Statistical Machine Learning and Its Applications

Lecture 4: Resampling Methods

KAIST Mark Mintae Kim

Department of Industrial & Systems Engineering
KAIST

OUTLINE

- Overview
- Validation set approach
- Cross-validation
  - Leave-one-out cross validation
  - K-fold cross-validation
- Bootstrap

OUTLINE

- Overview
- Validation set approach
- Cross-validation
  - Leave-one-out cross validation
  - K-fold cross-validation
- Bootstrap

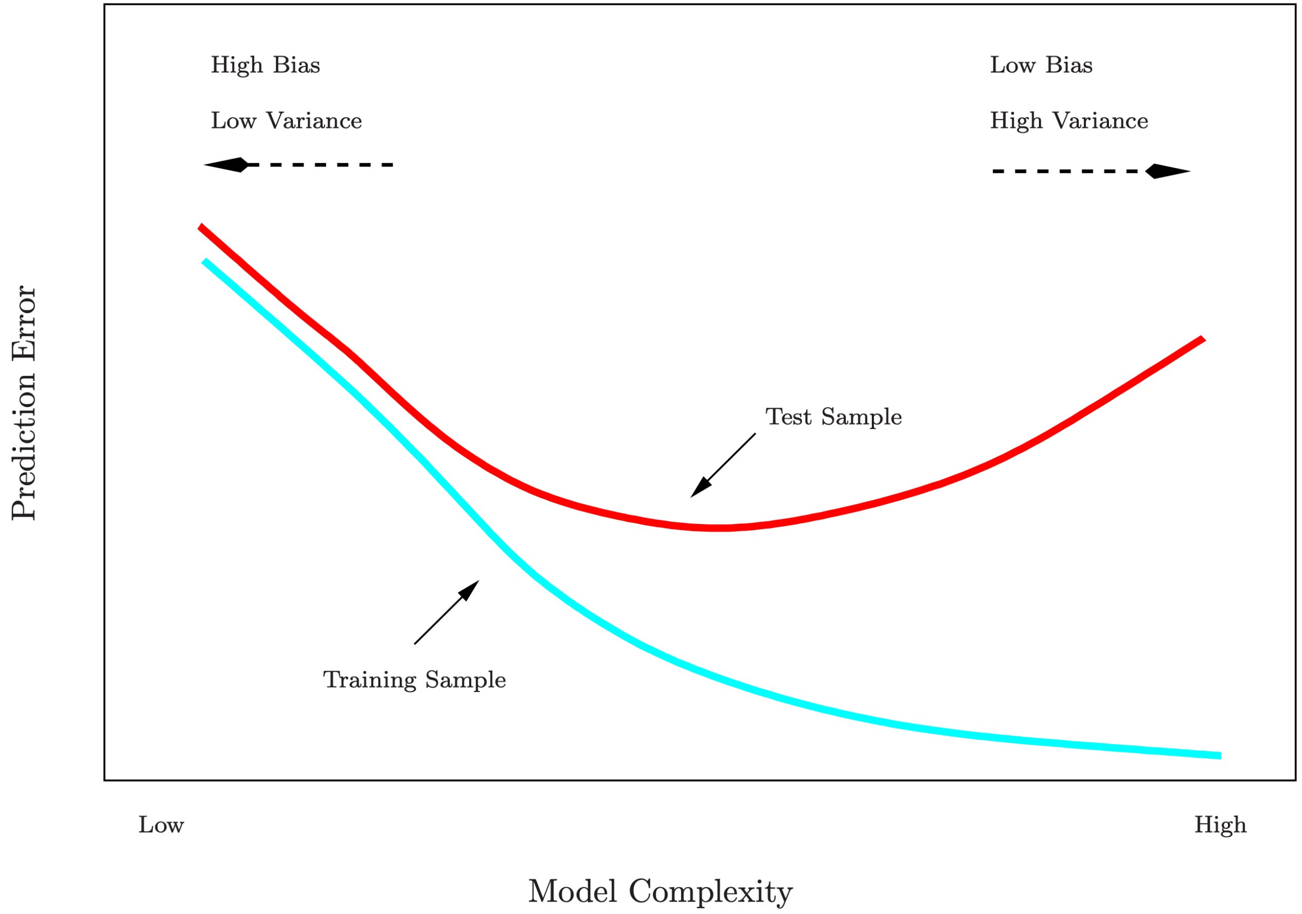
OVERVIEW: RESAMPLING

- Repeatedly drawing samples from training dataset
- Fitting a model of interest to these samples to obtain **additional information of the fitted model**
  - Example
    - Test-set prediction error
      - Using only training dataset is too optimistic
    - Variability of the estimated coefficients (bias and variance)
- Two most commonly used methods are **cross-validation** and **bootstrap**.
  - Cross-validation: Estimate the **test error** of models
  - Bootstrap: Quantify the **uncertainty** of estimators
- These both can be utilized in
  - Model assessment: The process of evaluating model performance.
  - Model selection: The process of selection the proper level of flexibility for a model.

TRAINING ERROR VS. TEST ERROR

- Recall the distinction between the training error, and the test error.
- **Training error**: Can be easily calculated by applying the ML method to the observations used in its training.
- **Test error**: The average error that results from using a ML method to predict the response on a new observation, one that was not used in training the method.
- But the training error rate often is quite different from the test error rate
  - In particular the training error can dramatically underestimate the test error.

TRAINING VS. TEST-SET PERFORMANCE



e.g., number of coefficients that we fit, polynomial degree

- Bias: How far off on the average the model is from the truth
- Variance: How much the estimates vary around their average?

Q. How can we find the best model?

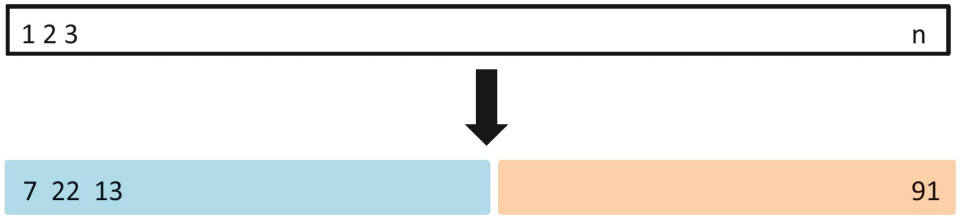
- **Best solution**: A large designated test set. But, often not available
- Alternatives: **Holding out** technique

OUTLINE

- Overview
- Validation set approach
- Cross-validation
  - Leave-one-out cross validation
  - K-fold cross-validation
- Bootstrap

VALIDATION SET APPROACH

- Randomly divide the available set of samples into two parts
  - **Training set** and **validation** or **hold-out set**

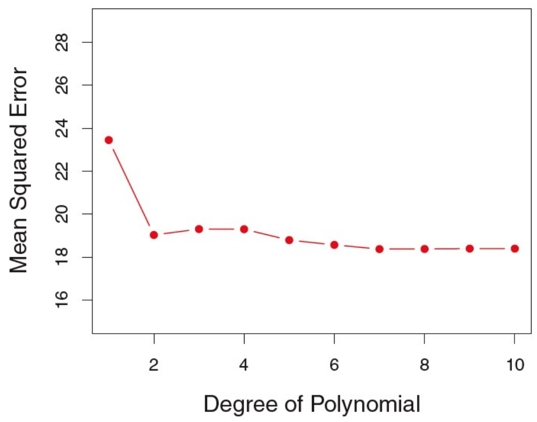


- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
- The resulting validation-set error provides an **estimate** of the test error
  - MSE for regression, and classification error for classification
- Drawbacks
  - The **test error rate** depends on which observations we used for training vs. testing
  - We are only training on a **subset** of the data
    - Validation set error may tend to **overestimate** the test error for the model fit on the entire data set (not enough training data)
- We need a new method!

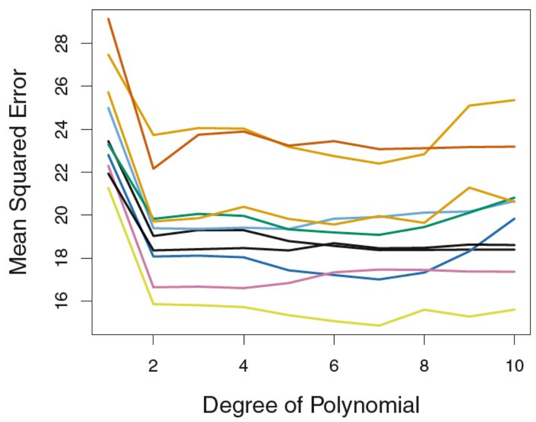
EXAMPLE: AUTOMOBILE DATA

- **Goal**: Compare linear vs high-order polynomial linear regression
- Recall
  - **Linear**: $y = \beta_0 + \beta_1 X$
  - **Quadratic**: $y = \beta_0 + \beta_1 X + \beta_2 X^2$
  - **Polynomials of degree p**: $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p$

Single split



Multiple split



How can we choose the best p?

Cross-validation!

- Validation set MSE (Single split)
  - Quadratic fit < linear fit
  - Cubic fit > quadratic fit
- **No consensus among the curves** as to which model results in the smallest validation set MSE.
  - The only thing we can be confident is Quadratic fit < linear fit

CROSS-VALIDATION

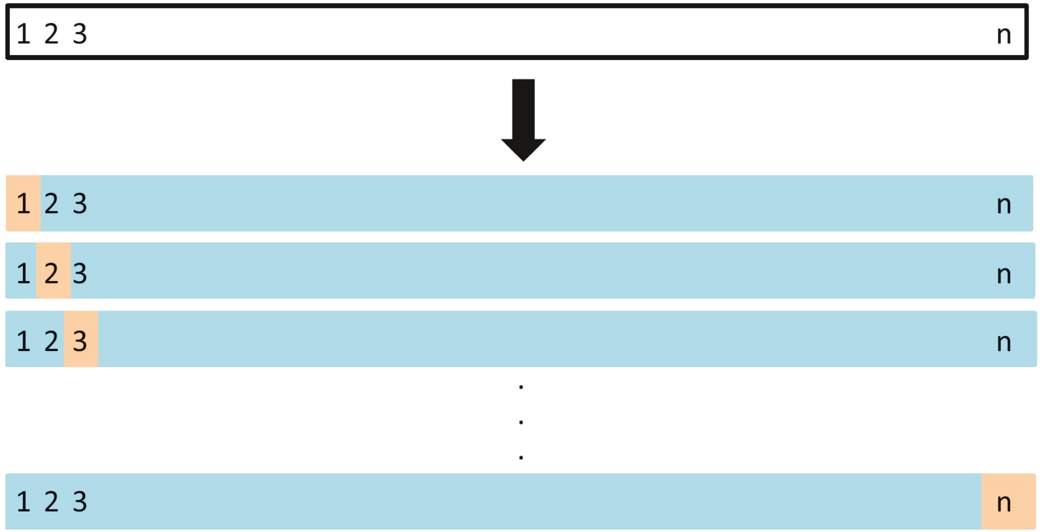
- Goal 1: Avoid sensitivity to test set selection
- Goal 2: Train on as much data as possible
- Approaches
  - Leave-one-out cross validation
  - K-fold cross validation

OUTLINE

- Overview
- Validation set approach
- Cross-validation
  - Leave-one-out cross validation
  - K-fold cross-validation
- Bootstrap

LEAVE-ONE-OUT CROSS VALIDATION (LOOCV)

- Suppose the data contain n data points.
- First, pick data point 1 as validation set, the rest as training set.
- Fit the model on the training set, evaluate the test error on the validation set → MSE_1
- … (repeat n times)
- Obtain an estimate of the test error by combining the MSE_i, $i = 1, \dots, n$

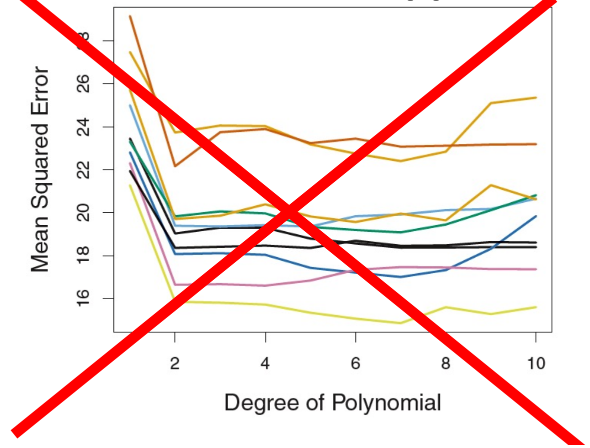


$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

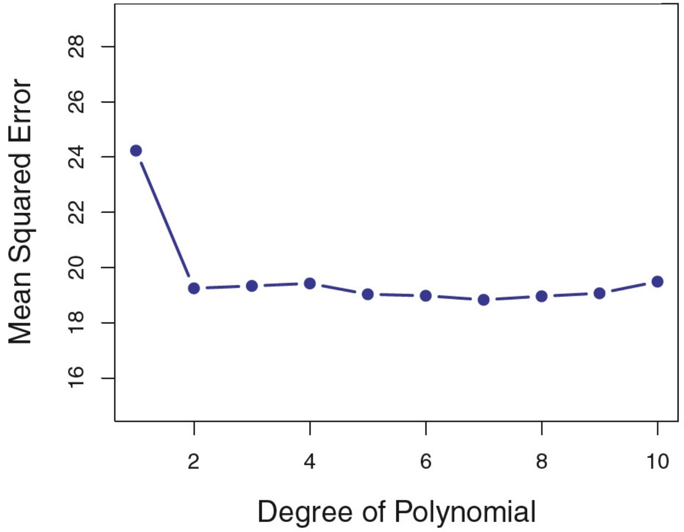
LEAVE-ONE-OUT CROSS VALIDATION (LOOCV)

- Advantage
  - Far less bias
    - Training data size $(n-1)$ is close to the entire data size (n)
    - In validation set approach, the training data size was $\frac{n}{2}$
    - LOOCV tends not to overestimate the test error (vs. validation set approach)
  - Low variance (variability) in the result (MSE)
    - No randomness in the training/validation set splits
- Disadvantage:
  - **Computationally expensive**
    - Especially if n is large, and each individual model is slow to fit
  - High variance in the model estimates
    - Doesn't shake up the data enough, which implies that the estimates from each fold are highly correlated
    - This means that the the estimate will vary if the training data changes

Validation set approach



LOOCV



CHEAP LOOCV FOR LEAST SQUARES REGRESSION (OPTIONAL)

- With least squares-based linear or polynomial regression, the following holds

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

  - y_i: Ground truth for ith sample
  - $\hat{y}_i$: Predicted value for ith sample using the whole dataset
  - $h_i \in [\frac{1}{n}, 1]$: Leverage statistic (Large value indicates an observation with high leverage)

How much an observation influences its own fit

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

h_i increases with the distance of x_i from $\bar{x}$

- A high leverage h_i implies that ith observation is influential
- The higher the leverage h_i, the more we penalize (make large) the MSE
- We can use h_i and MSE to calculate what the LOOCV error would be **without ever actually performing it!**
- But this only holds for least-squares regression

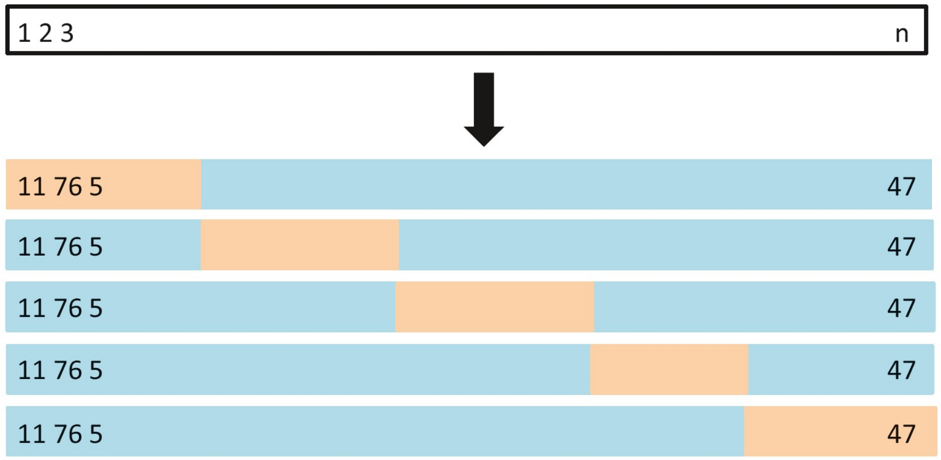
Check this link for proof https://robjhyndman.com/hyndsight/loocv-linear-models/

OUTLINE

- Overview
- Validation set approach
- Cross-validation
  - Leave-one-out cross validation
  - K-fold cross-validation
- Bootstrap

K-FOLD CROSS VALIDATION

- LOOCV is often too expensive on large datasets, but the same idea works even if we can't build n separate models
- Randomly divide n observations to K folds of approximately equal size
- Treat the first fold as a validation set, fit the model on each of the remaining $K - 1$ folds, compute MSE_1
- … (repeat K times)
- Obtain an estimate of the test error by combining the MSE_i, $i = 1, \dots, K$
- LOOCV is a special case of K-fold cross validation, actually n-fold cross validation.

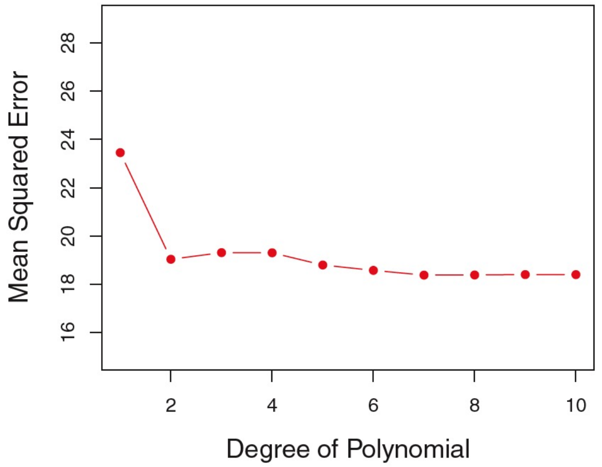


$$CV_{(k)} = \frac{1}{K}\sum_{i=1}^{K} MSE_i$$

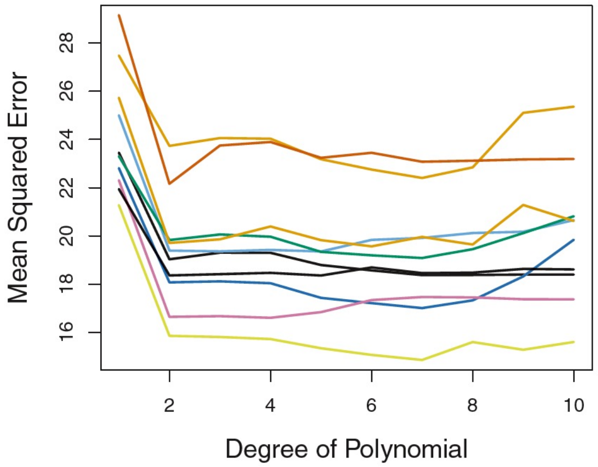
DIFFERENT APPROACHES TO VALIDATION

- **Validation set** vs. **LOOCV** vs. **10-fold CV**

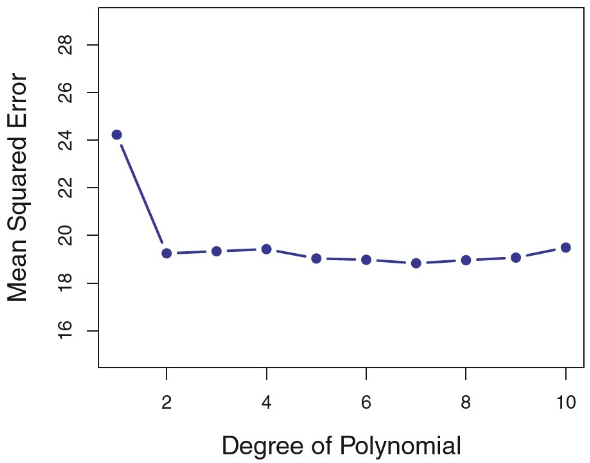
Validation set (Single split)



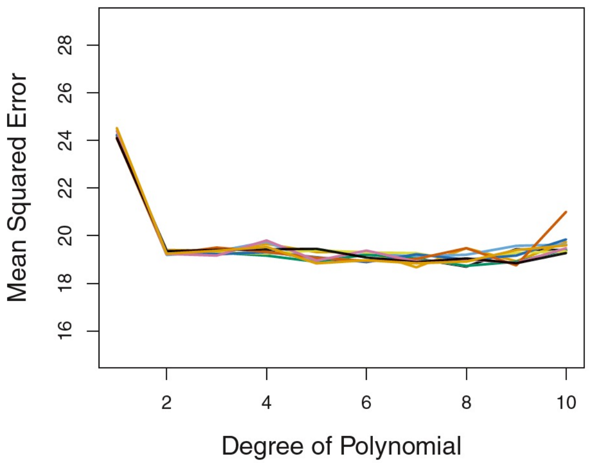
Validation set (Multiple splits)



LOOCV



10-fold CV

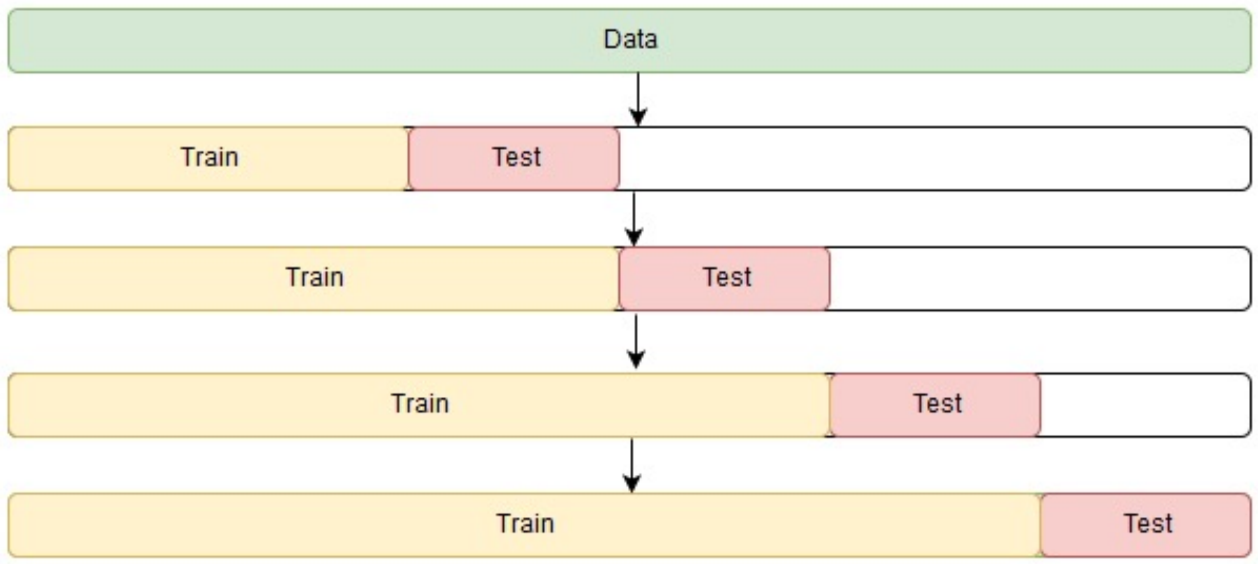


BIAS-VARIANCE TRADE-OFF FOR K-FOLD CROSS VALIDATION

- Recall bias-variance trade-off: $E\left(y_0 - \hat{f}(X_0)\right)^2 = Bias\left(\hat{f}(X_0)\right)^2 + Var\left(\hat{f}(X_0)\right) + Var(\epsilon)$
- LOOCV → Low bias
  - Gives approximately unbiased estimates of the test error, since each training set contains $n-1$ observations
    - Almost as many as the number of observations in the full data set.
- K-fold CV → Higher bias
  - Each training set contains $(k-1)n/k$ observations—fewer than in the LOOCV approach
- But, we also need consider variance!
  - LOOCV → High variance
  - ∵ Variance: highly correlated quantities > less correlated quantities
    - Var(X+Y) = Var(X) + Var(Y) + 2COV(X, Y)
  - K-fold CV → Low variance
- $K = 5\ or\ 10$ provides a good compromise for this bias-variance tradeoff.

5-FOLD CV FOR TIME SERIES DATA

- Time series
  - Example: Stock price



CROSS-VALIDATION ON CLASSIFICATION PROBLEMS

- So far we've only talked about regression
- Cross-validation works in classification in the same manner as in regression with the exception that MSE is replaced by the misclassification rate.
- LOOCV in classification

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

- K-fold CV in classification

$$CV_K = \frac{1}{n}\sum_{k=1}^{K} \frac{n_k}{n} \sum_{i \in C_k} I(y_i \neq \hat{y}_i)$$

  - $C_1, \dots, C_K$: K roughly equally divided dataset
  - n_k: number of samples in k-th fold

OUTLINE

- Overview
- Validation set approach
- Cross-validation
  - Leave-one-out cross validation
  - K-fold cross-validation
- Bootstrap

THE BOOTSTRAP: OVERVIEW

- Used to quantify the **uncertainty** associated with the estimated parameters
  - Example: Standard error of the estimated parameters

Linear model ➡ $SE\left(\hat{\beta}_0\right)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_i^n (X_i - \bar{X})^2}\right], SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_i^n (X_i - \bar{X})^2}$

Other models ➡ **Bootstrap**

- Repeatedly resample the original data to get a “new” dataset
- Perform resampling with replacement
  - Each observation may appear more than once in the resampled dataset
- Fit our model to each of the resampled dataset, and combine them

EXAMPLE: INVEST MONEY IN TWO FINANCIAL ASSETS

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y
- We will invest a fraction α of our money in X, and will invest the remaining $1-\alpha$ in Y
- Goal: Choose α to minimize the total risk, or variance, of our investment
  - That is, minimize $\boldsymbol{Var(\alpha X + (1-\alpha) Y)}$

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

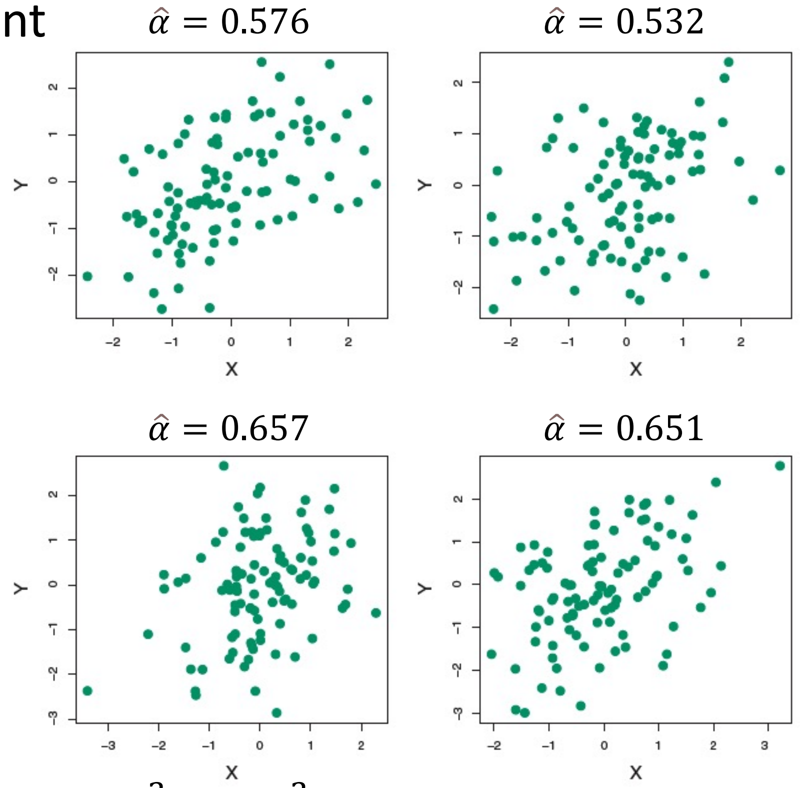
$$\sigma_X^2 = \mathrm{Var}(X)$$
$$\sigma_Y^2 = \mathrm{Var}(Y)$$
$$\sigma_{XY} = \mathrm{Cov}(X, Y)$$

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}$$

If we repeat this process 1,000 times

$$\bar{\alpha} = \frac{1}{1{,}000} \sum_{r=1}^{1{,}000} \hat{\alpha}_r = 0.5996$$

$$\sqrt{\frac{1}{1{,}000-1} \sum_{r=1}^{1{,}000} (\hat{\alpha}_r - \bar{\alpha})^2} = 0.083$$



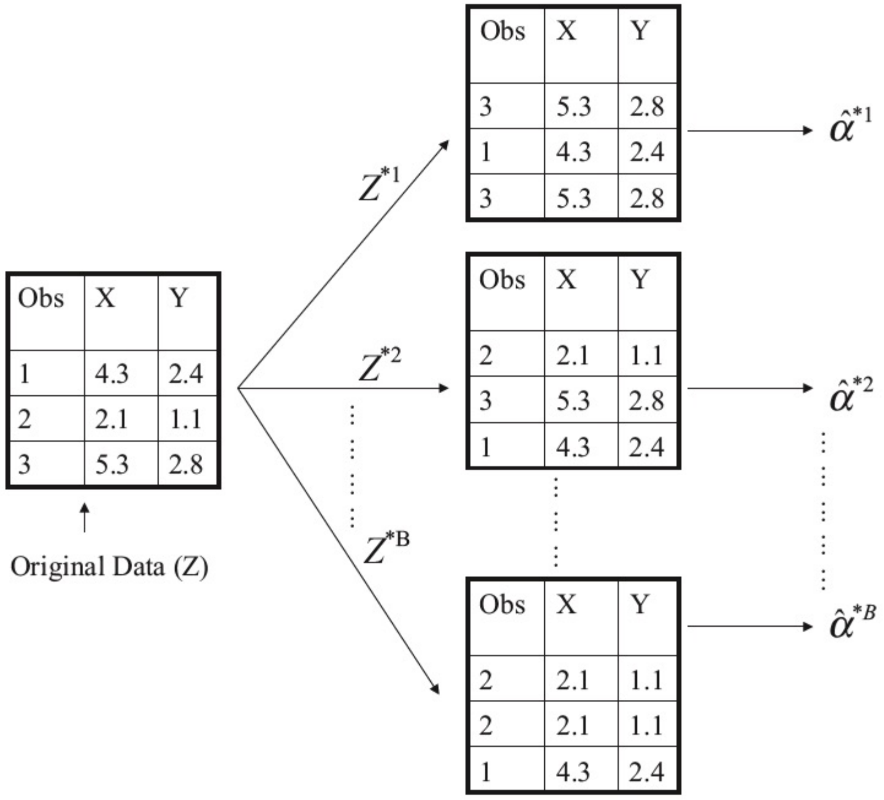
$\sigma_X^2 = 1, \sigma_Y^2 = 1.25,$ and $\sigma_{XY} = 0.5$

True $\alpha = 0.6$

However, we do not know the real distribution

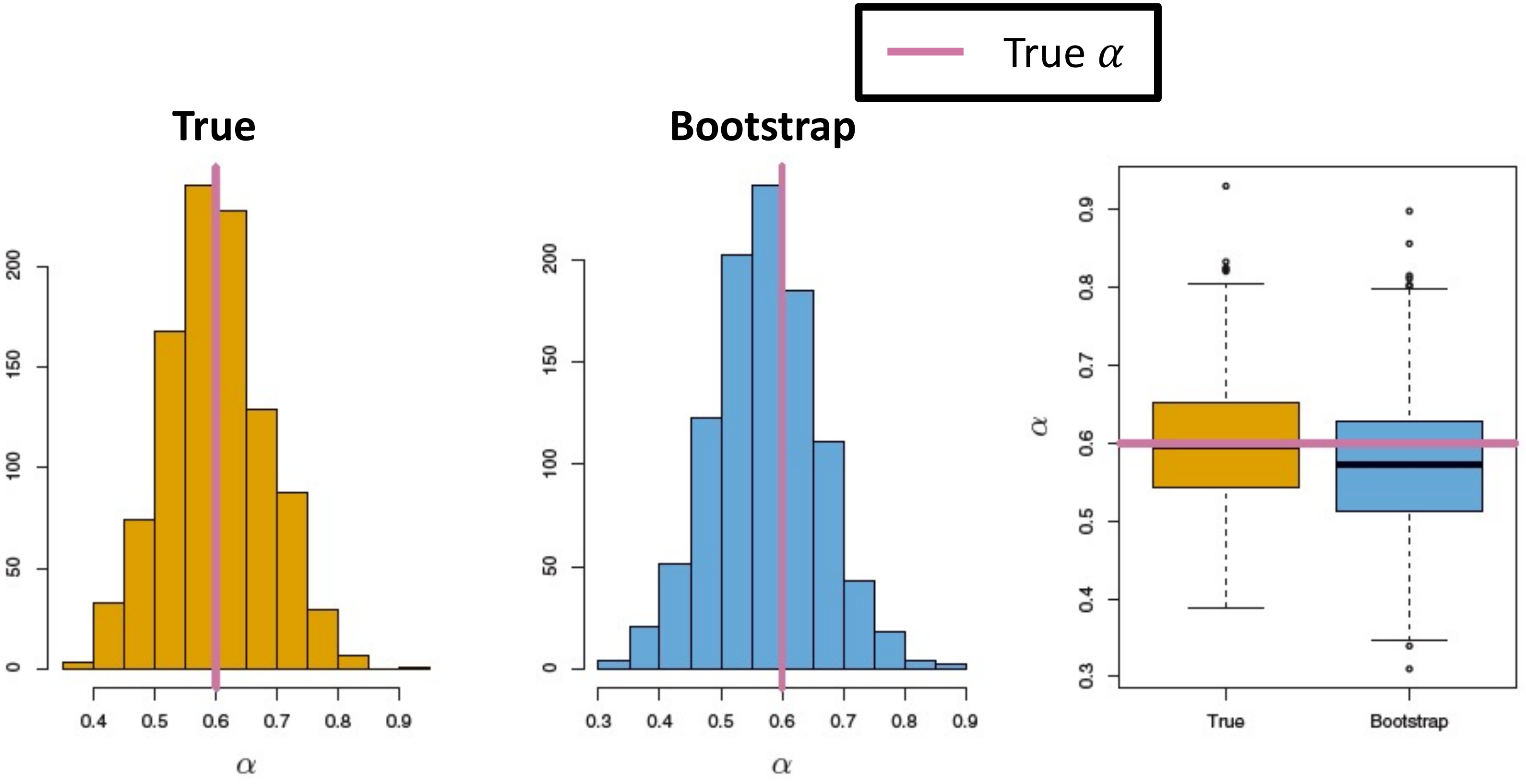
EXAMPLE: INVEST MONEY IN TWO FINANCIAL ASSETS

- We instead rely on **bootstrap**



$$\bar{\hat{\alpha}} = \frac{1}{B}\sum_{b=1}^{B} \hat{\alpha}^{*b}$$

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\alpha}^{*b} - \bar{\hat{\alpha}}\right)^2}$$



THE BOOTSTRAP: FORMAL DEFINITION

- Suppose we are interested in the standard error of an estimator $\hat{\theta}$ of the underlying parameter θ.
- The bootstrap estimate of the standard error is computed as follows:
  - From the initial sample of n observations, re-sample independently n observations **with replacement** B times, where B is then the number of bootstrap samples.
  - Estimate $\hat{\theta}_b$ from each bootstrap sample for each all $b = 1, \dots, B$
  - Bootstrap standard error is then

$$SE_B(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b - \bar{\hat{\theta}}\right)^2}$$

  - where $\bar{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$

CONCLUSION

- Validation set approach
  - Train vs. Test
- Cross-validation
  - Leave-one-out cross validation
  - K-fold cross-validation
  - Advantages and disadvantages
- Bootstrap

Coming up next:

Linear Model Selection and Regularization