BEE3066 – Machine Learning for Economics

Topic 4 : Re-sampling Methods

Pradeep Kumar
p.kumar@exeter.ac.uk

University of Exeter Business School

Introduction

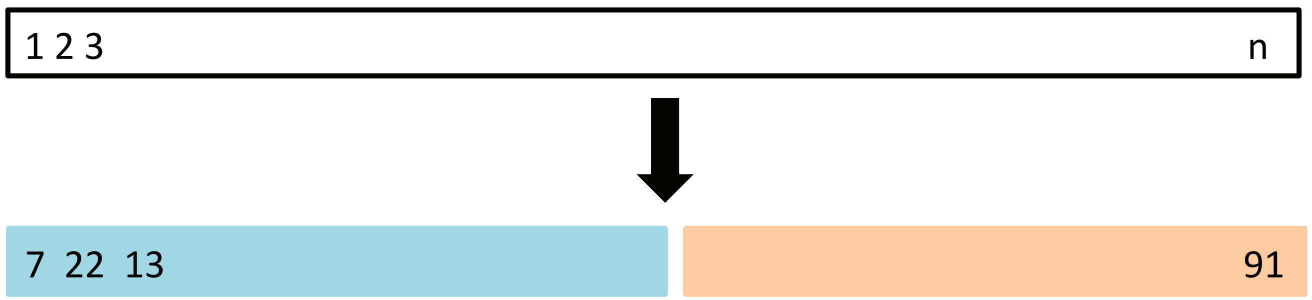
- **Resampling** involves drawing samples from the training data and re-fitting the model of interest.
- **Resampling methods** allow us to obtain information that is not available from fitting the model once e.g. variability in a regression model.
- Computationally intensive !
- Most commonly used re-sampling methods: **Cross-validation** and **Bootstrap**.
- **Cross-validation** is used for
  - **Model assessment:** Estimate the test error associated with a ML method
  - **Model selection:** Select the appropriate level of flexibility
- **Bootstrap** is most commonly used to measure accuracy of a parameter or of a ML method.

Cross-Validation

- **Cross-validation approach:** **Test error rate** can be calculated by
  1. holding out a subset of training observations while fitting the model
  2. applying the ML method to those held out observations and calculating Test error.
- Cross-validation works similar for both regression and classification problems.
- Why cross-validation?
  - Training error rate is not reliable.
  - A designated test data set is usually not available.
- Two cross-validation approaches:
  - Leave-One-Out Cross-Validation
  - k-Fold Cross-Validation
- Validation set approach is a simple ***hold-out*** method.
  - Cross-validation is a more refined version of the validation set approach.

Validation set approach

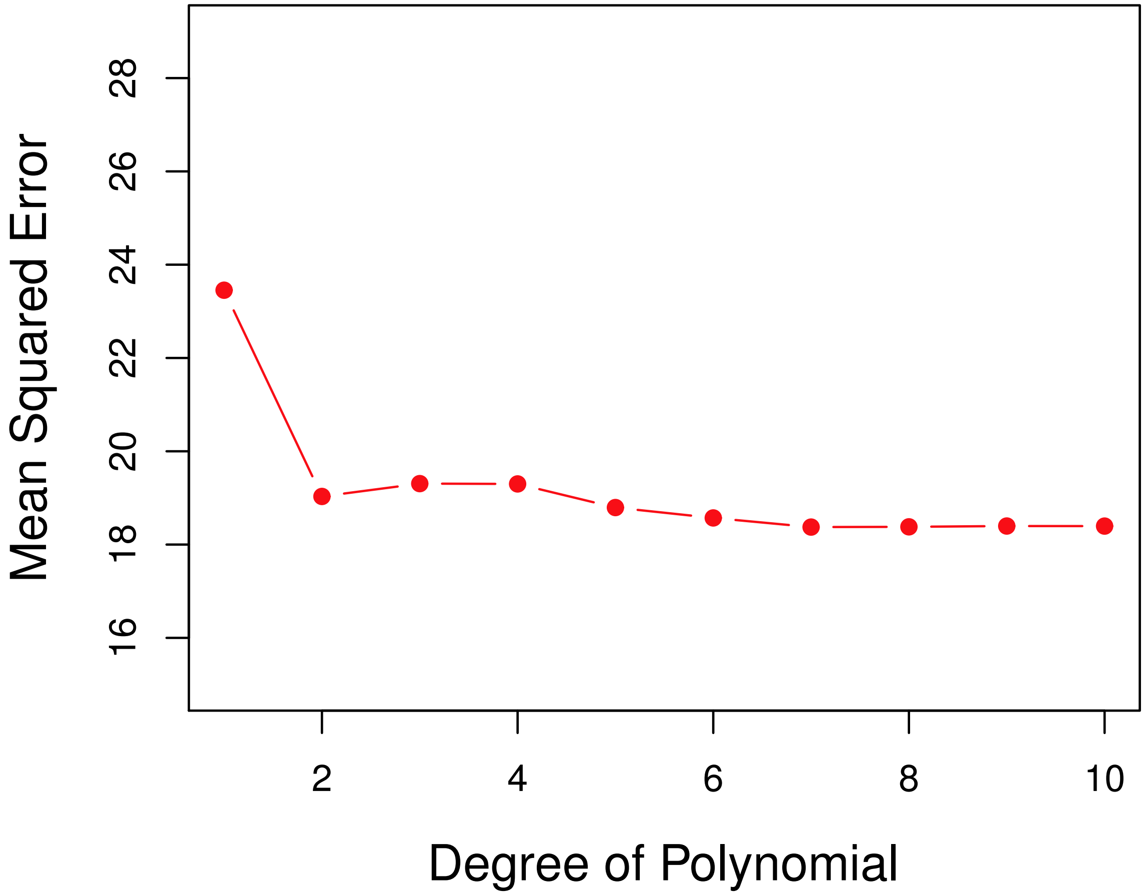
- It involves **randomly** dividing the available set of observations in two parts:
  1. **Training set:** Model is fit on this part.
  2. **Validation set:** Fitted model is used to predict the responses on this part.
- Validation set error rate (e.g. MSE in a KNN regression) provides an estimate of the test error rate.
- **Example:** A set of n observations are randomly split into a training set and a validation set:

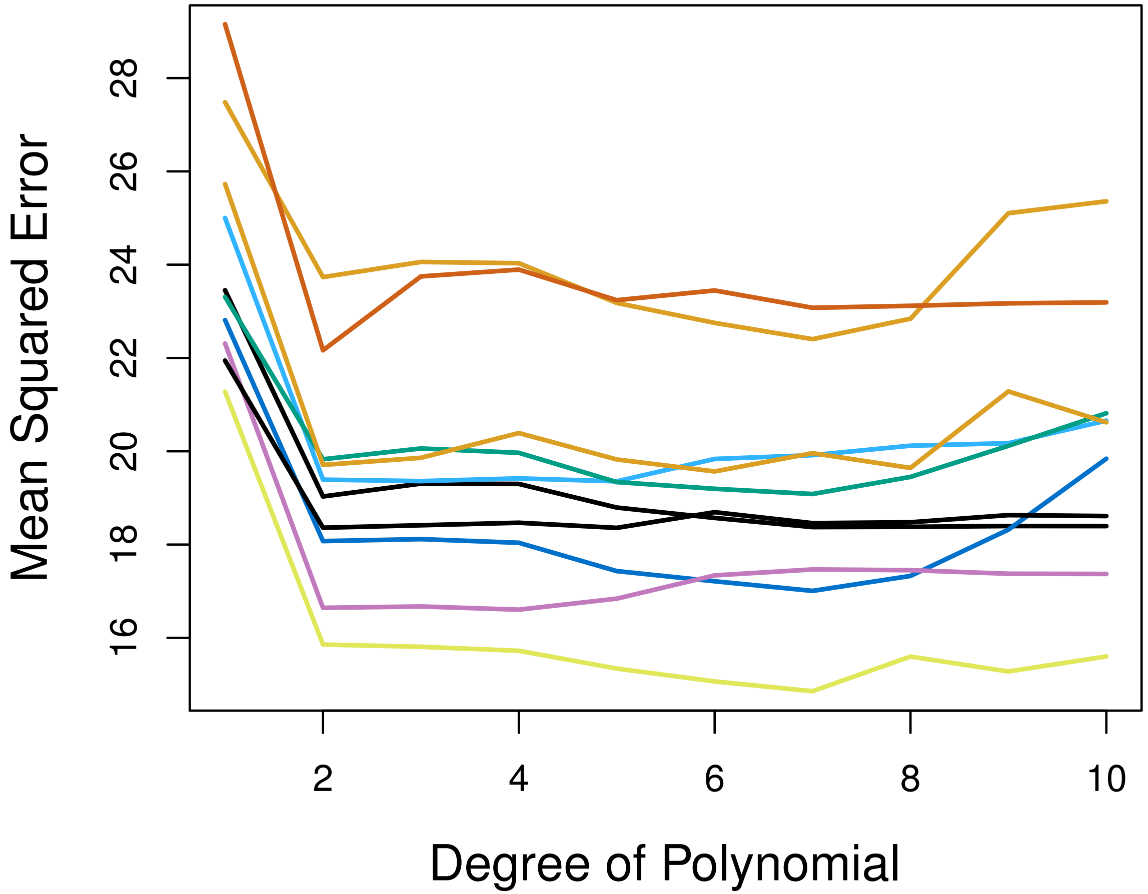


Validation set: An Example

- **Question:** How to decide the degree of non-linearity in a polynomial regression model?
- In the Auto data, we want to know the relationship between mpg and horsepower. Which degree polynomial will provide the best fit?
  - Quadratic? Cubic? or Higher order?
- We randomly split 392 obs. into two sets:
  - **Training set** with 196 observations
  - **Validation set** with 196 observations
- We fit polynomial models of varying degrees on the training set and compute their test error (MSE) using the validation set.
- The quadratic model is definitely preferred over the linear model. The cubic model has a slightly higher validation set MSE than the quadratic model.
- **Note:** If we repeat this procedure, we see a lot of variability in test MSE.

Predicting mpg using polynomials of horsepower





Validation set error for a single split

Validation set error for 10 random splits

Validation set approach

Drawbacks of validation set approach:

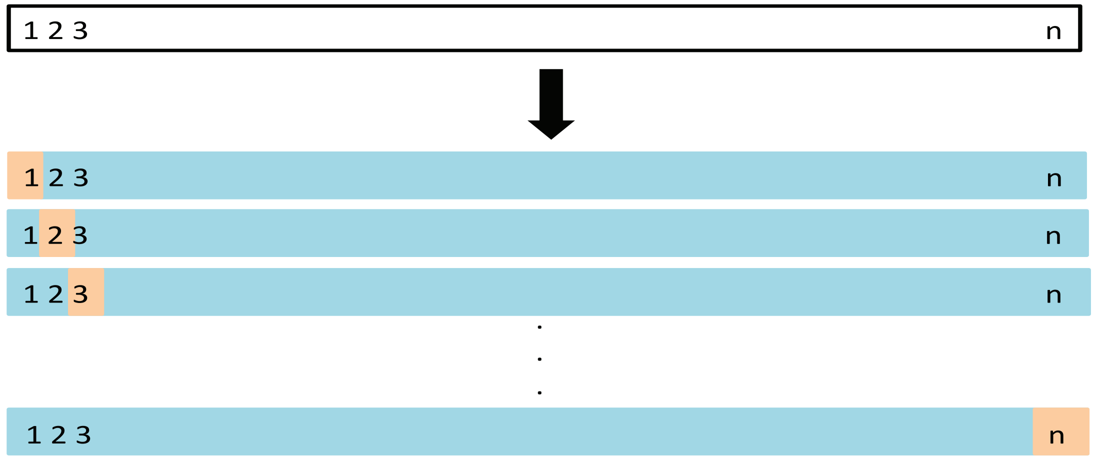
1. The validation estimate of the test error rate can be highly variable. It depends on which observations end up in the validation set.
2. Only a subset of the total observations are in the training set. ML methods perform worse when trained on fewer observations.

- **Question:** Validation set error has high variance or low variance?
- Validation set error rate may ***overestimate*** the test error rate for the model fit on the entire data.
- **Cross-validation** is a refinement of the validation set approach: Addresses both the drawbacks.

Leave-One-Out Cross-Validation (LOOCV)

- **LOOCV** is a very general method. It can be used with any ML method.
- Steps involved in **LOOCV**:
  1. A single obs. (x_1,y_1) is used for validation set and remaining observations (x_2,y_2) … (x_n,y_n) make up the training set.
  2. The ML method is fit on the (n-1) observations.
  3. The fitted model is used to make a prediction $\widehat{y_1}$ for the excluded observation (x_1,y_1).
  4. $MSE_1=(y_1-\widehat{y_1})^2$ is calculated.
  5. Steps 1-4 are repeated (n-1) times, each time excluding a single different observation.
  6. LOOCV estimate of test error:

$$\mathbf{CV_{(n)}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{MSE_i}$$



Leave-One-Out Cross-Validation (LOOCV)

- **LOOCV** has far less bias: It uses almost the whole data set.
  - Much better as compared to the validation set method.
  - Does not over-estimate the test error.
- Performing **LOOCV** multiple times yields the same result unlike the validation set approach.
- Major drawback of **LOOCV**:
  - Expensive to implement: Model has to be fit n times. Problematic if n is large!
- Special property (closed-form function) of **LOOCV** only for linear regression:

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

$\widehat{y_i}$ is i^{th} fitted value in the original regression

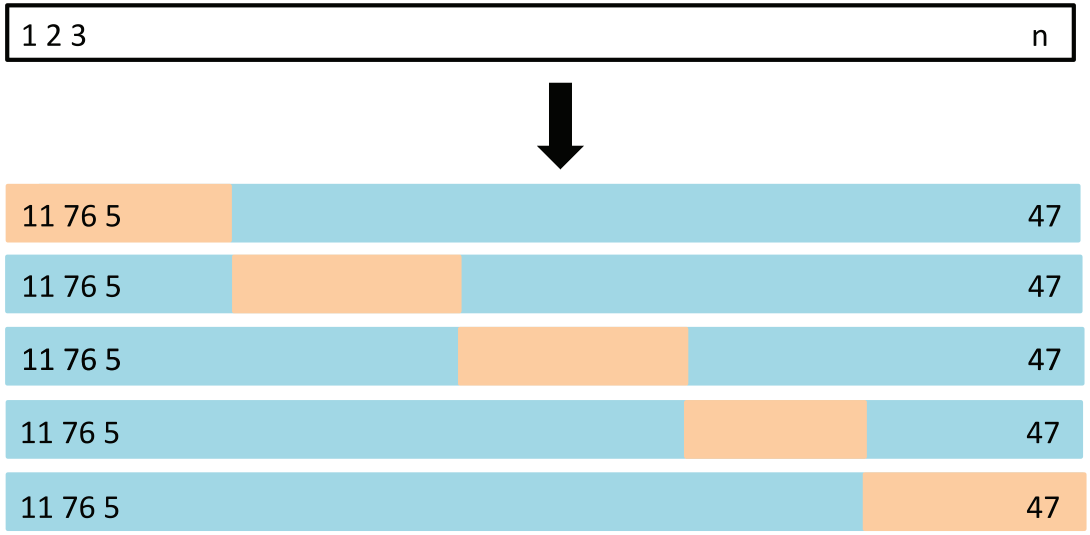
h_i is the leverage statistic (measure of the influence of single obs. on fit)

- LOOCV can be applied to any method such as LDA, KNN, Logistic model etc.

k-Fold Cross-Validation

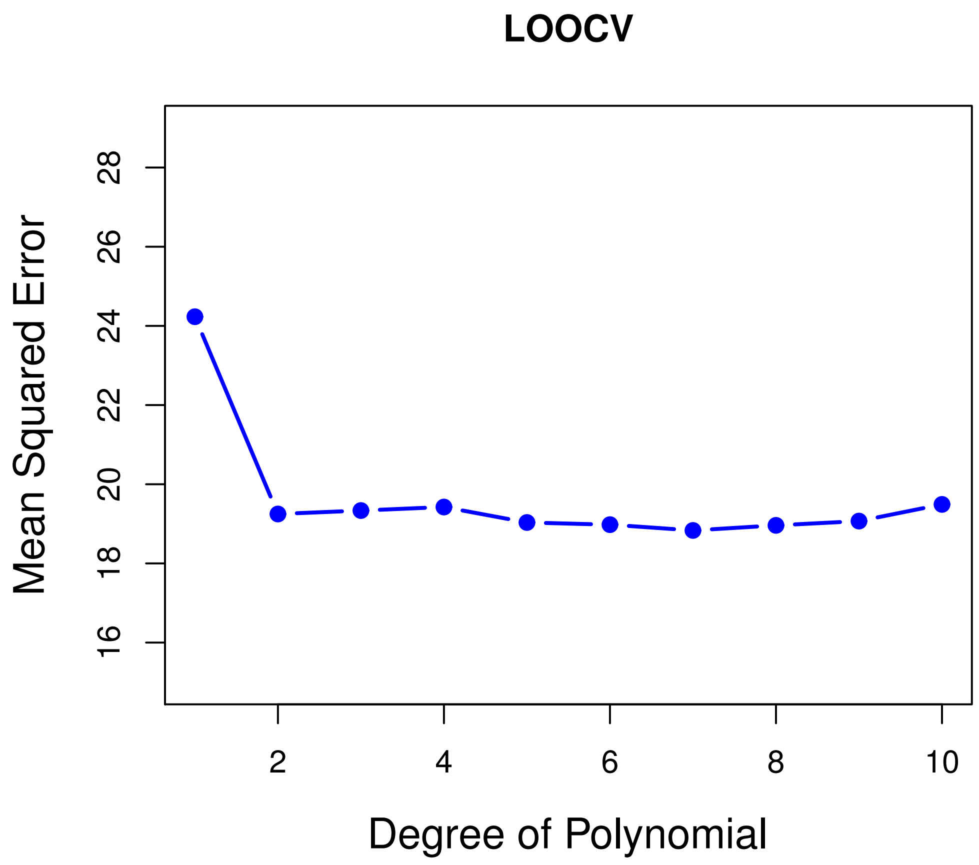
- Steps involved in **k-Fold CV**:
  1. Randomly divide a set of observations in k groups/folds of equal size (roughly).
  2. The first fold is treated as the validation set and ML method is fit on the rest (k-1) folds.
  3. MSE_1 is computed for the obs. in the held-out fold.
  4. Steps 2-3 are repeated (k-1) times, each time excluding a different set of observations.
  5. This process results k estimates of test error MSE_1, MSE_2 , …, MSE_k .
  6. k-fold CV is computed by averaging these values:

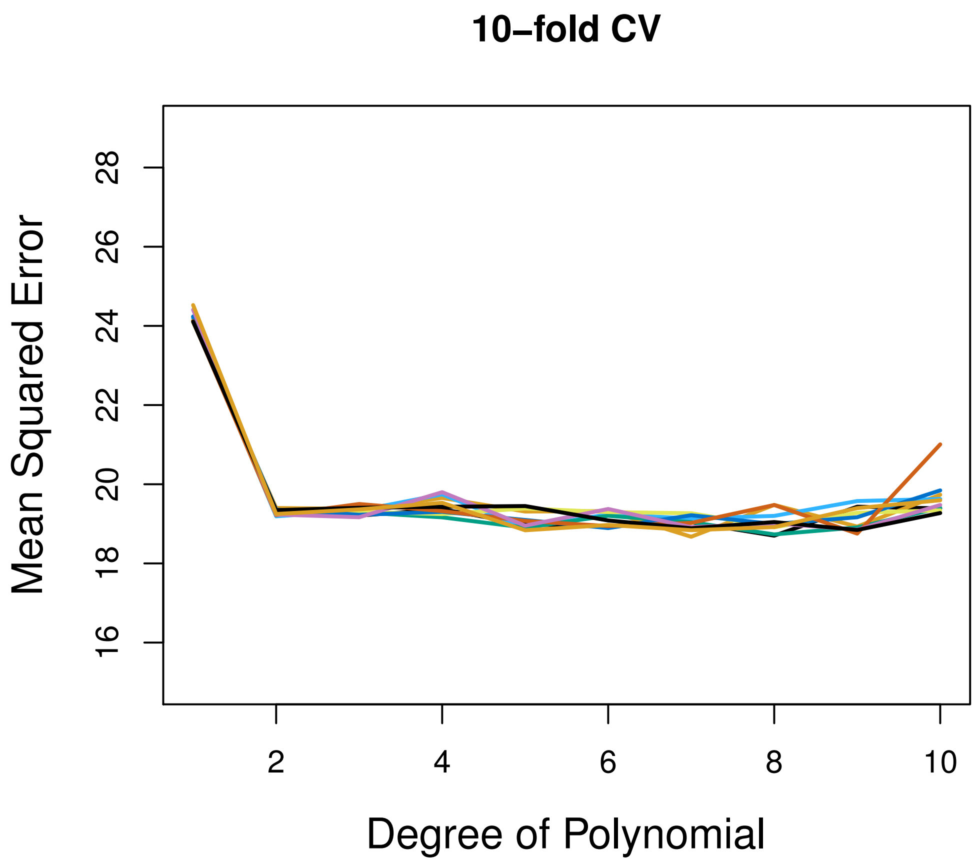
$$\mathbf{CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i}$$



Note: LOOCV is a k-fold CV with k=n.

Predicting mpg using horsepower



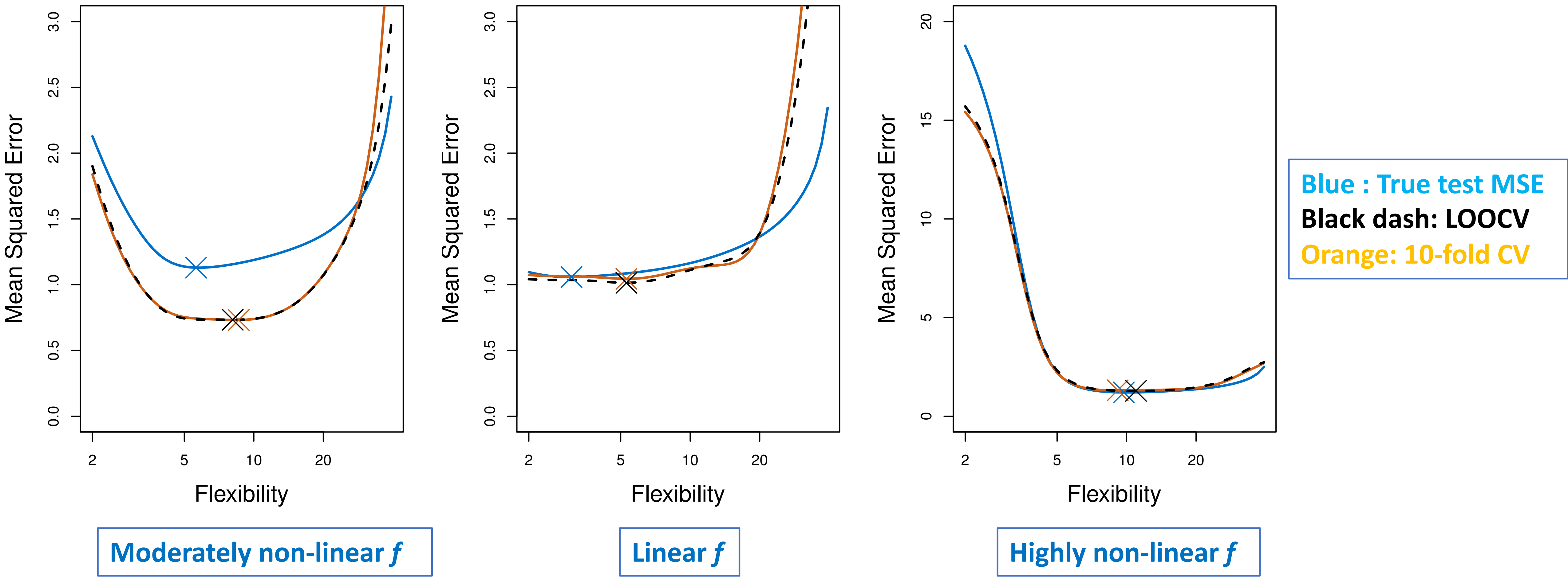


- Note that LOOCV has no variability.
- Variability in 10-fold CV << Validation set approach.
- k-Fold CV has an obvious (computational) advantage over LOOCV.

True Test MSE and Estimated Test MSE

- When we examine real data, we do not know the **True Test MSE** !
- We cannot say for sure whether the CV error is a good estimate of the True Test MSE.
- However, we can get more insight on this issue using simulated data:
  1. Estimate a fitted function $\hat{f}$ of a given flexibility using the given data D.
  2. Simulate new data D' using the true data generating function f.
  3. Compute True Test MSE on the new data D' using the fitted function $\hat{f}$ of given flexibility.
  4. Compute Estimated Test MSE (LOOCV and k-Fold CV) using the given data D.
  5. Compare True Test MSE and Estimated Test MSE.

True Test MSE and Estimated Test MSE



- CV curves are generally close to identifying the correct level of flexibility. **(Model Selection)**
- CV error may NOT correctly measure true test MSE in some cases. **(Model Assessment)**

Bias-Variance Trade-off for k-Fold Cross-Validation

- **k-Fold CV** often gives more accurate measures of test error rate than **LOOCV**.
- **Bias + Variance** is often less for a k-fold CV than the LOOCV method.
- **Bias reduction:** LOOCV $\succ$ k-Fold CV $\succ$ Validation set method
- **Variance reduction:** LOOCV $\prec$ k-Fold CV if $k < n$
  - In LOOCV, we are averaging outputs over almost identical n models. (High correlation)
  - LOOCV = mean of highly correlated quantities = High variance
  - In k-fold CV, the outputs of k fitted models are somewhat less correlated as there is less overlap between the training sets in each model.
  - k-fold CV = mean of lesser correlated quantities = Low variance

- Typically, k=5 or k=10 are used in practice. These values are shown to have neither high bias, nor high variance.

Cross-Validation on Classification Problems

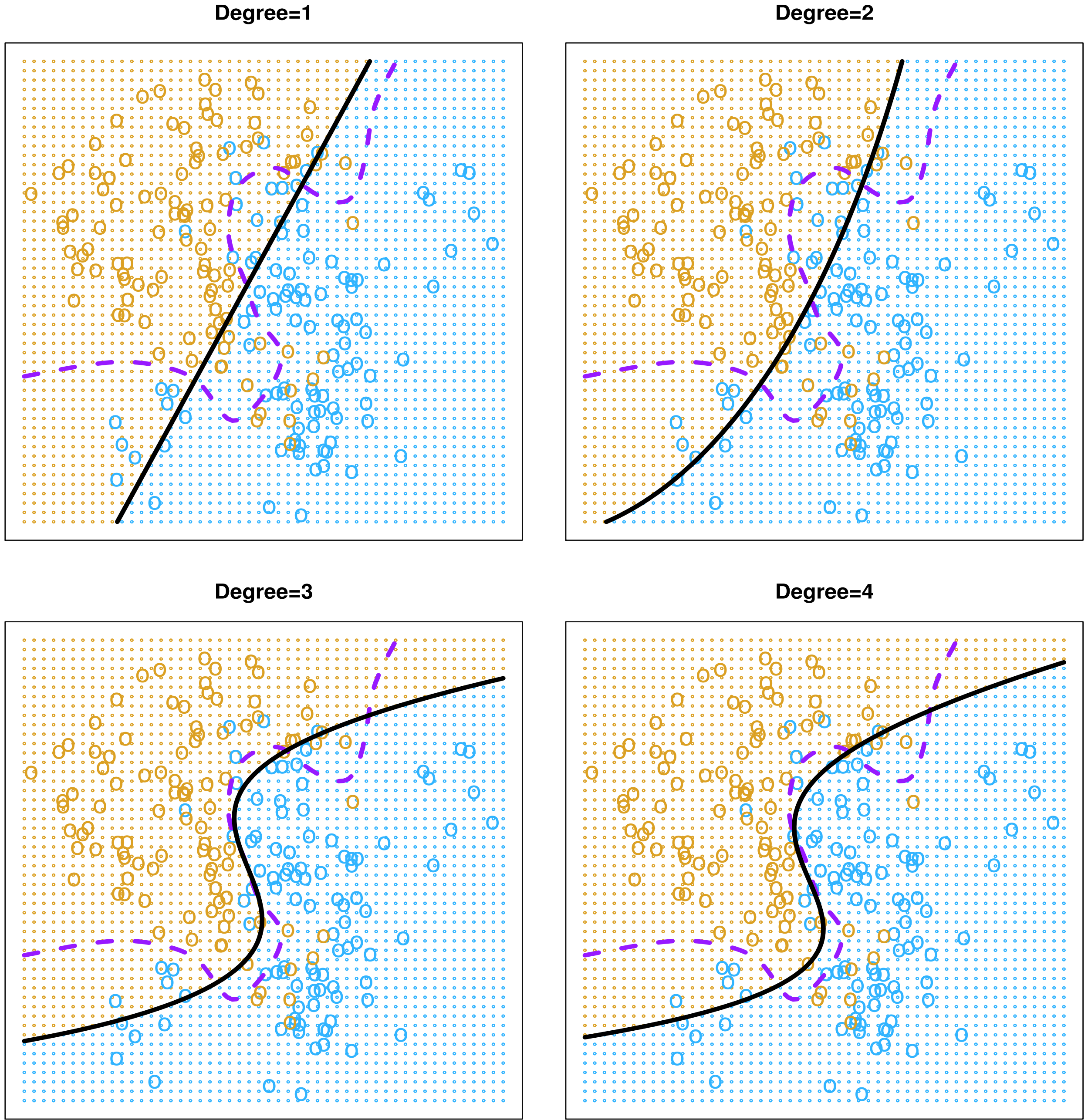
- CV works similarly in the classification setting - where Y is qualitative.
- Instead of MSE, we use the number of mis-classified observations.
- **LOOCV** error rate is measured as

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Err}_i$$

where $\mathrm{Err}_i = \mathrm{I}[y_i \neq \widehat{y_i}]$.

- k-Fold error rate and Validation Set error rates are defined similarly.
- **A classification example:**
  - Y = 2 classes and X = 2 predictors.
  - Logistic regression is used to classify Y.
  - We use polynomial functions of predictors to make the decision boundary flexible.
  - How flexible the decision boundary we want?
  - In other words, what level of flexibility will minimize the test error?

Logistic Regression (Polynomial Predictors)



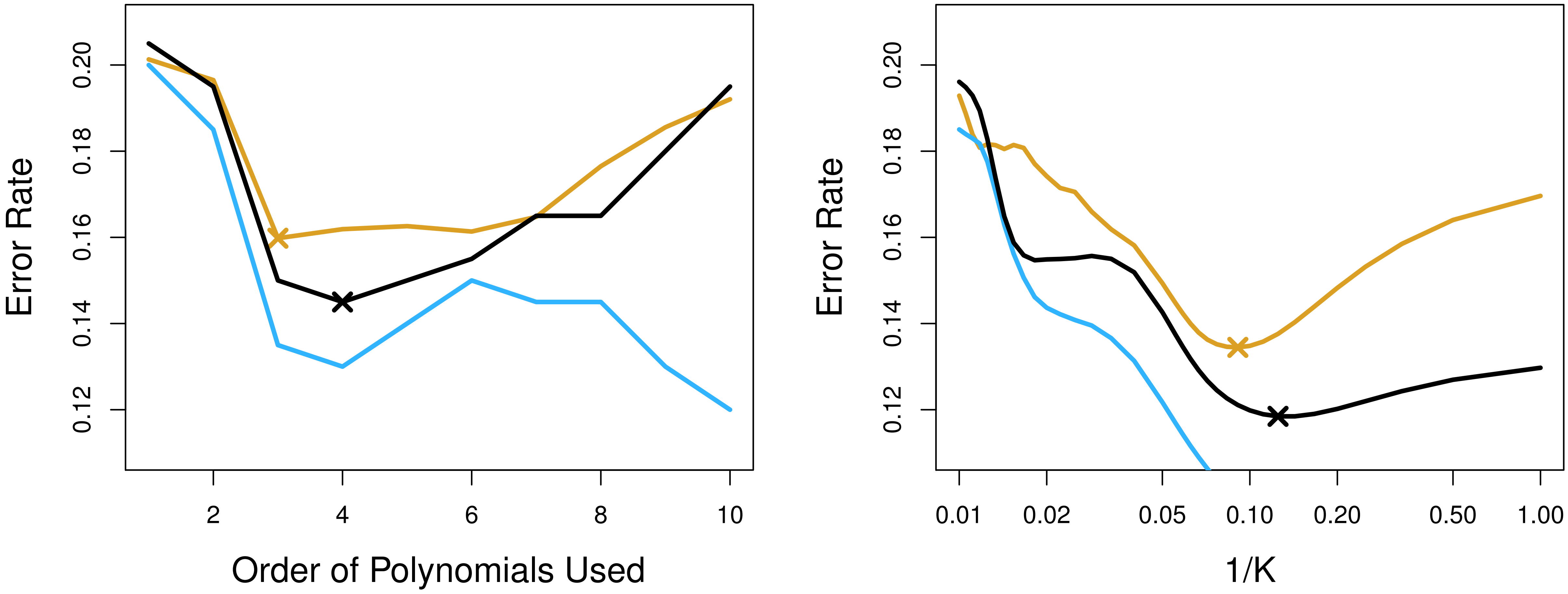
Bayes decision boundary = purple dashed line.
Logistic decision boundaries = Black line

- **Bayes error rate = 0.133**
- **True test error rate = 0.201 (Degree=1)**
- **True test error rate = 0.197 (Degree=2)**
- **True test error rate = 0.160 (Degree=3)**
- **True test error rate = 0.162 (Degree=4)**
- **A third degree polynomial seems to be optimal.**

Cross-Validation and Classification

- In real data, we neither have Bayes error rate nor do we have the luxury of measuring true test error rate.
- We turn to cross-validation for making the decision of the poly. flexibility level.
- We compute 10-fold CV errors from fitting ten logistic regression models using polynomial functions of the predictor up to tenth order.
  - Training error decreases with flexibility
  - Test error rate has the characteristic *U-shape.*
  - 10-fold CV provides a good approximation to the true test error. Bit under-estimation.
  - 10-fold CV chosen flexibility level leads to good test set performance. True test error = similar for degree 3 and 4.
- Fitting a KNN on the same data: Use cross-validation to find an optimal k.
- The value of k chosen by CV for KNN is similar to k chosen by the true test error rate.

Cross validation and Classification



True test error = Brown
Training error = Blue
10-fold CV = Black

The Bootstrap

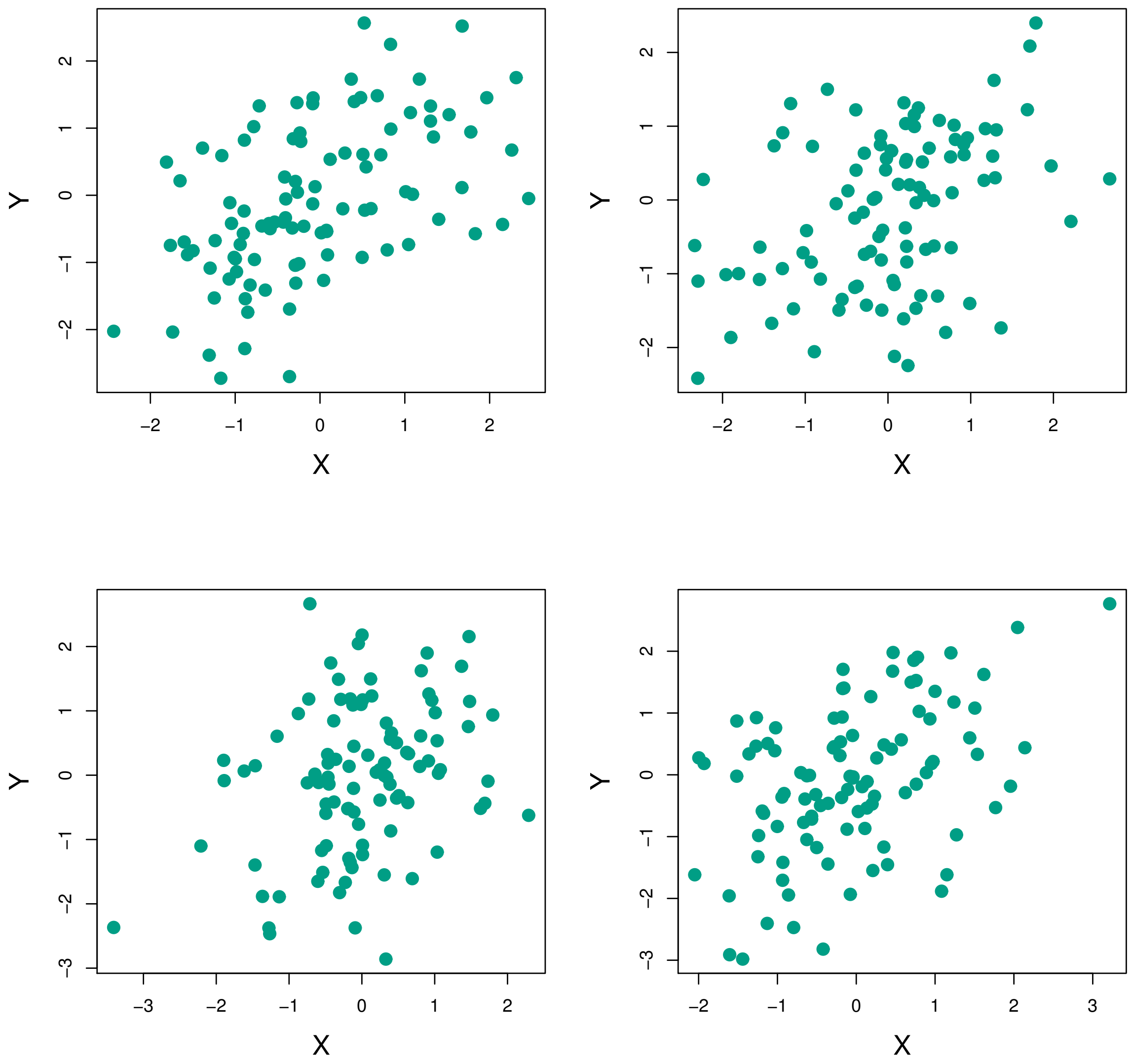
- **Bootstrap** is used to quantify the uncertainty associated with any ML method.
- The appeal of bootstrap is in its wide applicability.
- An investment risk example:
  - We wish to invest a fixed sum of money in two financial assets that yield return of **X** and **Y**, where **X** and **Y** are random quantities.
  - We will invest a fraction α in **X** and remaining **(1- α)** in **Y**.
  - The risk associated with our investment **= Var(αX + (1- α)Y)**
  - The risk is minimized at $\boldsymbol{\alpha} = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$, where σ_X^2=Var(X), σ_Y^2=Var(Y), and σ_{XY}=Cov(X,Y).
  - We can compute the estimates of σ_X^2, σ_Y^2 and σ_{XY} using past data, which gives us $\hat{\alpha}$,

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

  - It is natural to wish to quantify the accuracy of $\hat{\alpha}$. This is where bootstrap comes in.

The Bootstrap: An investment example

- We simulate 100 pairs of X and Y (4 times) with σ_X^2=1, σ_Y^2=1.25 and σ_{XY}=0.5.



- We use them to estimate σ_X^2, σ_Y^2 and σ_{XY} and eventually a measure of $\hat{\alpha}$.
- The resulting estimates of α are 0.576, 0.532, 0.657 and 0.651.
- How accurate is our measure? True α =0.6
- We drew the sample 1000 times to measure accuracy. The mean $\hat{\alpha}$ over 1000 samples=0.599
- The std. deviation of the estimates =0.083
- Roughly speaking, we expect estimated $\hat{\alpha}$ to differ from true α by 0.083, on average.
- We cannot apply this procedure for real data as we only have one sample of data.
- Bootstrap for rescue !

The Bootstrap method

- We obtain distinct data sets by repeatedly sampling observation from the original data set.
- The sampling is performed **with replacement** i.e. one observation can appear twice.
- If an observation is sampled, both its X and Y are included.
- Suppose, we obtain B distinct bootstrap data sets. Each data set can be used to measure $\hat{\alpha}$. Hence, we will have estimates $\hat{\alpha}_1, \ldots, \hat{\alpha}_B$.
- The std. error using B=1000 bootstrap samples is 0.087, which is very close to 0.083 obtained using simulated data sets.
- Hence, bootstrap can be used to effectively measure the variability associated with $\hat{\alpha}$ or risk associated with the investment.

Sampling in bootstrap

Original Data (Z)

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Z^{*1} → $\hat{\alpha}^{*1}$

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Z^{*2} → $\hat{\alpha}^{*2}$

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

⋮

Z^{*B} → $\hat{\alpha}^{*B}$

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

Fun fact: On an average, 1/3 of the observations are not used in a bootstrapped sample.

Summary

- Importance of re-sampling
- Validation Set Approach
- Cross-Validation:
  - Leave-one-out cross-validation (LOOCV)
  - k-fold cross-validation
- The Bootstrap