ML for Finance:
Cross Validation, Model and Forecast Evaluation

S. Yanki Kalfa

JHU-SAIS

June 30, 2022

Outline

Cross Validation: Motivation

- Machine Learning models are different than standard forecasting methods
- We can solve standard OLS models easily

$$\hat{\beta} = \underset{\beta}{\arg\min} (Y - X\beta)^2 = (X'X)^{-1}X'Y$$

- The solution to most ML models is not as simple as above
- For example the LASSO:

$$\hat{\beta} = \underset{\beta}{\arg\min} (Y - X\beta)^2 + \lambda|\beta|$$

- No analytical solution
- How do we pick the best λ ?
- For each λ we have a $\hat{\beta}$

Cross Validation: How to choose θ

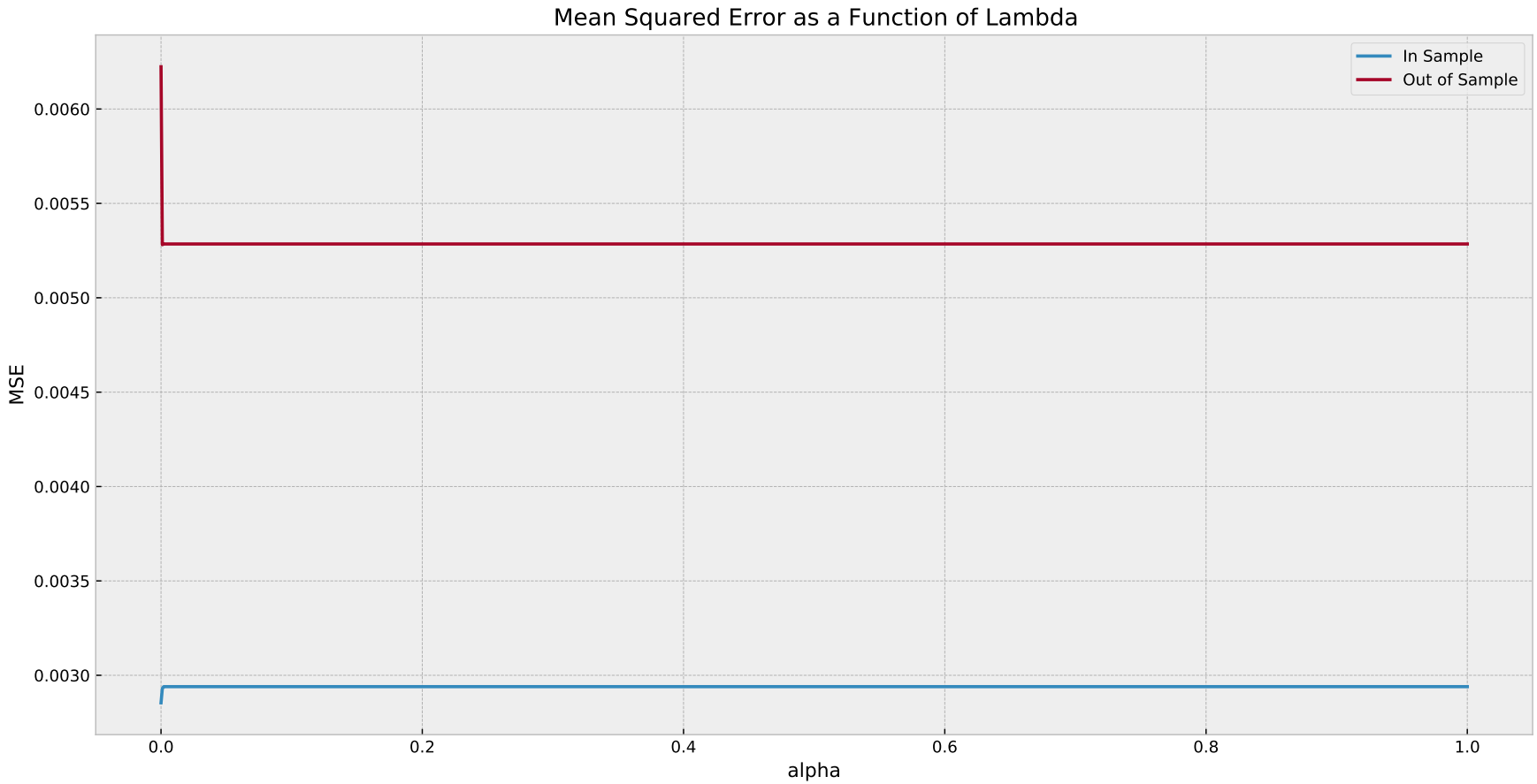
- Let θ be the hyperparameter, or a vector of hyperparameters
- Should we choose θ based on In Sample Error ?
- Should we choose θ based on Out of Sample Error ?
- We can choose the θ that minimize either the in sample or the out of sample error. However, we need to keep in mind that minimizing in sample error leads to overfitting. Furthermore, in sample error is a very poor estimate of out of sample error.

$$\bar{\varepsilon}^{IS} = \frac{1}{T}\sum_{t=1}^{T} y_t - \hat{f}(X_{t-1})$$

$$\varepsilon^{OoS} = y_{T+1} - \hat{f}(X_T)$$

- What is the problem here?

Cross Validation: How to choose θ



Cross Validation: How to choose θ

- In sample MSE is smaller than out of sample MSE for all λ
- As penalty increases Out of Sample MSE decreases
- As penalty increases In Sample MSE increases
- We care about out of sample MSE, we tune the hyperparameter to achieve smaller out of sample MSE

Validation Set

- We divide the sample into a traning set and a validation set
- We fit our model on the training set and predict in the validation set
- The validation set error gives us an estimate of the out of sample estimate

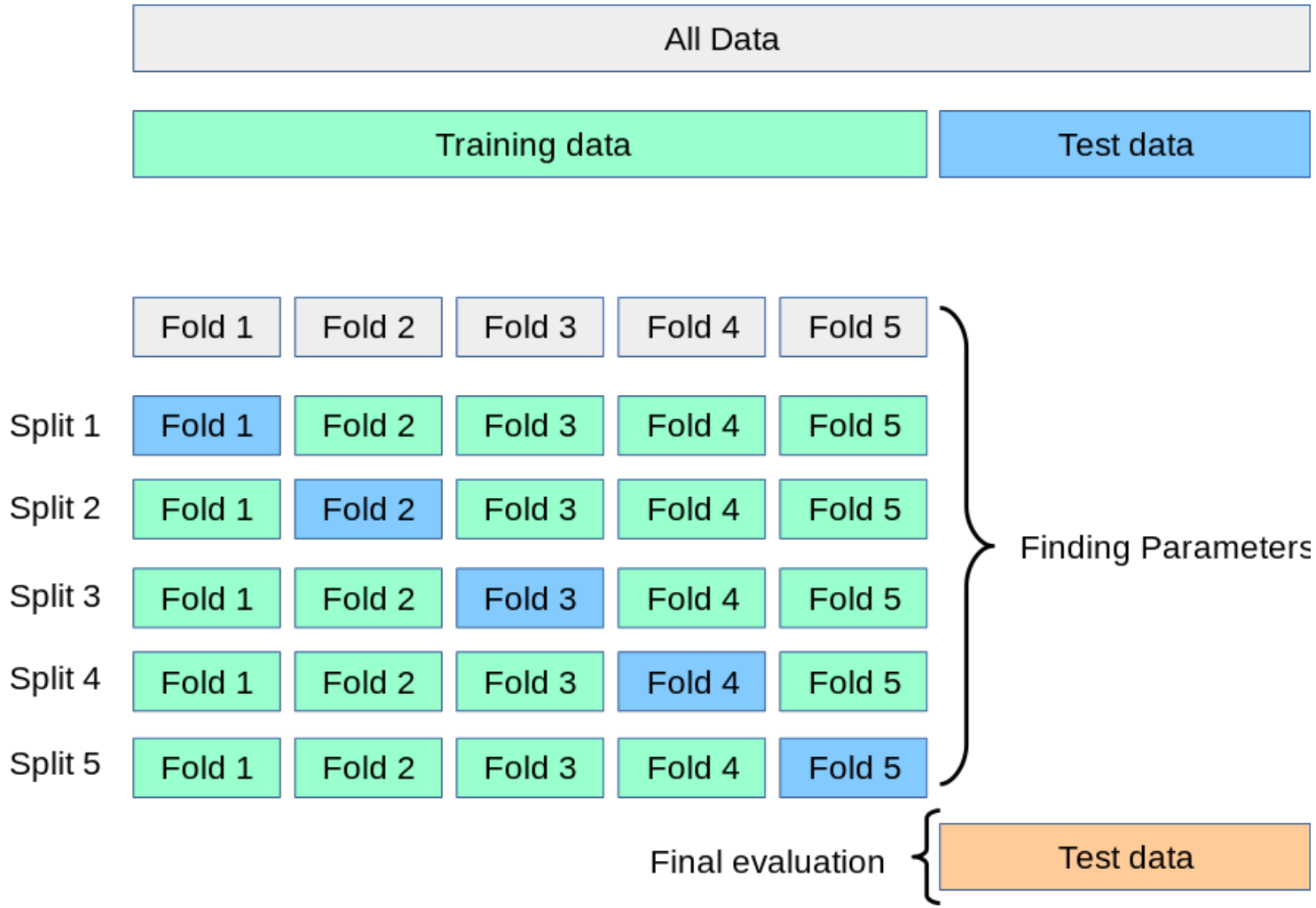
Drawbacks

- Validation error are sensitive to observations included in the training sample and validation sets
- This is because with the validation technique we fit the model to only a subset of the observation
- This means that there is a chance that validation errors may over estimate the test error.

K-Fold Cross Validation

- We randomly split the sample into K parts
- Leave out the k^{th} part
- Estimate the model on the $K-1$ parts
- predict the k^{th} part
- Repeat the process for each $k = 1, 2, \ldots, K$

K-Fold Cross Validation



Mathematical Details

- We define a model with a hyperparameter as: $\hat{f}(x, \alpha)$
- The Cross Validation estimation of errors are given by :

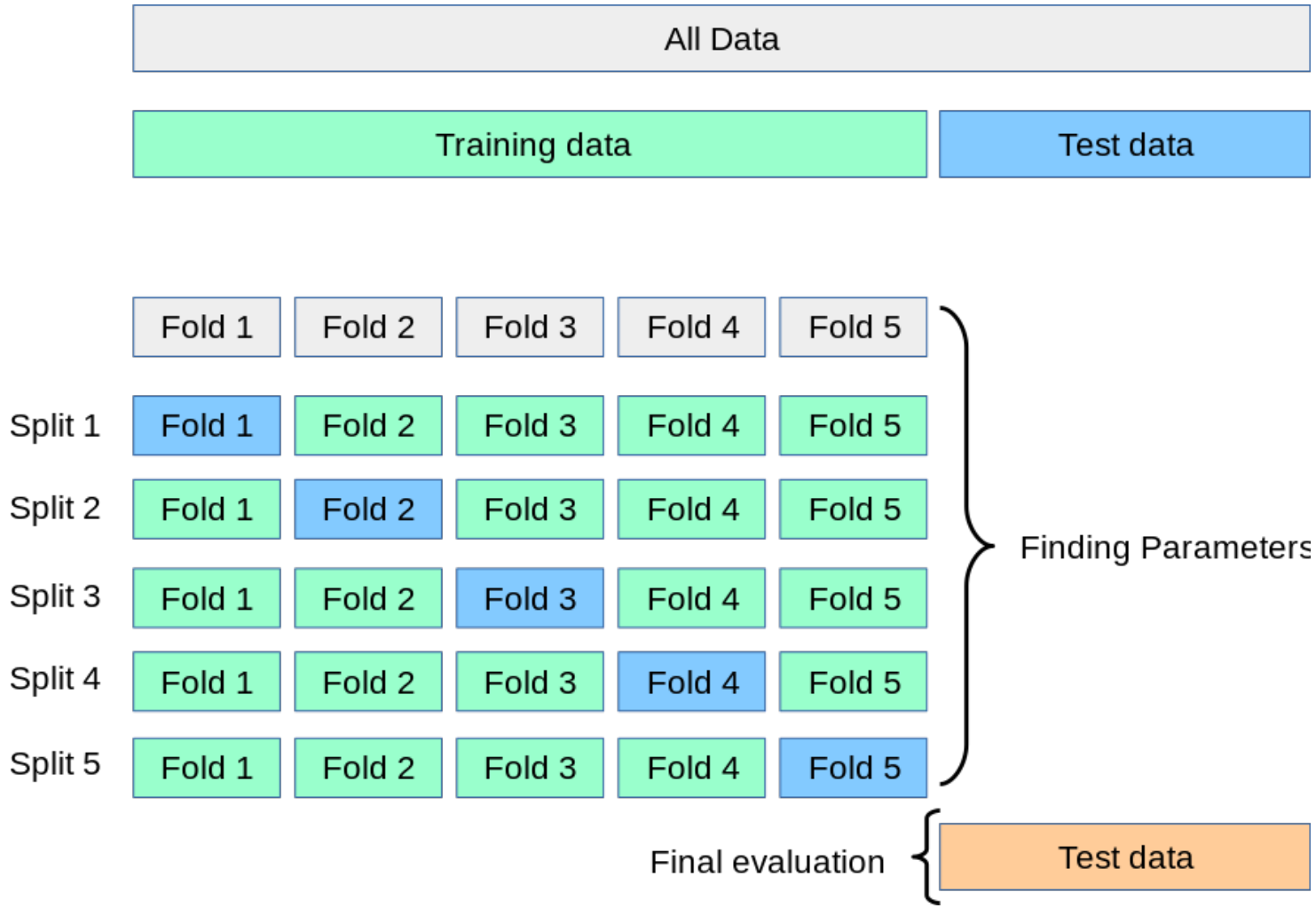
$$CV(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{\kappa(i)}(x_i, \alpha))$$

- We choose the hyperparameter that minimizes the above function and fit it to the full sample
- Typical values for K are 5 or 10
- We want to test the model against different combination of validation sets to get a more accurate expected test error

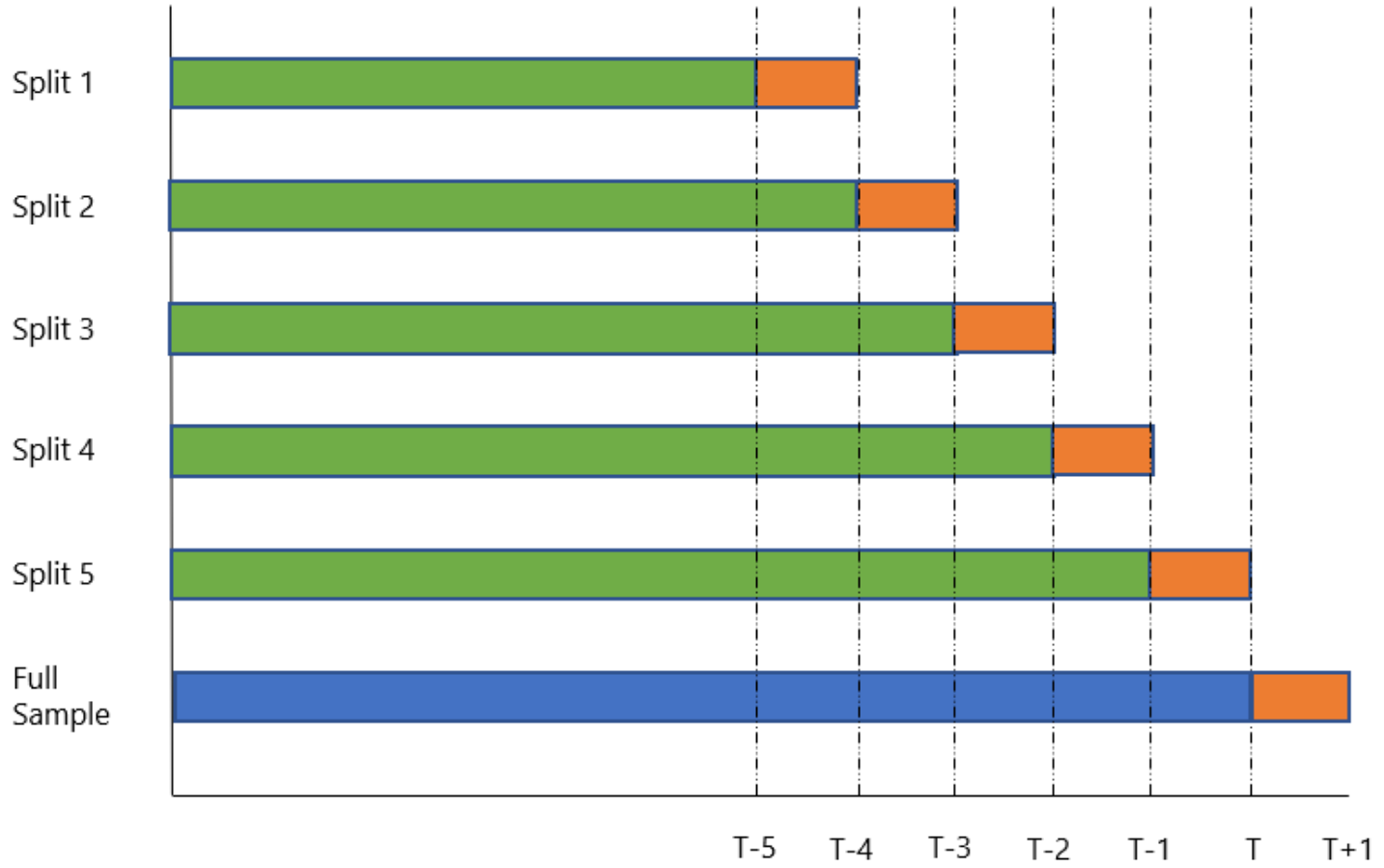
Time Series Cross Validation: Introduction

- We know that K-Fold CV is widely used
- Is it appropriate for forecasting? NO!
- K-Fold uses data from the future to tune data from the past
- This breaks the concept of forecasting

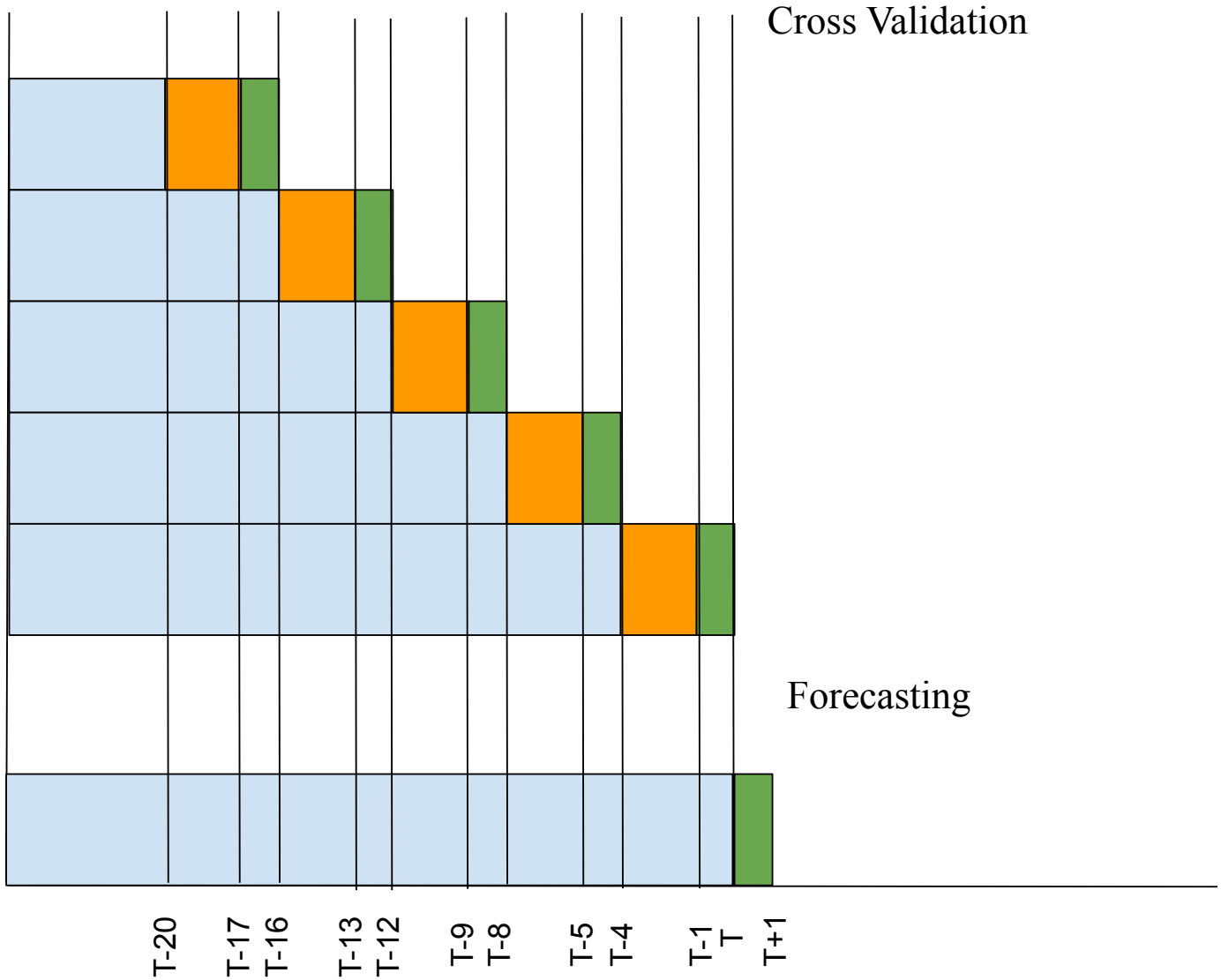
Time Series Cross Validation: K-Fold Comparison



Time Series Cross Validation: Expanding Window



Time Series Cross Validation: Expanding Window with Gap



Model Instability

- There are a few reasons why forecasts don't work:
- The relationship between explanatory variables and the dependent variable changes
- What do I mean by that?
  - Model coefficients change
  - Some variables become less important
  - Signs of coefficients flip
- We definitely cannot ignore this problem

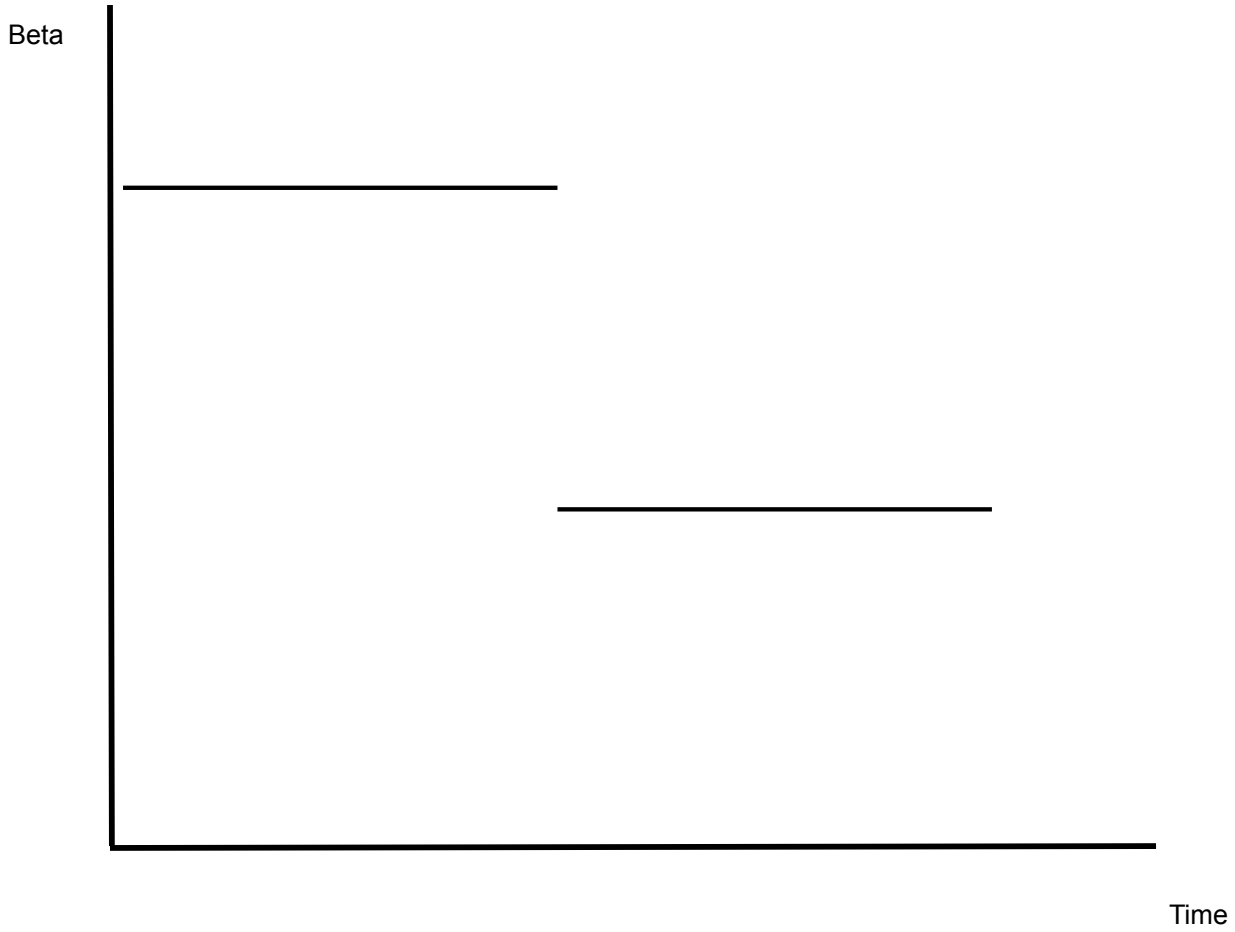
Model Instability

- To have a reliable forecasting model we need to have a stable model
- Analyzing the evolution of coefficients or variable importance across time is important
- If variable importance changes across time we should not expect to have reliable forecasts

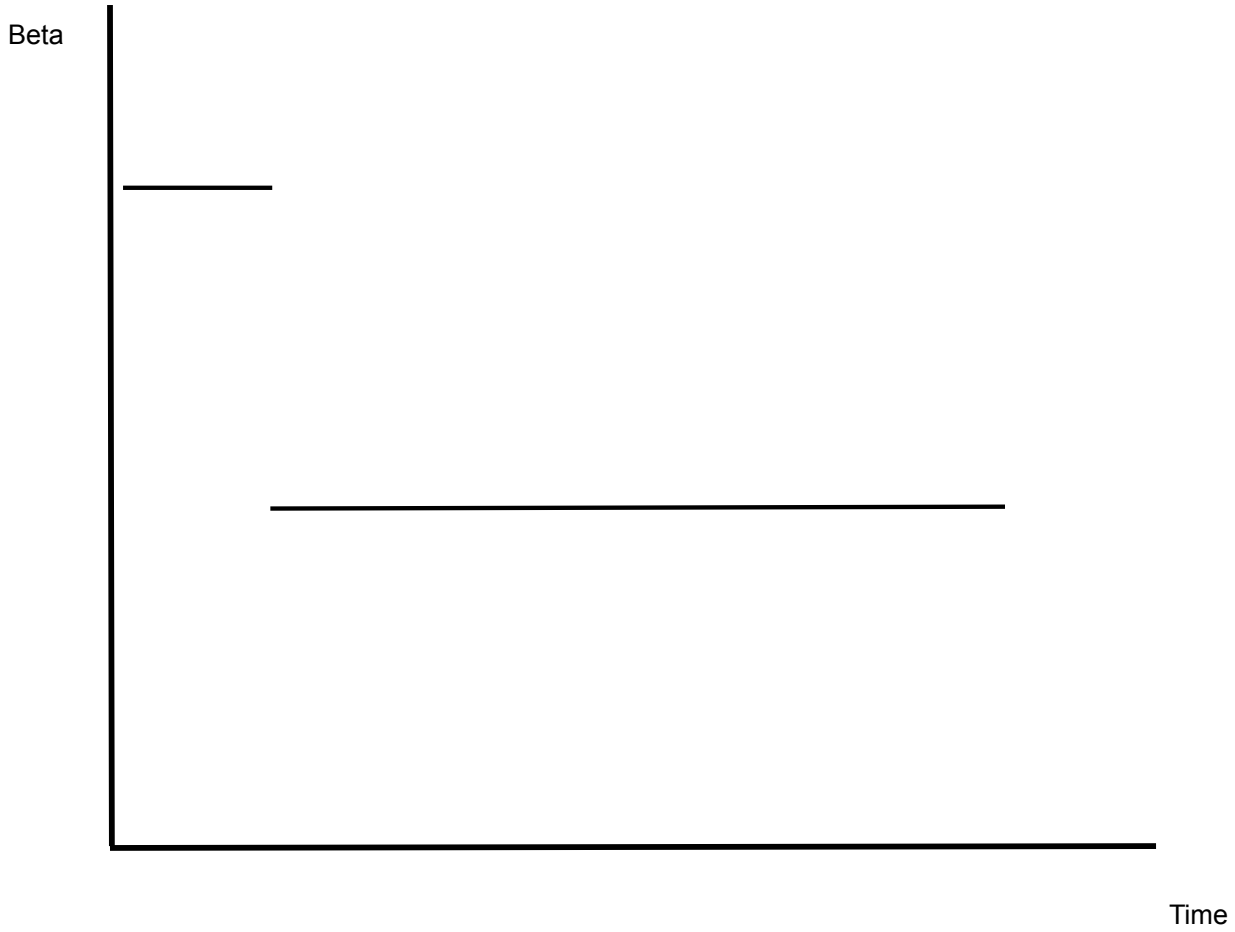
Model Instability: Structural Breaks

- Can we test for breaks
- Yes
- Different types of break tests:
  - Know single break
  - Unknown single break
  - Unknown multiple breaks
- We are not going to cover classical methods
- Covered in the Spring semester by Dr. Campbell and Dr. Ericsson

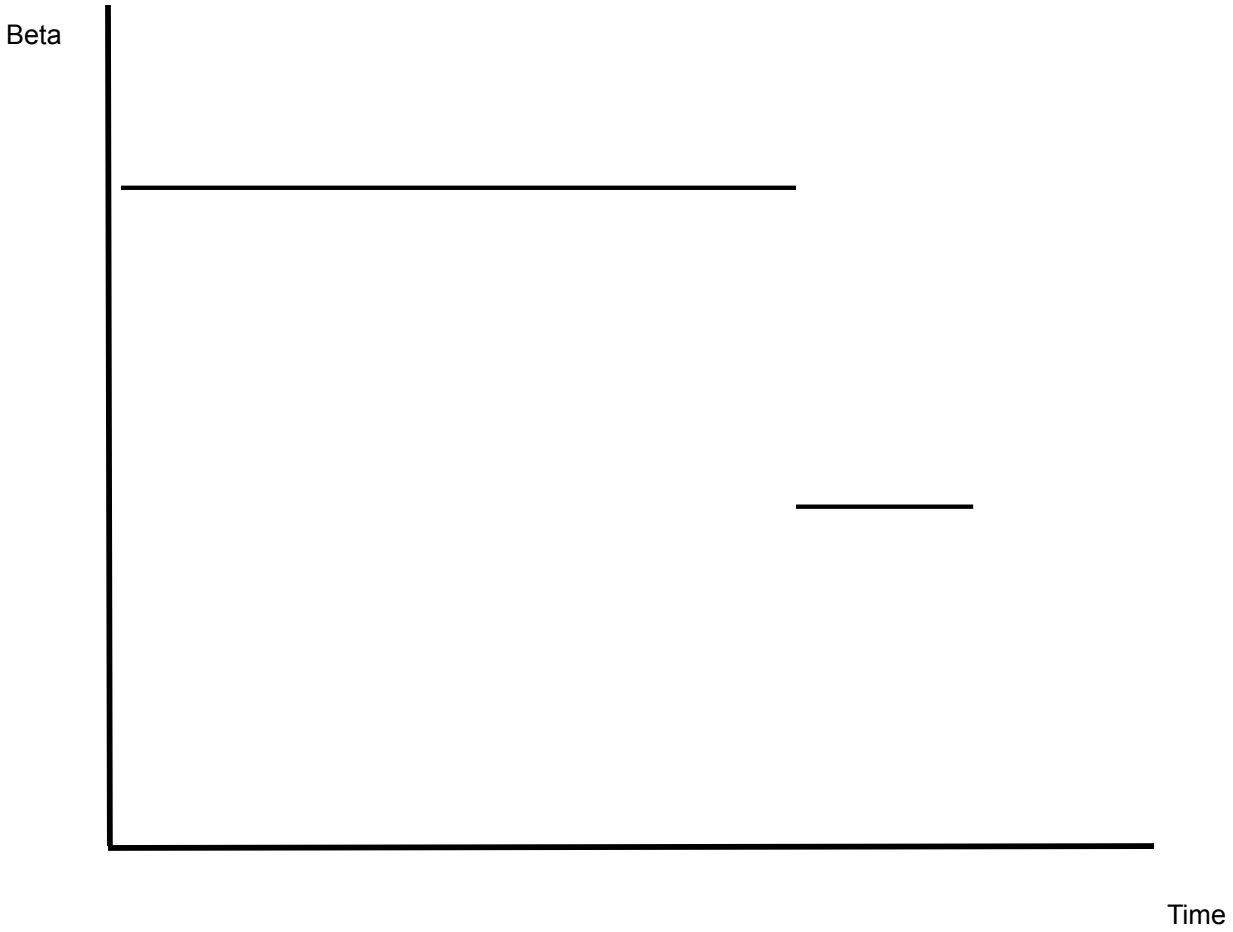
Parameter Instability: Mid Sample Break



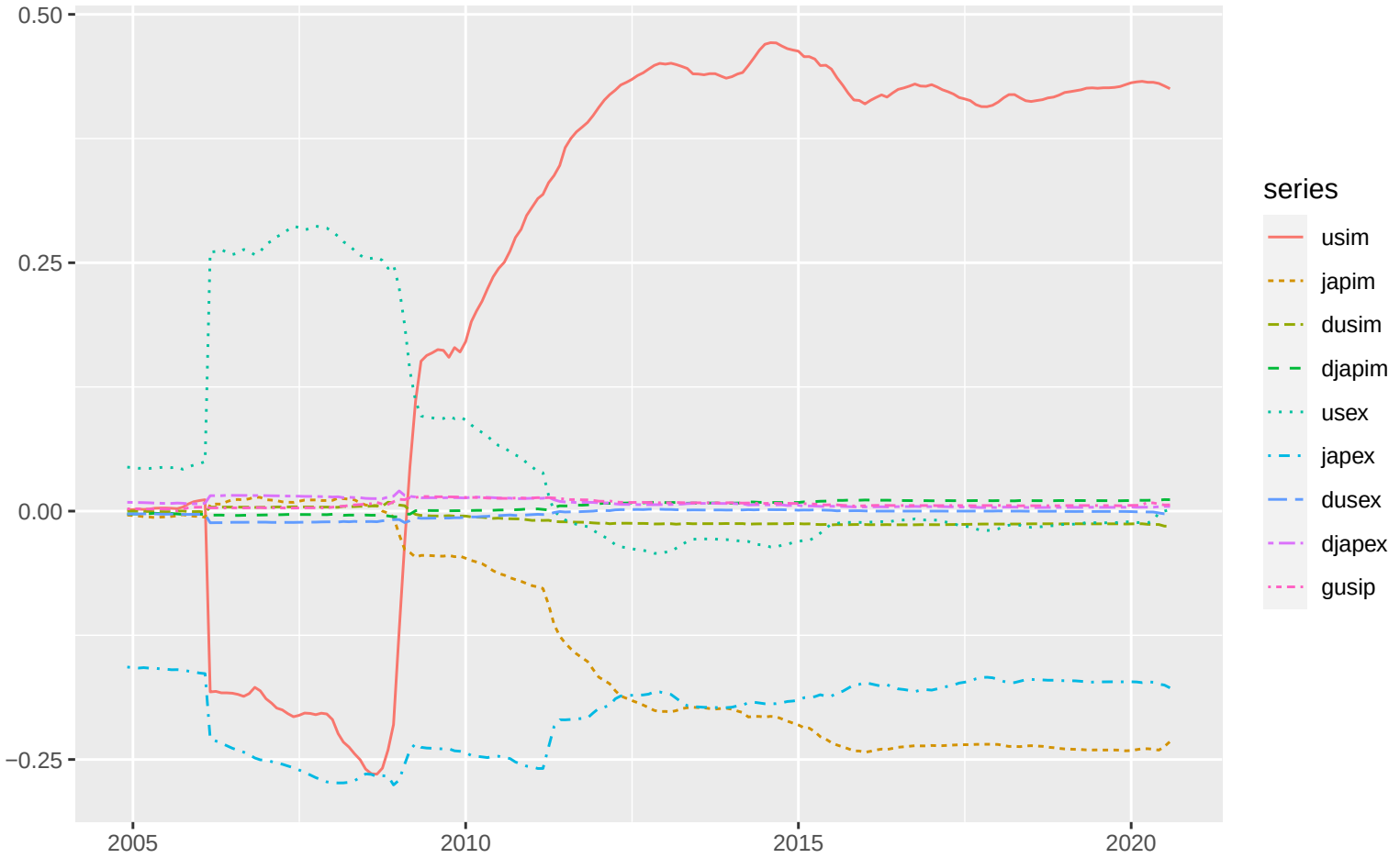
Parameter Instability: Early Break



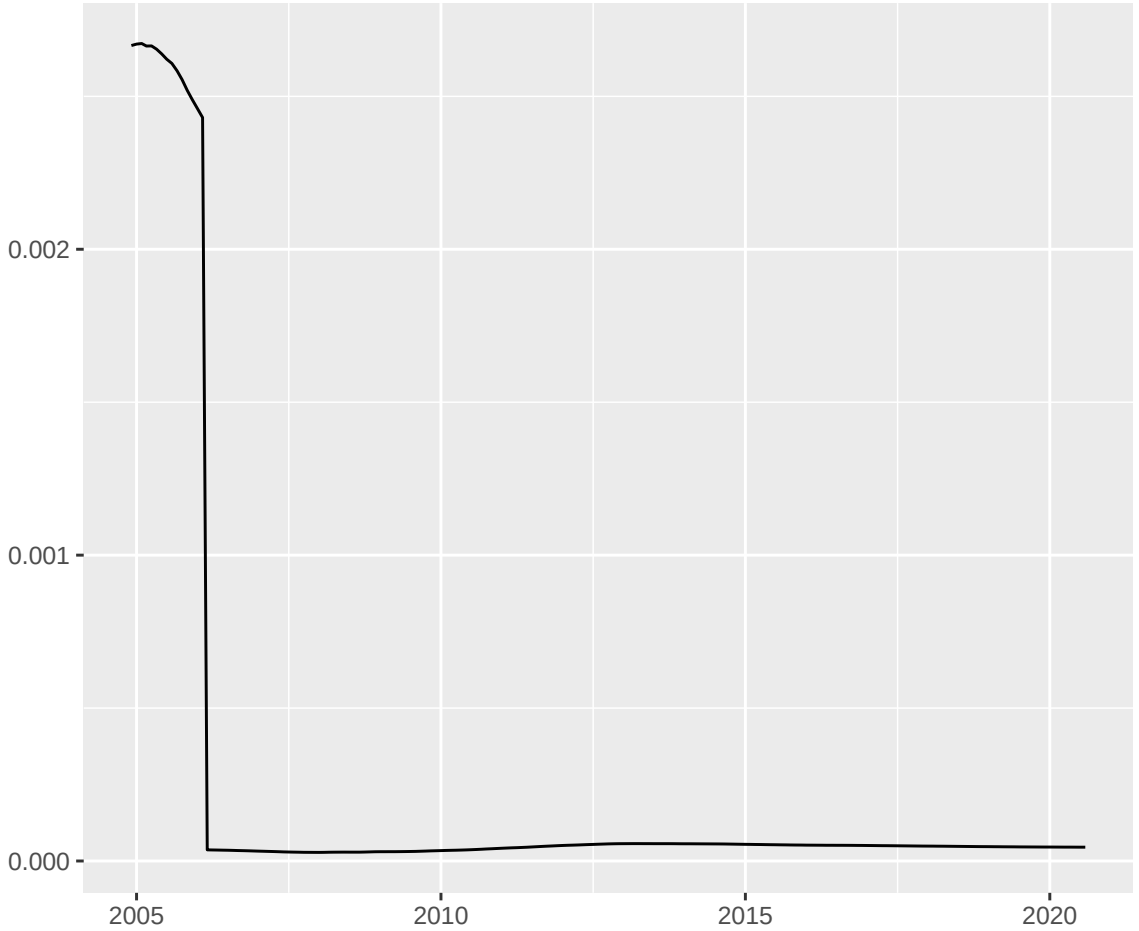
Parameter Instability: Late Break



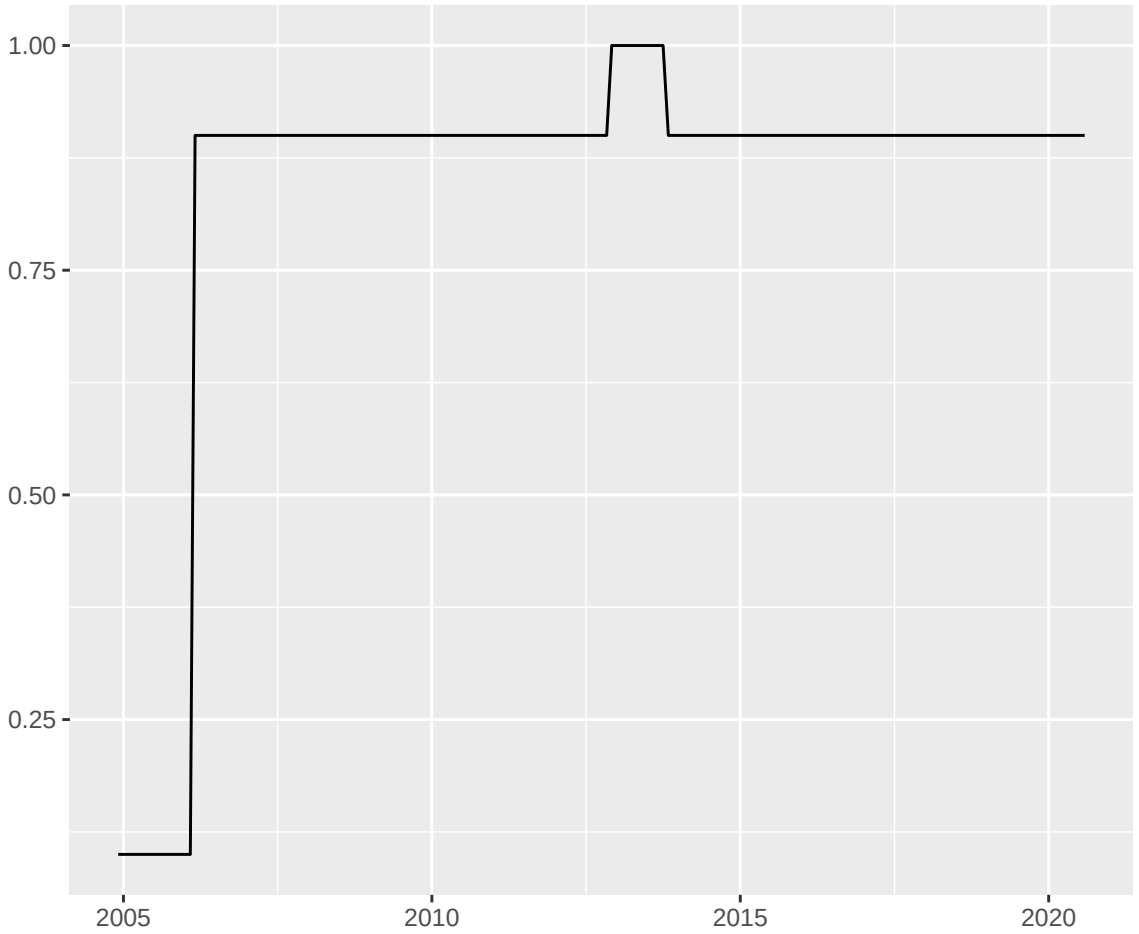
Model Instability: Forecasting Exchange Rate: βs



Model Instability: Forecasting Exchange Rate: λ



Model Instability: Forecasting Exchange Rate: α



Lessons from Model Instability

- Model instability may occur because of changes in the underlying dynamics between the explanatory and dependent variable
- Classical methods for breaks may be used to estimate the point of break
- Intercept adjustment (IIS or SIS) can be used to detect and correct for breaks (AutoMetrics)
- The above solutions generally do not work for ML methods
- We tune hyperparameters using cross validation and therefore we the parameters in LASSO and Elastic Net are a function of the hyperparameters
- It is generally a good idea to recursively forecast to capture the potential shifts in hyperparameters

Forecast Evaluation: Introduction

- In the previous econometrics classes you focused on the properties of the residual
- In our case we still need to look at the errors, however, our focus are the forecast errors
- We will discuss the properties of optimal forecasts

Optimal Forecasts Under MSE Loss

- Optimal forecasts under MSE loss need to satisfy the following three rules

1. Unbiasedness:

$$\mathrm{E}[e_{t+1|t}] = 0$$

2. h period ahead forecasts are uncorrelated with information at time t. 1 step ahead forecasts are serially uncorrelated

$$\mathrm{E}[e_{t+1|t} e_{t|t-1}] = 0$$

   If this condition is not satisfied, then forecast errors are predictable

3. The variance of the forecast errors increase with the horizon

$$var(e_{t+h+1|t}) \geq var(e_{t+h|t}), \text{ for } h \geq 1$$

Unbiasedness Tests

Simple test:

$$e_{t+1|t} = \alpha$$

$$H_0 : \alpha = 0 \text{ Unbiased}$$

Under the null of unbiasedness:

$$e_{t+1|t} = 0$$

$$y_{t+1} - f_{t+1|t} = 0$$

$$y_{t+1} = 0 + f_{t+1|t}$$

So we run the following regression:

$$y_{t+1} = \alpha + \beta f_{t+1|t} + u_{t+1}$$

And test the joint hypothesis:

$$H_0 : \alpha = 0 \ \& \ \beta = 1$$

This is called the Mincer Zarnowitz Regression

Unpredictable Forecast Errors

By the second condition

$$\mathrm{E}[e_{t+1|t}e_{t|t-1}] = 0,$$

forecast errors should be unpredictable. We can test this condition a few different ways.

1. We can test for serial correlation in forecast errors using the Ljung Box test
2. We can run a AR(1)

   $$e_{t+1|t} = \beta e_{t|t-1}$$

   We need to have $\beta = 0$

Forecast Comparison: Introduction

- Normally we end up with more than one forecasting model
- How can we compare the forecasting performance of two or more forecasting models

Forecast Comparison: Standard Metrics

- There are two metrics that are widely used in forecast comparison:
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- These metrics provide a tool to compare two or more forecasting model's out of sample performance

Forecast Comparison: RMSE and MAE

$$RMSE_m = \sqrt{\frac{1}{T}\sum_{t}^{T} e_{m,t}^2}$$

$$MAE_m = \frac{1}{T}\sum_{t}^{T}\sum_{t}^{T} |e_{m,t}|$$

- RMSE and MAE do not mean much by themselves
- They gain relevance when comparing two or more models

Equal Predictive Accuracy Test

The Equal predictive accuracy test is also known as the Diebold-Mariano (DM) test.

- Compare the losses of two different forecast generating models
- It is model agnostic, meaning we do not need to know anything about the model to compare two of them
- Vert simple idea: if your model is only as good as the other one, then you do not need to complicate things

Equal Predictive Accuracy Test

We construct the statistic of interest:

$$d_{t+h} = L(f_{1,t+h}, y_{t+h}) - L(f_{2,t+h}, y_{t+h})$$
$$d_{t+h} = e^2_{1,t+h} - e^2_{2,t+h}$$

And test:

$$d_{t+h} = \alpha + \varepsilon_{t+h}$$
$$H_0 : \alpha = 0, \text{ equal predictive accuracy}$$
$$H_a : \begin{cases} \alpha > 0 & \text{Model 2 is more accurate} \\ \alpha < 0 & \text{Model 1 is more accurate} \end{cases}$$

When running the regression we need to adjust for serial correlation in the error term, so we use Newey-West(1987) standard errors

Benchmarks in Forecasting

- In most cases, we are not the first researchers to come up with a forecasting model
- It is important to test the out of sample performance of our model against a certain benchmark
- Usually, we want to compare our forecast to an industry benchmark
- Which model is widely used?
- Which model is proven to be hard to beat?

Some Useful Benchmarks

- Almost in all cases, one of the following benchmarks is used:
  1. Random Walk: Exchange rate forecasting, stock prices
  2. AR(1): used across the board in finance and economics
  3. Prevailing Mean: Forecasting returns, profits
- If our model is not beating these models, then we should think of a new model, or just use the benchmark
- Simplicity is very important
- Additional complexity must be justified

Out of Sample R^2

- Out of Sample R^2 (R^2_{OoS}) is a non-formal way to compare the out of sample performance of a forecasting model against a benchmark
- It has been popularized by Goyal and Welch (2008)

$$R^2_{OoS} = 1 - \frac{MSE_{model}}{MSE_{benchmark}}$$

Positive values of R^2_{OoS} indicates that the model out performs the benchmark, and negative values indicates that the benchmark beats the model

Δ Cumulative SSE

- R^2_{OoS} gives us one number
- Could there be periods in which our model beats the benchmark?
- To visualize the periods in which our model outperforms or underperforms a benchmark we use the Δ Cumulative SSE
- Positive and increasing values represent moments where the model beats the benchmark, and vice versa

$$\Delta CumSSE_m = \sum_{\tau=\underline{t}}^{T} (e_{\tau,benchmark})^2 - \sum_{\tau=\underline{t}}^{T} (e_{\tau,m})^2$$