STAT5003

Week 6 : Cross validation and bootstrapping

Dr. Justin Wishart
Semester 2, 2020

THE UNIVERSITY OF SYDNEY

Readings

- Cross validation and bootstrap covered in Chapter 5 in James, Witten, Hastie, and Tibshirani (2013)

Training error vs test error

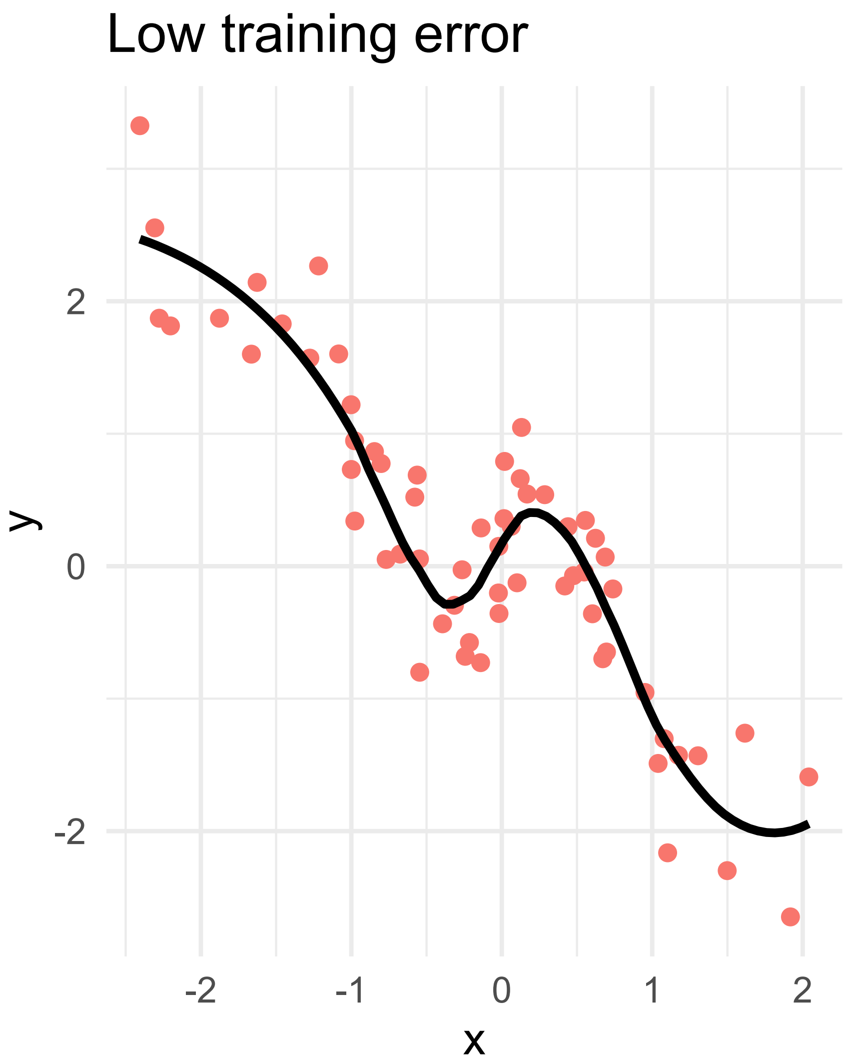
Training error vs test error

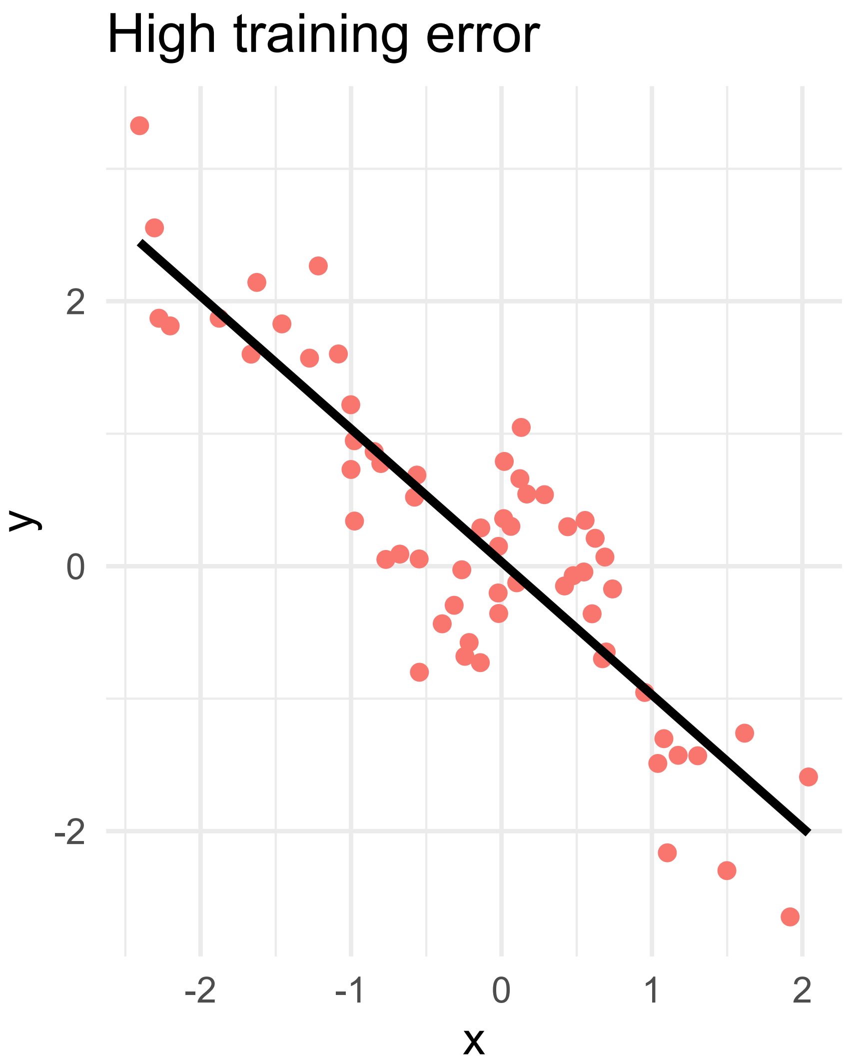
Training error is the performance metric applied to the observations used to train the model.

Test error is the average error when applying a model to predict the response on new (test) observations that were not used in the training of the model.

- Training error is usually very different in magnitude to the test error.
  - Training error can **underestimate** the test error.

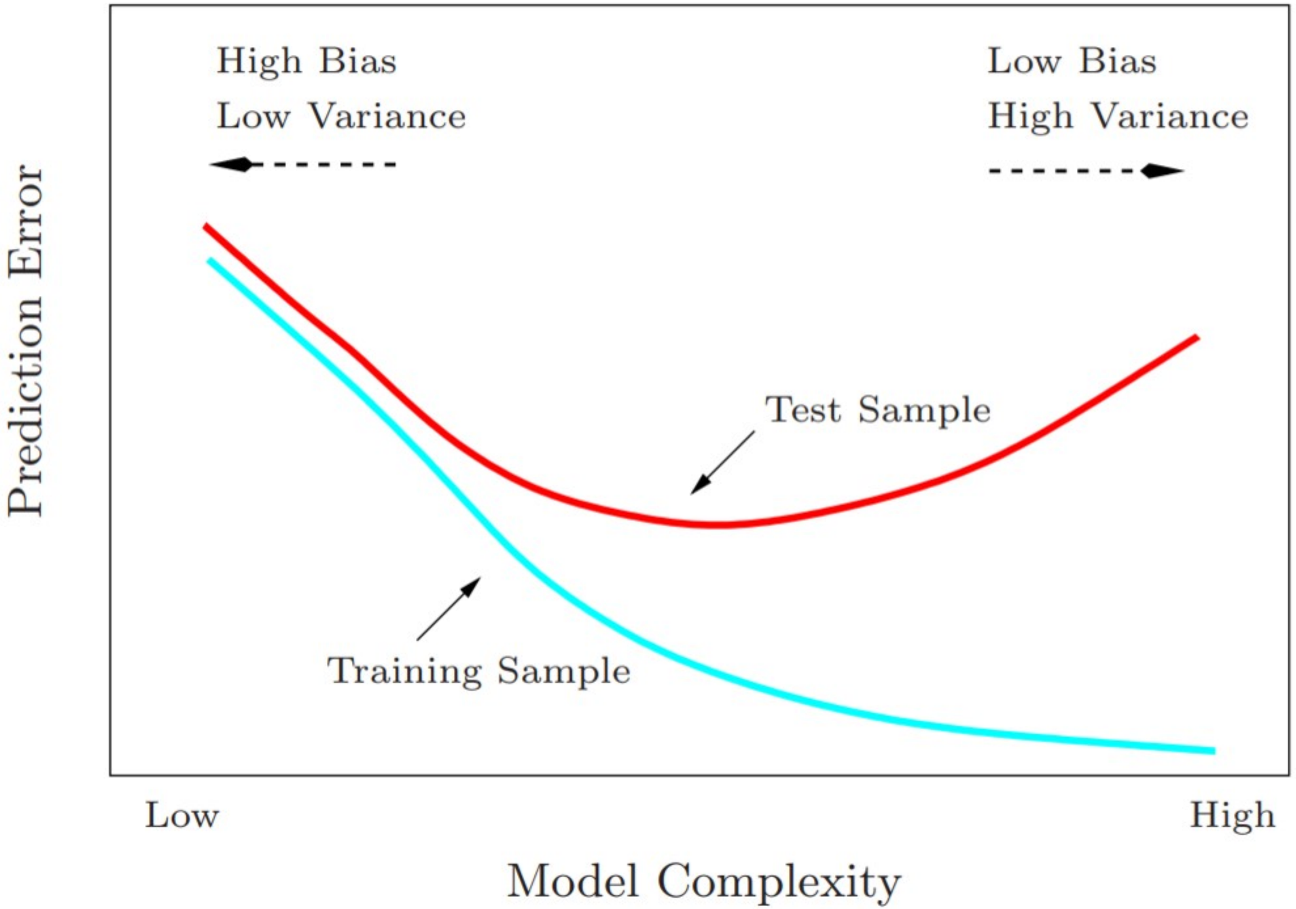
Pick the better model





Pick the better model

Training set vs Test set error



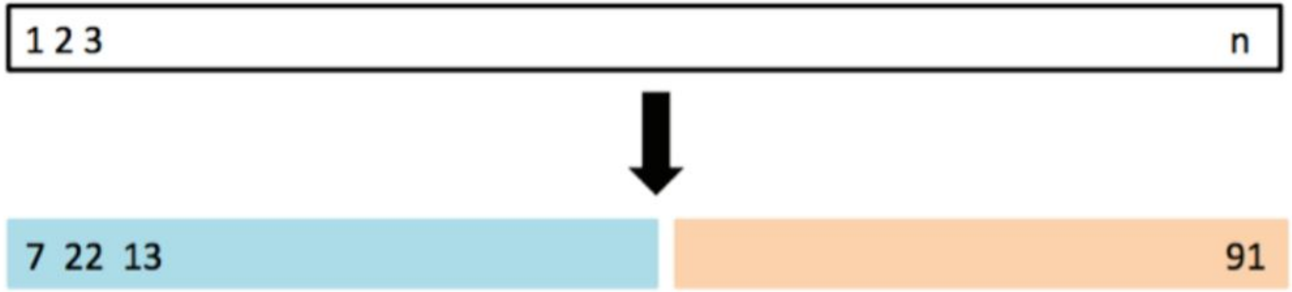
Estimate the test error

- Gold standard:
  - Use a large designated test set. Often not available
- Adjust the training error to estimate the test error
  - Common to add a penalty term to the model
    - BIC
    - Adjusted R^2
- Cross validation
  - Remove or hold out a subset of observations (test set) and use the rest to train the model.
  - Assess model performance on the test set.

Test Set approach

- Here we randomly divide the available set of samples into two
    - a training set
    - test set
- The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the test set.
- The resulting test-set error provides an estimate of the test error. Typically assessed using
    - MSE in the case of a quantitative response
    - Misclassification rate in the case of a qualitative (discrete) response.

Example of the training and test split



- Random split of the data into two halves
    - The left is the training indices
    - The right is the test indices

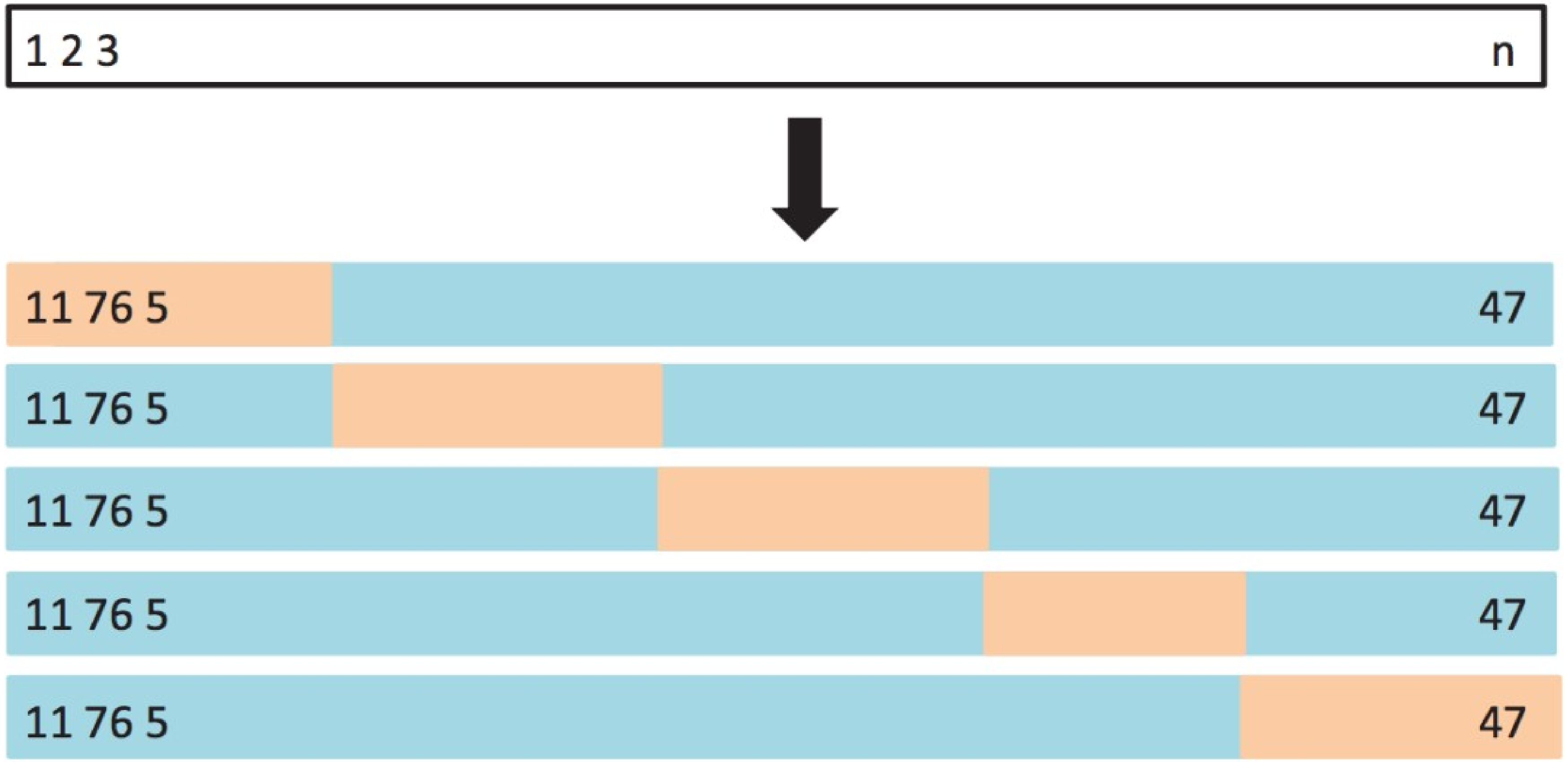
Drawbacks of test set approach

- The estimate of the test error can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the test set.
- In the test set approach, only a subset of the observations are used to fit the model.
    - This suggests that the test set error may tend to overestimate the test error for the model fit on the entire data set.

K-fold cross validation

- Widely used approach for estimating test error.
  - Estimates can be used to select best model, and to give an idea of the test error of the final chosen model.
- Idea is to randomly divide the data into K equal-sized parts.
  - We leave out part k, fit the model to the other $K - 1$ parts (combined), and then obtain predictions for the left-out k^{th} part.
- This is done in term for each part $k = 1, 2, \ldots, K$ and then the results are combined.

Example: 5-fold



Cross-validation formula

- Let the K parts be $C_1, C_2, \ldots, C_K$, where C_k denote the indices of the observations in part k.
  - There are n_k observations in part k:
  - if n is a multiple of K, then $n_k = \frac{n}{K}$
- Compute

$$CV_k = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

  - where $MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$
  - $\hat{y}_i$ is the fit for observation i obtained from the data with part k removed.

Cross-validation for classification problems

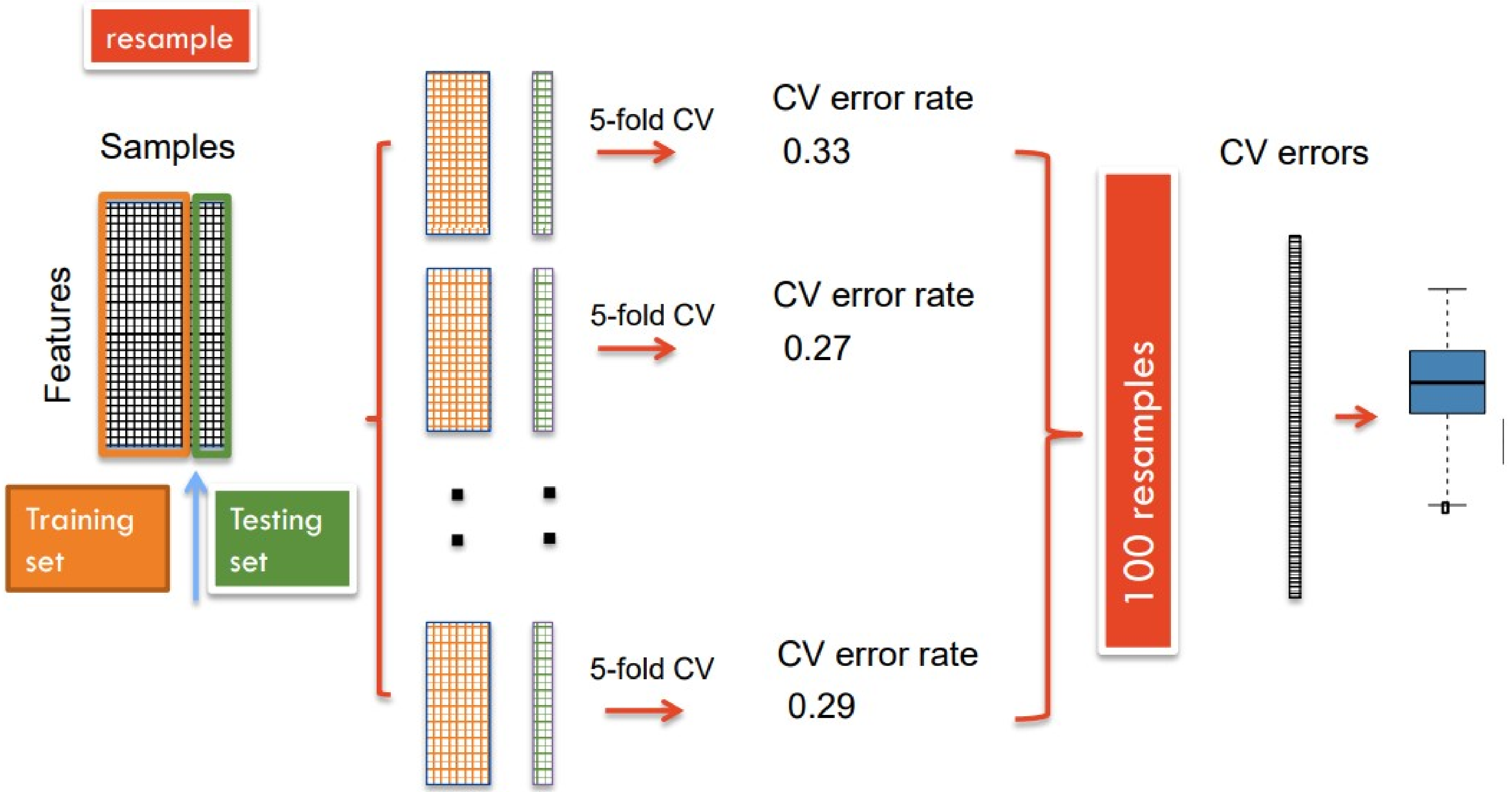
- For classification problems, we can compute the accuracy for each fold by calculating:

$$CV_K = \sum_{k=1}^{K} \frac{n_k}{n} A_k$$

  where the terms are

  - n : The total number of observations in the dataset
  - n_k : The number of observations in the belonging to class k
  - A_k : The accuracy of the classifier in fold k
    - e.g. $A_k = \frac{1}{n_k} \sum_{i \in C_k} 1_{\{\hat{y}_i = y_i\}}$

Repeated Cross validation

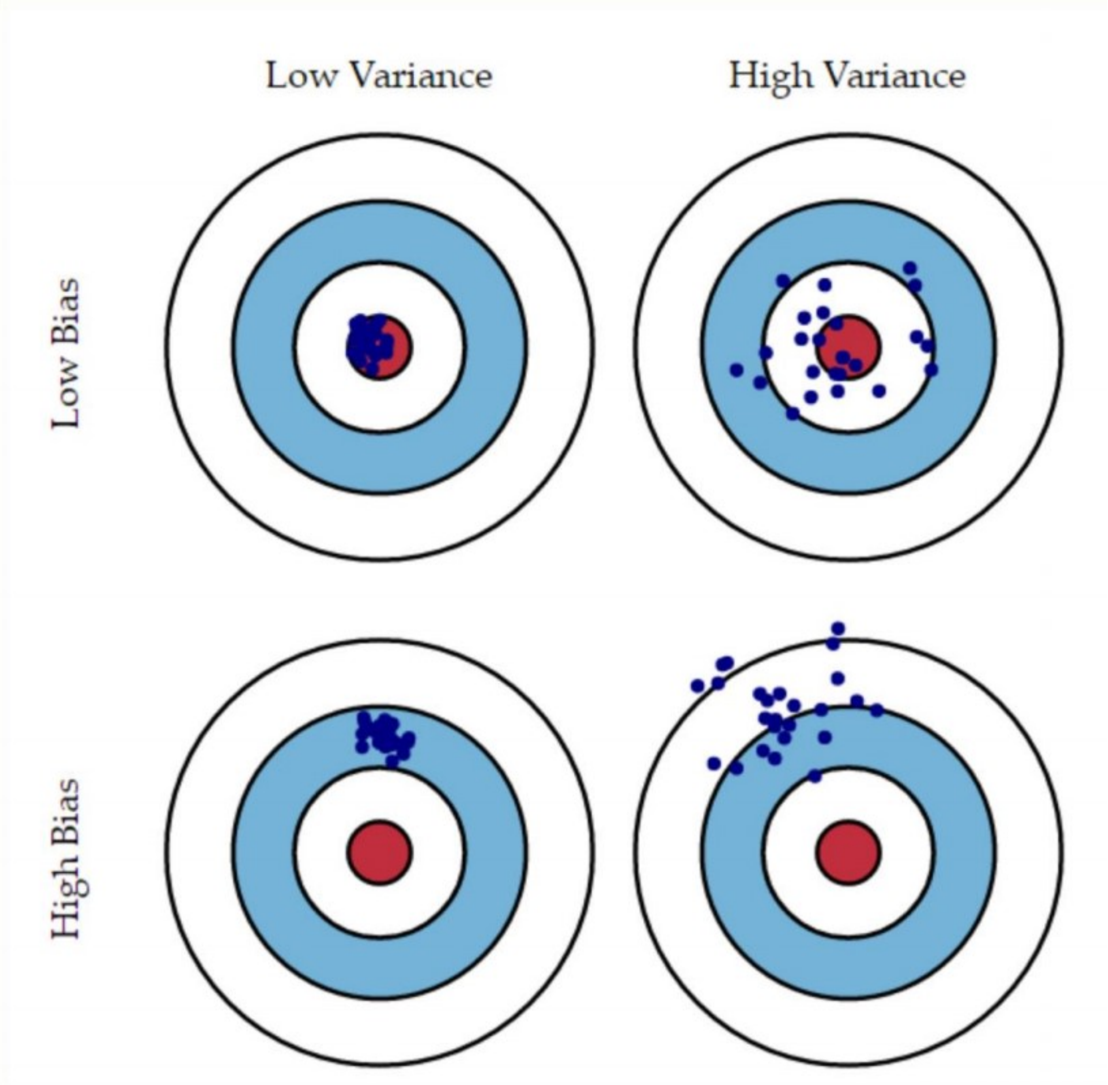


Repeated cross validation properties

In general, repeated CV provides a less biased CV error estimate

- Repeated CV also gives you the variance of the CV error
- However, it comes with a computational cost
- Implemented in the `caret` package in R

Dart board interpretation of bias & variance



Example of CV procedure

Consider a problem where you have a high dimensional data set, all entirely numeric, and need dimension reduction to proceed.

- You decide to reduce the dimensions of the data and use the following CV procedure:

1. Compute correlation matrix, select the top 50 variables that have the highest correlation with the response.
2. Use these 50 variables as features and perform K-fold cross validation

Issue with the previous slide

- Variable selection performed once using both the training and the test datasets
- Information can leak from the test to the training set
- Hence, the CV error estimate is likely to be biased.
- Ideally you shouldn't use the test data in any way in the training step.
  - If absolutely necessary some pre-processing on the features can be done so long as it doesn't involve the response variables.

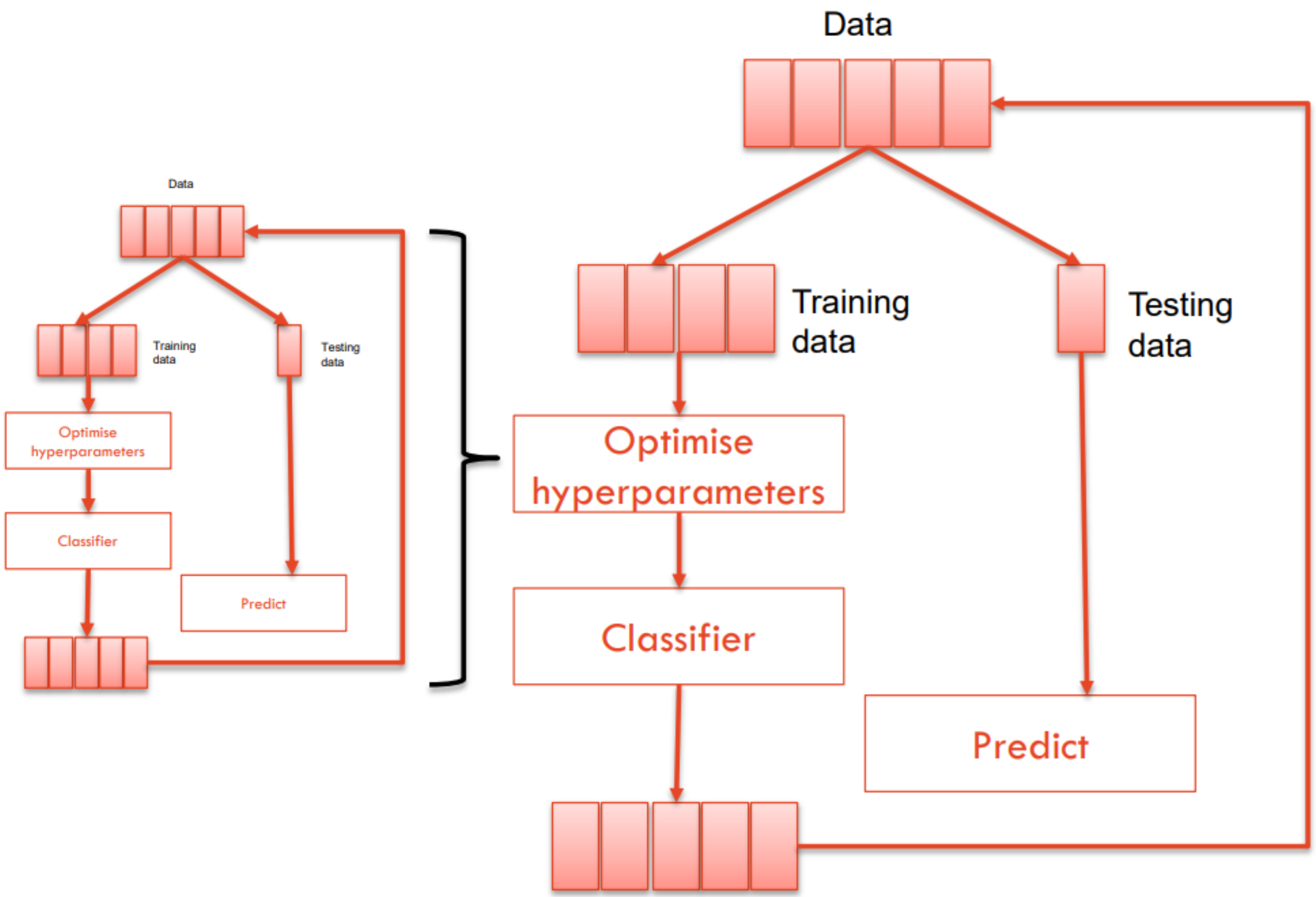
Corrected CV procedure

- Split the dataset into K folds
- For each $k = 1, 2, \ldots, K$
  - Determine the variables that correlate the best with the response using all the data except the data in fold k
  - Train your model using the selected variables above.
  - Run your classification algorithm and record accuracy against the test set.

Other information leakage to check

- Other things you should not do once but do it within with CV loop
  - Feature selection
  - Hyperparameter optimization
  - Missing data imputation
- Another method is nested cross validation

Nested cross validation



Final model building

- The reason for doing cross-validation is to evaluate the different models by estimating their performance on unseen data
- Example. If you need to choose between kNN, LDA and logistic regression and SVM, then you can run each of these classification algorithms with cross-validation, and pick the one with the highest CV accuracy
- But then, you can go back to use all the data to build a final model

Classification accuracy

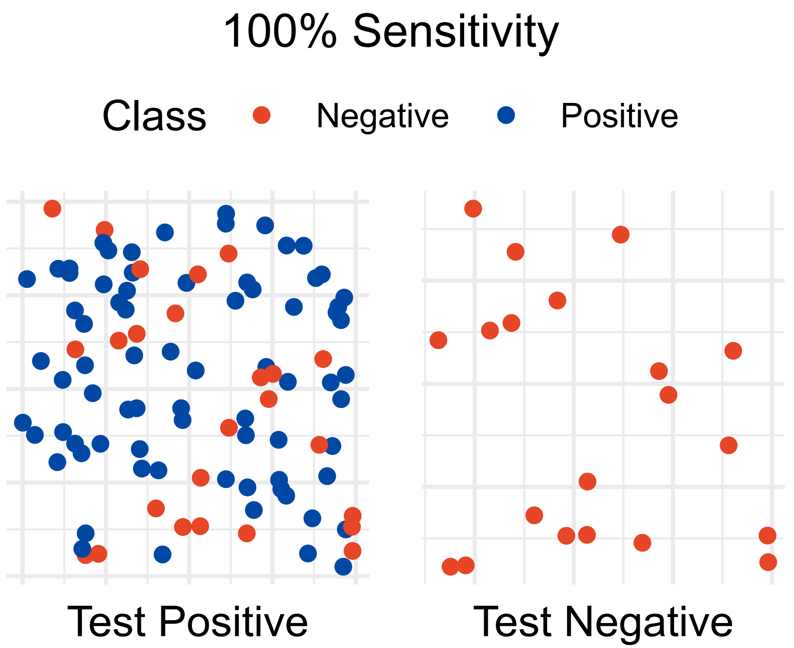
- Overall classification accuracy:
- Disadvantages:
  - Makes no distinction about the type of errors being made.
    - In spam filtering, the cost of erroneous deleting an important email is likely to be higher than incorrectly allowing a spam email past a filter.
  - Does not consider the natural frequencies of each class

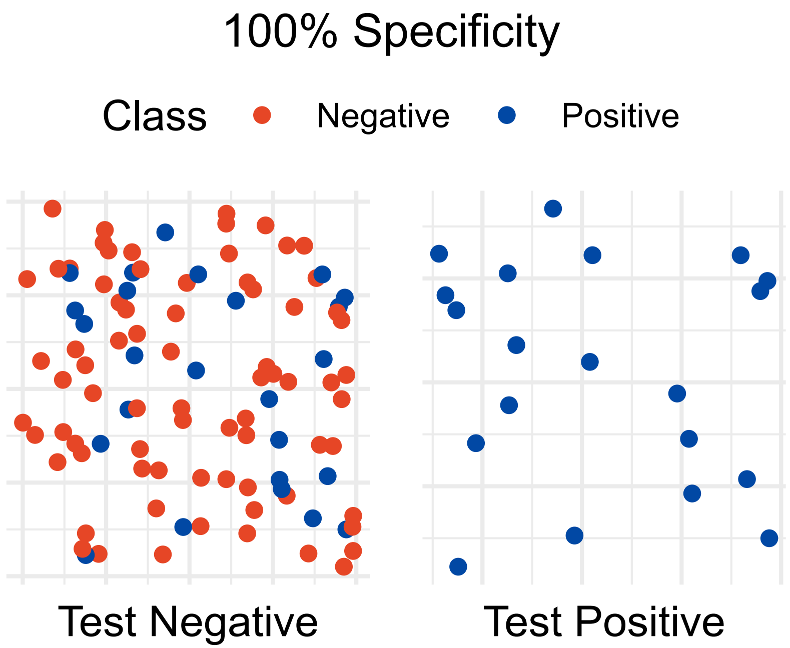
Confusion Matrix

	Actual	
	True	False
Predicted		
True	True Positive	False Positive
False	False Negative	True Negative

- True positive: Are positive class and predicted to be positive class
- False positive: Are negative class but predicted to be positive class
- False negative: Are positive class but predicted to be negative class
- True negative: Are negative class and predicted to be negative class

Sensitivity and Specificity

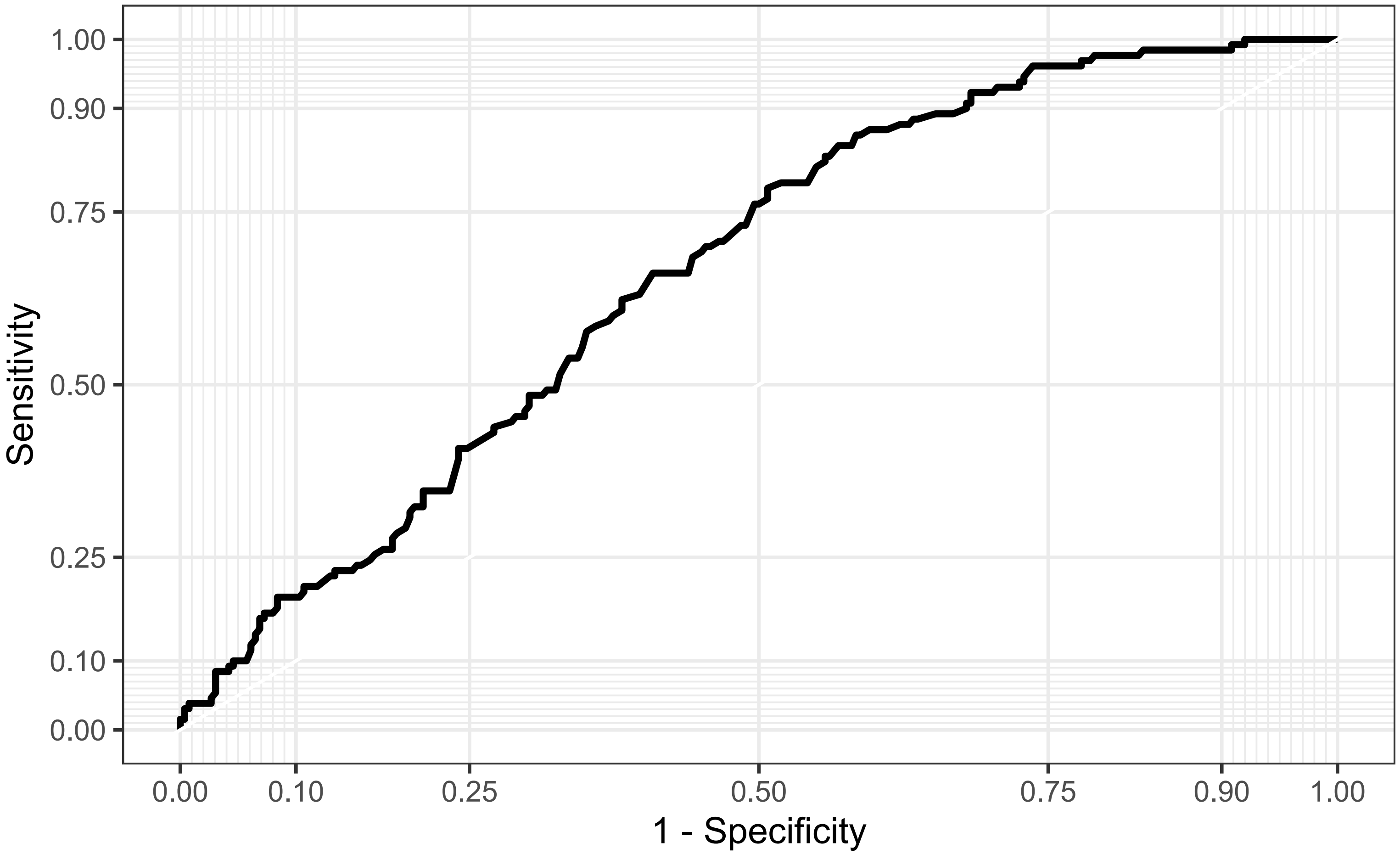




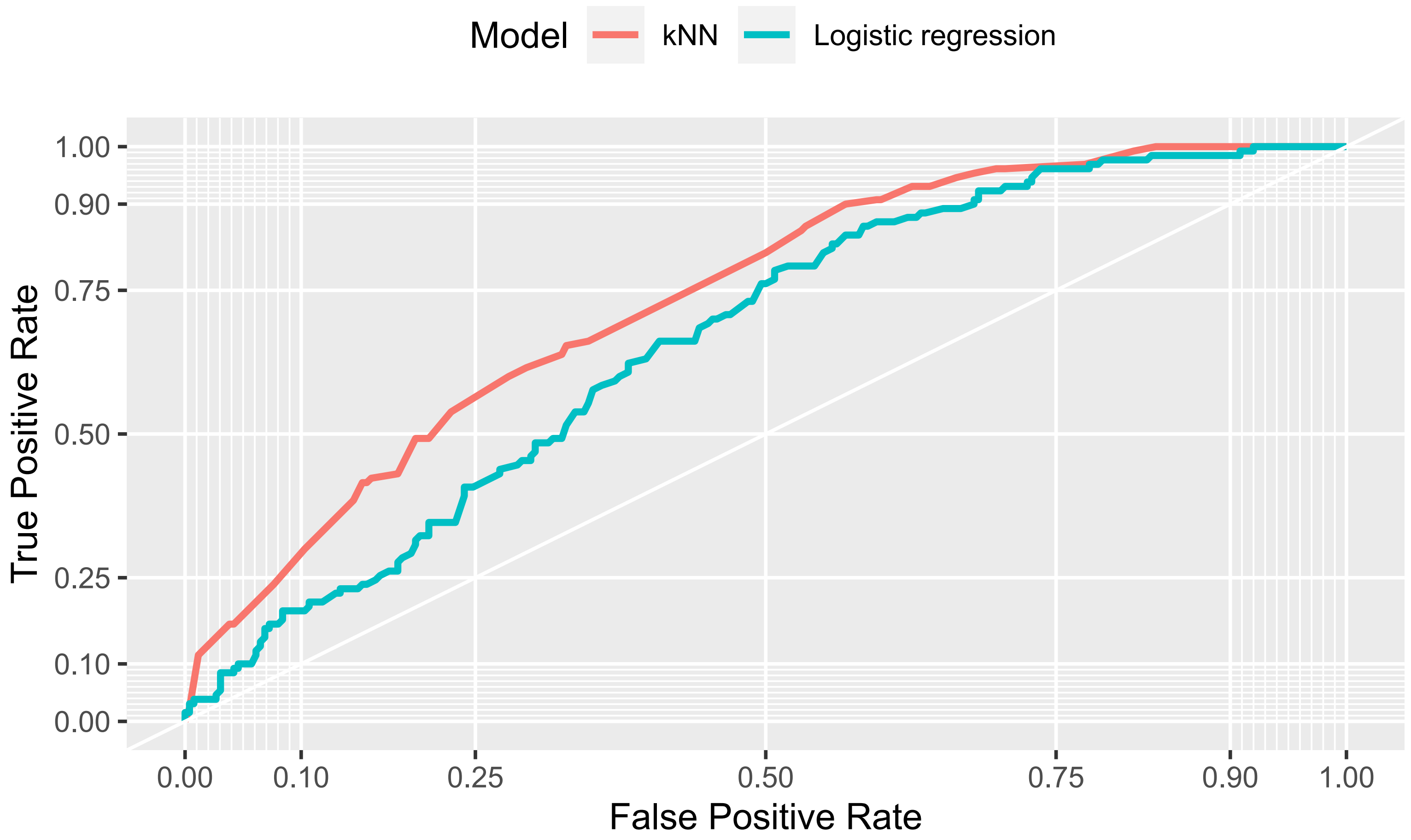
- Accuracy $= \frac{(TP+TN)}{(TP+FP+FN+TN)}$
- Sensitivity $= \frac{TP}{(TP+FN)} = \frac{TP}{P}$
- Specificity $= \frac{TN}{(TN+FP)} = \frac{TN}{N}$

- Precision $= \frac{TP}{(TP+FP)}$
- Recall $= \frac{TP}{(TP+FN)} = \frac{TP}{P}$
- $F_1 = \frac{2\,\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (Harmonic mean)
- GM $= \sqrt{\text{Precision} \times \text{Recall}}$ (Geometric mean)

Receiver Operating Characteristics (ROC) curve



Comparing ROC curves



Bootstrap

Bootstrap

- The bootstrap is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method.
- For example, it can provide an estimate of the standard error of a coefficient, or a confidence interval for that coefficient.

Bootstrap resampling algorithm

- Essentially sampling with replacement

Simple example

- Suppose that we wish to invest a fixed sum of money in two financial assets that yield returns of X and Y where X and Y are random quantities.
- The goal is to create a portfolio by investing fraction α of our wealth in X and $(1-\alpha)$ in Y.
- Want to choose to minimise the total risk of the investment. Mathematically this involves minimising $Var(\alpha X + (1-\alpha)Y)$
- The solution to this problem (calculus) is,

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \tag{1}$$

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

Example

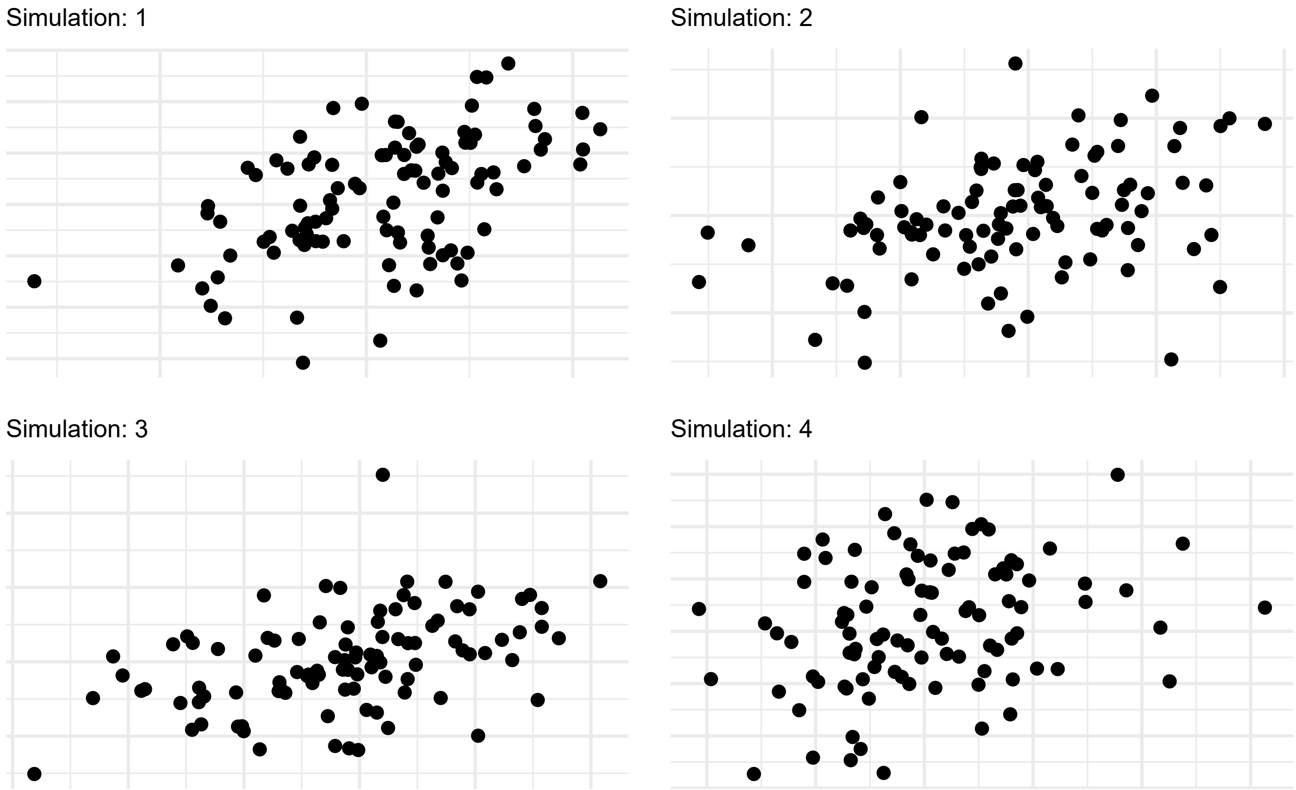
- The values of σ_X^2, σ_Y^2 and σ_{XY} are unknown but estimates can be computed from the data.
- The estimate of α that minimises the variance of the investment can then be computed with

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

- Suppose that X and Y can be sampled from the population repeatedly
- To estimate the standard deviation of $\widehat{\alpha}$, paired observations (X, Y) can be repeated simulated, say 100 pairs to get a single estimate of α. Repeat this process to get 1,000 estimates for α.
- Denote these estimates $\widehat{\alpha}_1, \widehat{\alpha}_2, \ldots, \widehat{\alpha}_{1000}$

Bootstrap simulations

- Consider example with $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.5$ and $\sigma_{XY} = 0.5 \Rightarrow \alpha = 2/3$.



- Each panel shows 100 simulated returns. From left to right, top to bottom, the estimates for α are 0.659, 0.683, 0.726, 0.68.

Parameter estimates

- Consider the mean of all the parameter estimates

$$\overline{\widehat{\alpha}} = \frac{1}{1,000}\sum_{k=1}^{1000}\widehat{\alpha_k} = 0.6662595$$

  - This is close to the true value of 0.6666667

- Estimate of the standard error using the standard deviation of all the estimates.

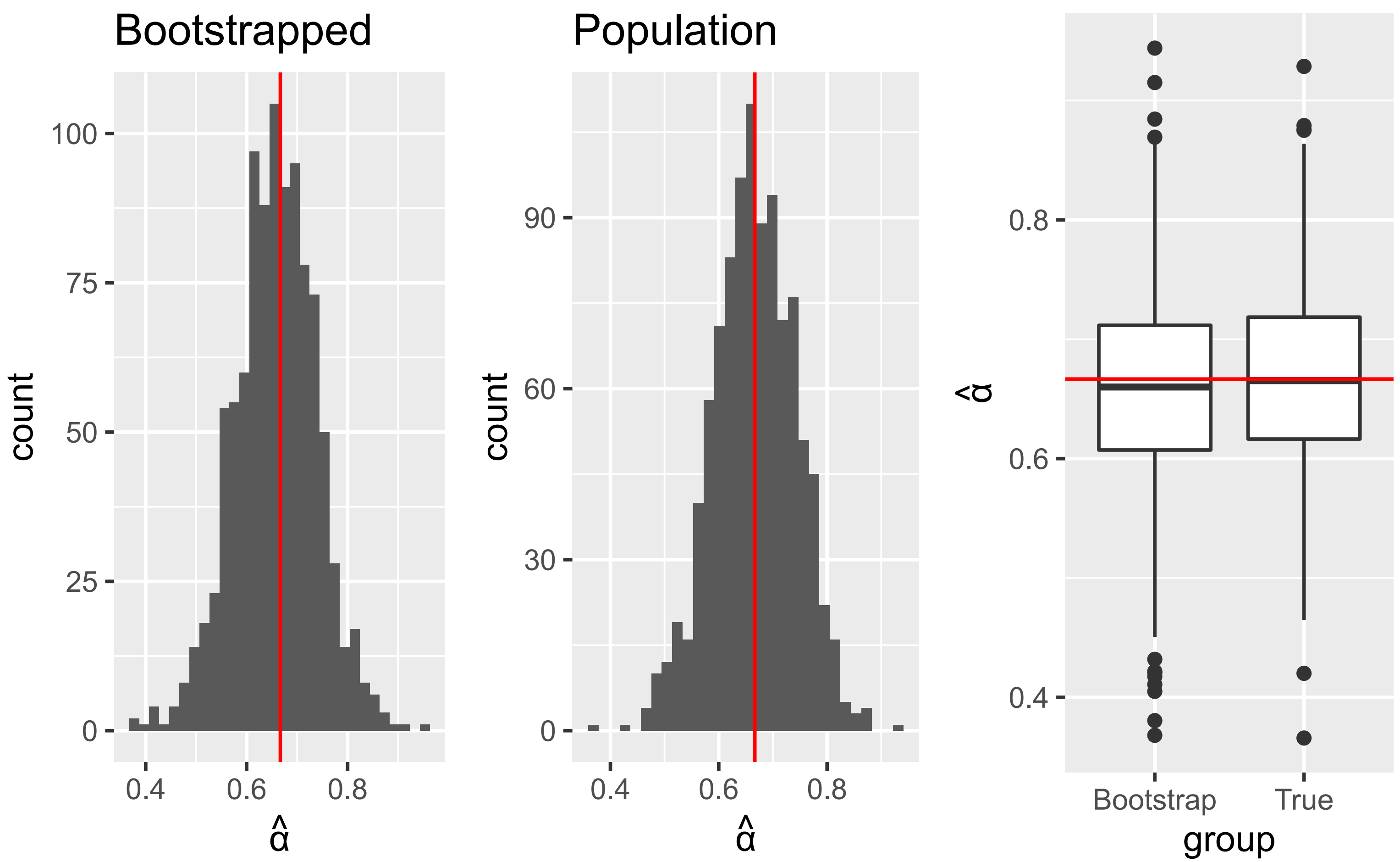
$$\sqrt{\frac{1}{1000-1}\sum_{k=1}^{1000}\left(\widehat{\alpha_k} - \overline{\widehat{\alpha}}\right)^2} = 0.0760217$$

- This gives an intuitive description of the reliability of the estimator.
  - For a random sample the estimate would vary around the true value by 0.0760217

Application in reality

- Cannot apply this directly in reality
    - cannot generate new observations from the population model.
- Bootstrap attempts to mimic this process
- Instead of sampling new independent observations from the population
    - Re-sample observations from the data *with replacement*
- Some observations appear more than once and some not at all

Results bootstrap vs population



References

James, G, D. Witten, T. Hastie, et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.