariate produces a substantial mean return, it should also have substantial return volatility.
- If such links between expected returns and covariances were absent, economically implausible near-arbitrage opportunities would exist.

# Assumptions



- Assume $K$ vectors of covariates used to predict returns are equal to $K$ eigenvectors of $\Sigma$.
- Let $\boldsymbol{\Sigma} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'$
- Let $\boldsymbol{Q}_K$ be a selection of K columns of the orthogonal matrix $\boldsymbol{Q}$
- Let $\boldsymbol{X} = \boldsymbol{Q}_K$, which implies $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_K$.
- Since the portfolio weights are based on eigenvectors, or principal components, of the return covariance matrix, we label them principal component (PC) portfolios. We can get that these PC portfolios have expected returns and variance

$$\mathbb{E}[\boldsymbol{Q}_K'\boldsymbol{r}_t] = \boldsymbol{g}, \qquad \mathrm{var}(\boldsymbol{Q}_K'\boldsymbol{r}_t) = \boldsymbol{\Lambda}_K,$$

# Assumptions



- Sharpe Ratios associated with each of these $K$ PC portfolio returns, $r_p = Q_K' r_t$, are $\Lambda_K^{-1/2} g$
- An assumption about these Sharpe ratios that has some economic K plausibility is that high Sharpe ratios are concentrated among PC portfolios that have relatively high variance.
- We can get the prior beliefs

$$\mathbf{g} \sim \mathcal{N}(0, \gamma^{-1} \boldsymbol{\Lambda}_K^2), \qquad 0 < \gamma < 1,$$

- The hyperparameter $\gamma$ now has an economic interpretation as controlling the expected squared Sharpe ratio under prior beliefs.
- The maximum squared Sharpe ratio attainable from the assets is $g' Q_K' \Sigma^{-1} Q_K g$. Taking expectations under the prior distribution, we get

$$\mathbb{E}[\mathbf{g}' \mathbf{Q}_K' \boldsymbol{\Sigma}^{-1} \mathbf{Q}_K \mathbf{g}] = \mathbb{E}[\mathbf{g}' \boldsymbol{\Lambda}_K^{-1} \mathbf{g}] = \frac{1}{\gamma} \operatorname{tr}(\boldsymbol{\Lambda}_K).$$

# Assumptions

Given the prior beliefs in (3.23), the Bayesian regression of $\bar{\boldsymbol{r}}$ on $\boldsymbol{X} = \boldsymbol{Q}_K$ in (3.21) becomes

$$\hat{\boldsymbol{g}} = \left(\boldsymbol{\Lambda}_K^{-1} + \frac{\gamma}{\tau}\boldsymbol{\Lambda}_K^{-2}\right)^{-1} \boldsymbol{\Lambda}_K^{-1}\boldsymbol{Q}_K'\bar{\boldsymbol{r}}$$

$$= \left(\boldsymbol{I}_K + \frac{\gamma}{\tau}\boldsymbol{\Lambda}_K^{-1}\right)^{-1} \boldsymbol{Q}_K'\bar{\boldsymbol{r}}, \qquad (3.25)$$

- This type of shrinkage expresses the prior beliefs that these low-variance portfolios are unlikely to be the source of high Sharpe ratios.
- The assumptions here that the covariates are exactly equal to K eigenvectors of $\Sigma$ is not quite realistic.
- the analysis in this transparent special case illustrates how one can use the Bayesian approach to bring economic priors into the analysis and give an economic interpretation to shrinkage parameters.

# Table of Contents

# Return Covariances and Portfolio Aggregation



- Covariances of prediction errors play a much bigger role in asset pricing than in typical ML applications.
- While we have assumed so far that the covariance matrix is known, in practice it must be estimated. This introduces an additional layer of estimation errors that can lead to substantial problems in portfolio construction.
- Estimating this covariance matrix would be difficult without imposing substantial constraints on its functional form or some form of shrinkage estimation applied to the covariance matrix.

# portfolio aggregation approach



- It can make sense to first aggregate stocks into port- folios based on the covariates of the return prediction model
- We want to aggregate into these portfolios, and hence get these benefits, without negatively affecting the investment opportunities in terms of the maximum squared Sharpe ratio

# portfolio aggregation approach



- It can make sense to first aggregate stocks into port- folios based on the covariates of the return prediction model
- We want to aggregate into these portfolios, and hence get these benefits, without negatively affecting the investment opportunities in terms of the maximum squared Sharpe ratio
- If and only if the covariance matrix is of this form, then aggregation by X does not lead to a loss of any investment opportunities for a mean- variance investor.

  A necessary and sufficient condition for this equality to hold (Amemiya (1985), Theorem 6.1.1) is that the covariance matrix takes the form

$$\boldsymbol{\Sigma} = \boldsymbol{X\Psi X}' + \boldsymbol{U\Phi U}' + \sigma^2 \boldsymbol{I}_N, \tag{3.27}$$

  for some conformable matrices $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$, and a matrix $\boldsymbol{U}$ such that $\boldsymbol{U}'\boldsymbol{X} = 0$.
- More in the next chapter

# Table of Contents

# Nonlinearity



TABLE 3.2
Return prediction with a polynomial of lagged returns

| Method | Scaling | CV criterion | $\gamma$ (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 |

- Nonlinearities play a big role in many ML applications. Howevr, the inclusion of second- and third-order terms does not contribute much to return predictability

# Nonlinearity



TABLE 3.2
Return prediction with a polynomial of lagged returns

| Method | Scaling | CV criterion | $\gamma$ (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 |

- Nonlinearities play a big role in many ML applications. Howevr, the inclusion of second- and third-order terms does not contribute much to return predictability
- Adjusting the covariate scaling so that ridge regression downweights these nonlinear terms actually improves the predictive performance of the model

# Nonlinearity

TABLE 3.2
Return prediction with a polynomial of lagged returns

| Method | Scaling | CV criterion | $\gamma$ (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 |

- Nonlinearities play a big role in many ML applications. Howevr, the inclusion of second- and third-order terms does not contribute much to return predictability
- Adjusting the covariate scaling so that ridge regression downweights these nonlinear terms actually improves the predictive performance of the model
- Cannot prove that nonlinearities are not important, but as least not as helpful as in many ML applications.

# Nonlinearity



- Interactions between covariates, rather than additive nonlinearity of individual predictors, are perhaps a more plausible source of nonlinear- ities in return prediction models

# Nonlinearity



- Interactions between covariates, rather than additive nonlinearity of individual predictors, are perhaps a more plausible source of nonlinear- ities in return prediction models
- Interaction Effect Examples - Some return predictability patterns could be stronger
  - among smaller, illiquid stocks
  - among stocks that are particularly strongly exposed to investor sentiment or to macroeconomic risks

# Nonlinearity



- Interactions between covariates, rather than additive nonlinearity of individual predictors, are perhaps a more plausible source of nonlinear- ities in return prediction models
- Interaction Effect Examples - Some return predictability patterns could be stronger
  - among smaller, illiquid stocks
  - among stocks that are particularly strongly exposed to investor sentiment or to macroeconomic risks
- More in next chapter

# Table of Contents

# Sparsity

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 |

- Many efforts to bring ML methods into asset pricing have focused on lasso-type methods that allow for sparsity

# Sparsity



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 |

- Many efforts to bring ML methods into asset pricing have focused on lasso-type methods that allow for sparsity
- Even thought it's successful in ML, it is not obvious that sparsity-inducing priors have much justification in asset pricing applications

# Sparsity

TABLE 3.2
Return prediction with a polynomial of lagged returns

| Method | Scaling | CV criterion | $\gamma$ (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 |

- We can see higher *IS* $R^2$, but a bit worse *CV* $R^2$. Thus, in terms of the predictive performance, lasso does worse than ridge regression

# Sparsity



TABLE 3.2
Return prediction with a polynomial of lagged returns

| Method | Scaling | CV criterion | $\gamma$ (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 |

- We can see higher $IS\ R^2$, but a bit worse $CV\ R^2$. Thus, in terms of the predictive performance, lasso does worse than ridge regression
- Lasso does somewhat worse than ridge in terms of the mean return, but the standard deviation of the portfolio return is lower as well, and so the Sharpe ratio ends up slightly higher

# Sparsity



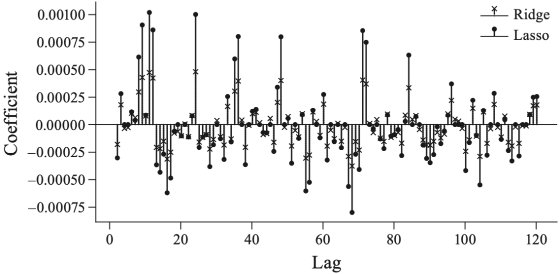
TABLE 3.2
Return prediction with a polynomial of lagged returns

| Method | Scaling | CV criterion | $\gamma$ (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 |

- We can see higher *IS* $R^2$, but a bit worse *CV* $R^2$. Thus, in terms of the predictive performance, lasso does worse than ridge regression
- Lasso does somewhat worse than ridge in terms of the mean return, but the standard deviation of the portfolio return is lower as well, and so the Sharpe ratio ends up slightly higher
- Sparsity is not as helpful for predictive performance in asset pricing as it can be in many other ML applications

# Sparsity



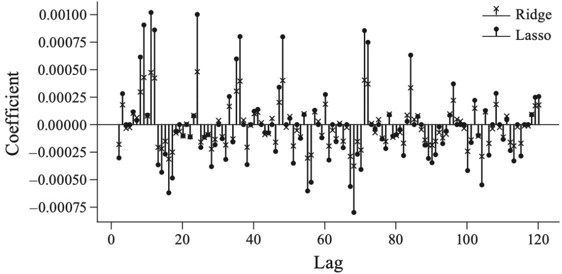



(a) Coefficients for past returns

- The estimates reveal the general tendency of lasso to shrink coef- ficients less than ridge, unless they are shrunk all the way to zero

# Sparsity



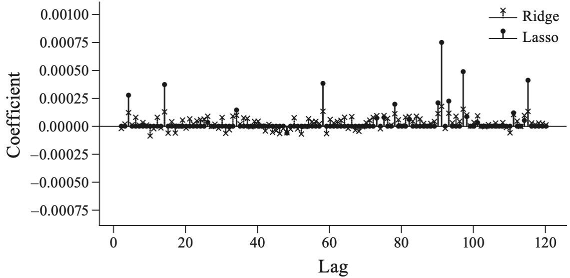



(a) Coefficients for past returns

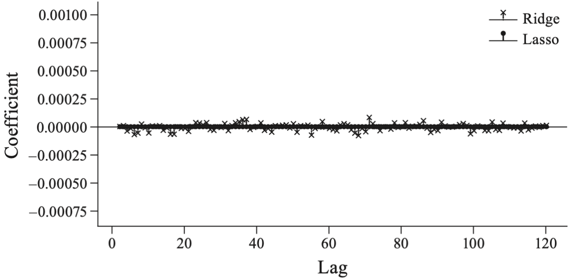
- The estimates reveal the general tendency of lasso to shrink coef- ficients less than ridge, unless they are shrunk all the way to zero
- We do not have strong a priori reasons to expect sparsity

## Sparsity





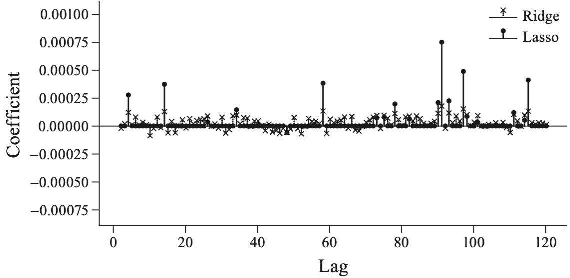

(b) Coefficients for squared past returns



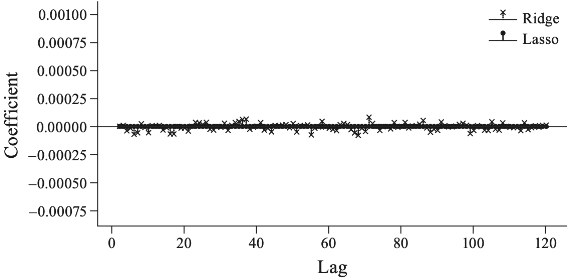

(c) Coefficients for third power of past returns

- Most of the coefficients that have been set to zero are all coefficients for second- and third-order terms

# Sparsity




(b) Coefficients for squared past returns

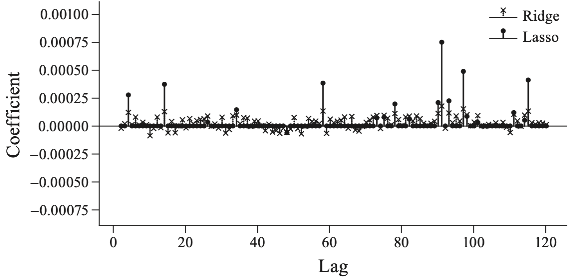



(c) Coefficients for third power of past returns

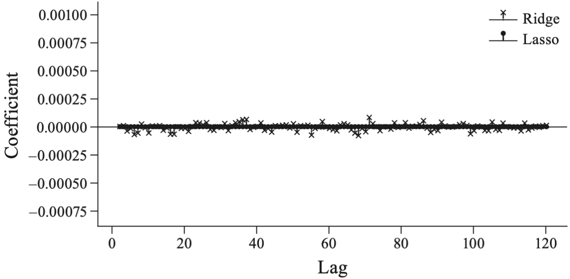
- Most of the coefficients that have been set to zero are all coefficients for second- and third-order terms
- Lasso almost completely discards the nonlinear terms from the model

## Sparsity





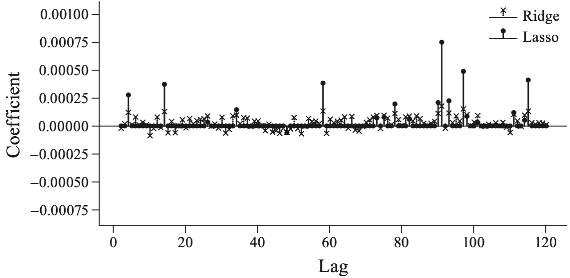
(b) Coefficients for squared past returns



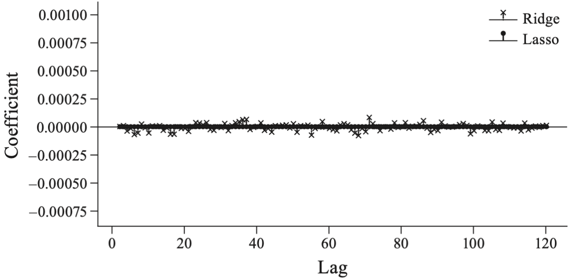
(c) Coefficients for third power of past returns

- Most of the coefficients that have been set to zero are all coefficients for second- and third-order terms
- Lasso almost completely discards the nonlinear terms from the model
- No a priori reason to expect particular types of additive nonlinearities

# Sparsity




(b) Coefficients for squared past returns




(c) Coefficients for third power of past returns

- Most of the coefficients that have been set to zero are all coefficients for second- and third-order terms
- Lasso almost completely discards the nonlinear terms from the model
- No a priori reason to expect particular types of additive nonlinearities
- One should not take it for granted that sparsity is helpful for predictive performance in asset pricing

# Table of Contents

# Structural Change



- Data-generating process in financial markets is likely undergoing continuous structural change

# Structural Change



- Data-generating process in financial markets is likely undergoing continuous structural change
  - Economy overall is undergoing structural change

# Structural Change



- Data-generating process in financial markets is likely undergoing continuous structural change
  - Economy overall is undergoing structural change
  - Investors learn from data

# Structural Change



- Data-generating process in financial markets is likely undergoing continuous structural change
  - Economy overall is undergoing structural change
  - Investors learn from data
- In the ML literature, the structural change problem is known as concept drift

# Structural Change



- Data-generating process in financial markets is likely undergoing continuous structural change
  - Economy overall is undergoing structural change
  - Investors learn from data
- In the ML literature, the structural change problem is known as concept drift
- Solution in ML is a weighting scheme that gives more weight to recent data in the training of an algorithm

# Structural Change



- Data-generating process in financial markets is likely undergoing continuous structural change
  - Economy overall is undergoing structural change
  - Investors learn from data
- In the ML literature, the structural change problem is known as concept drift
- Solution in ML is a weighting scheme that gives more weight to recent data in the training of an algorithm
  - Rolling window approach

# Structural Change



- Data-generating process in financial markets is likely undergoing continuous structural change
  - Economy overall is undergoing structural change
  - Investors learn from data
- In the ML literature, the structural change problem is known as concept drift
- Solution in ML is a weighting scheme that gives more weight to recent data in the training of an algorithm
  - Rolling window approach
  - Exponential weighting (little work has been done bringing from ML into AP)

# Structural Change

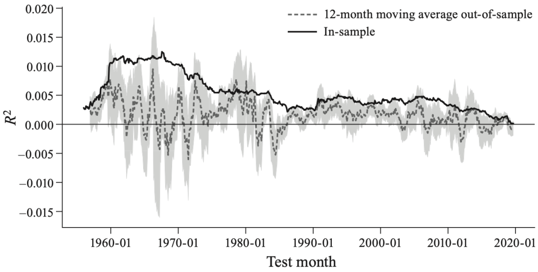



Figure 3.3. Rolling forward out-of-sample prediction $R^2$

- Rolling window approach with 20-year rolling windows, Ridge regression

# Structural Change



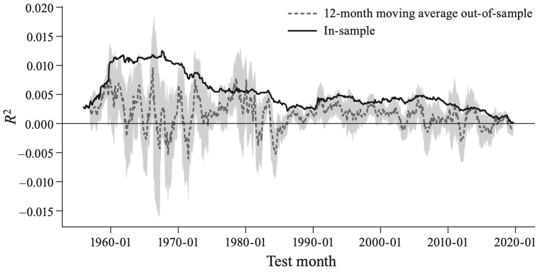



Figure 3.3. Rolling forward out-of-sample prediction $R^2$

Findings:

- *OOS* $R^2$ is almost everywhere smaller than the *IS* $R^2$
  - Cross-validated *IS* $R^2$ still seems to be an upward biased estimate of the *OOS* $R^2$
  - Structural change also accounts for a substantial part of the $R^2$-decay from IS to OOS.

# Structural Change

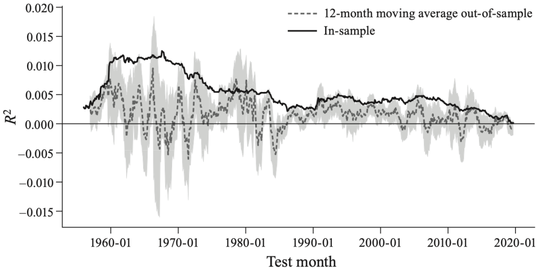



Figure 3.3. Rolling forward out-of-sample prediction $R^2$

Findings:

- *OOS* $R^2$ is almost everywhere smaller than the *IS* $R^2$
  - Cross-validated *IS* $R^2$ still seems to be an upward biased estimate of the *OOS* $R^2$
  - Structural change also accounts for a substantial part of the $R^2$-decay from IS to OOS.
- In the last 10–15 years of the sample, the average *OOS* $R^2$ is close to zero.

# Structural Change: Issues



- Issue is not only how to **track structural change** in the parameters of the prediction model, but also whether and **how to adapt the values of the penalty hyperparameters** over time
- In the previously discussed example, we kept the penalty parameter $\gamma$ fixed, but there is not necessarily a good reason to keep it fixed over time
- Reestimating the penalty hyperparameters every period may be computationally too expensive
- Monti, Anagnostopoulos, and Montana (2018) propose a **recursive updating scheme** for penalty hyperparameters

# Structural Change: Issues



- Issue is not only how to **track structural change** in the parameters of the prediction model, but also whether and **how to adapt the values of the penalty hyperparameters** over time
- In the previously discussed example, we kept the penalty parameter $\gamma$ fixed, but there is not necessarily a good reason to keep it fixed over time
- Reestimating the penalty hyperparameters every period may be computationally too expensive
- Monti, Anagnostopoulos, and Montana (2018) propose a **recursive updating scheme** for penalty hyperparameters
- Structural change considerations also raise questions about the suitability of CV methods for model validation and hyperparameter tuning (More in Chapter 5)

# Table of Contents

# Conclusion

- While ML techniques can be useful, off-the-shelf application of ML methods without careful adaptation to the specific conditions of an asset pricing application is unlikely to produce good results
- We need an analytical framework that allows us to inject a limited amount of economic reasoning when we set up ML tools to tackle asset pricing problems