STATISTICAL LEARNING

CHAPTER 5: RESAMPLING METHODS

INSTRUCTOR: SEOKHO LEE

HANKUK UNIVERSITY OF FOREIGN STUDIES

2015 SPRING

- **Resampling methods**
  - They involve repeatedly drawing samples from a training set and refitting a model of interest on each sample in order to obtain additional information about the fitted model
  - Resampling approaches can be computationally expensive
  - They include **cross-validation** and **bootstrap**
- Cross validation
  - Cross-validation is used to estimate the test error associated with a given statistical learning method
  - **Model assessment**: Cross-validation can be used to evaluate its performance
  - **Model selection**: Cross-validation can be used to select the appropriate level of flexibility
- Bootstrap
  - Bootstrap provides a measure of accuracy of a parameter estimate or of a given statistical learning method

Cross-Validation

- Test error rate
  - The average error that results from using a statistical learning method to predict the response on a new observation
  - The test error can be easily calculated if a designated test set is available. Unfortunately, this is usually not the case
  - Training error, which can be easily calculated, cannot be used for an estimate for test error rate because the training error rate can dramatically underestimate the test error rate
- Simple approach to estimate the test error rate
  - We can consider a class of methods that estimate the test error rate by *holding out* a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations

The Validation Set Approach

- **Validation set approach**
  - Randomly divide the available set of observations into two parts, a **training set** and a **validation set** (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 rate (MSE for a quantitative response or misclassification rate for a qualitative response) provides an estimate of the test error rate
  - See Figure 5.1
- Auto data set for polynomial regression
  - Randomly split the 392 observations into two sets, a training set containing 196 of the data points, and a validation set containing the remaining 196 observations
  - See Figure 5.2

The Validation Set Approach

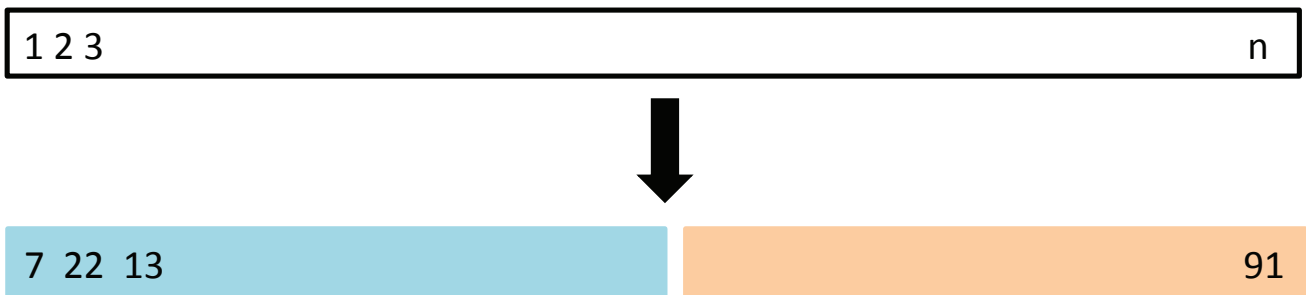

Figure 5.1: A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.

The Validation Set Approach

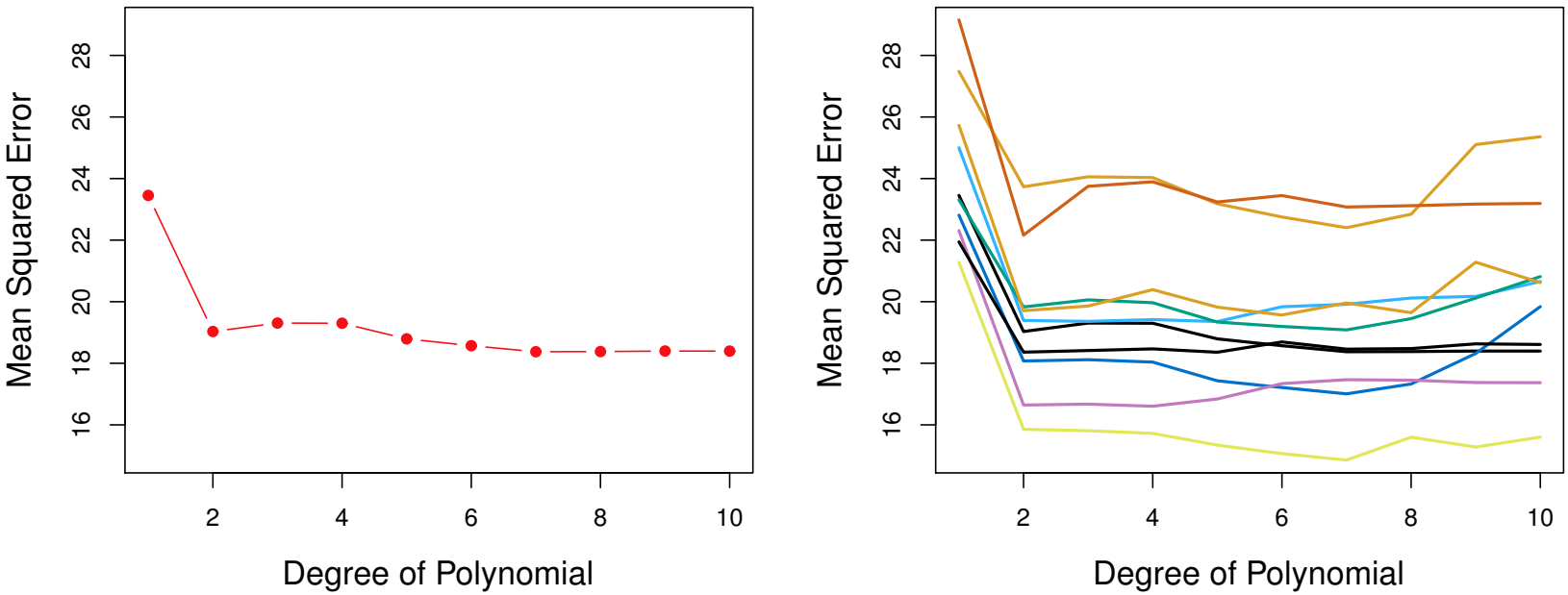

Figure 5.2: The validation set approach was used on the Auto data set in order to estimate the test error that results from predicting mpg using polynomial functions of horsepower. Left: Validation error estimates for a single split into training and validation data sets. Right: The validation method was repeated ten times, each time using a different random split of the observations into a training se hand a validation set. This illustrates the variability in the estimated test MSE that results from this approach.

The Validation Set Approach

- Two potential drawbacks for validation set approach:
  1. As in shown in the right-hand panel of Figure 5.2, the validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set
  2. In the validation approach, only a subset of the observations–those that are included in the training set rather than in the validation set–are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to *overestimate* the test error rate for the model fit on the entire data set

Leave-One-Out Cross-Validation

- **Leave-one-out cross-validation** (LOOCV)
  - LOOCV is closely related to the validation set approach but it attempts to address its drawbacks
  - Procedure (See Figure 5.3)
    1. A single observation (x_1, y_1) is used for the validation set and the remaining observations $\{(x_2, y_2), \ldots, (x_n, y_n)\}$ make up the training set. Compute $\text{MSE}_1 = (y_1 - \hat{y}_1)^2$
    2. Repeat 1 through n and compute $\text{MSE}_i = (y_i - \hat{y}_i)^2$ for $i = 1, 2, \ldots, n$
    3. Compute the LOOCV estimate for the test MSE as the average of these n test error estimates

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \text{MSE}_i \tag{5.1}$$

- Advantages of the LOOCV over the validation set approach
  - It has far less bias. It uses $n - 1$ observations to train the statistical learning method
  - There is no randomness in the training/validation set splits

Leave-One-Out Cross-Validation

- LOOCV has the potential to be expensive to implement, since the model has to be fit n times
- For least squares linear or polynomial regression, an amazing shortcut makes the cost of LOOCV the same as that of a single model fit!

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 \tag{5.2}$$

  where $\hat{y}_i$ is the ith fitted value from the original least squares fit, and h_i is the leverage defined in chapter 3

- LOOCV is a very general method, and can be used with any kind of predictive modeling including logistic regression or linear discriminant analysis. The magic formula (5.2) does not hold in general, in which case the model has to be refit n times

Leave-One-Out Cross-Validation

Figure 5.3: A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.

Leave-One-Out Cross-Validation

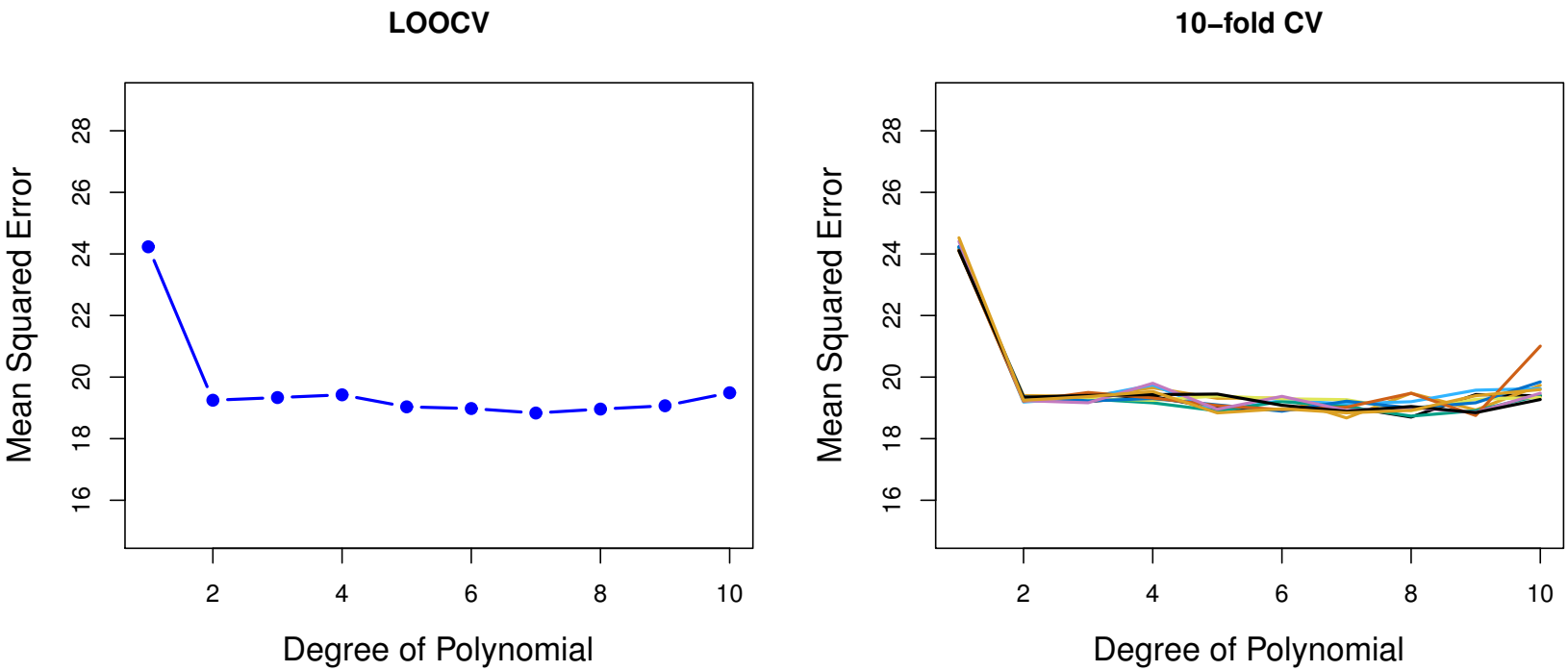

Figure 5.4: Cross-validation was used on the Auto data set in order to estimate the test error that results from predicting mpg using polynomial functions of horsepower. Left: The LOOCV error curve. Right: 10-fold CV was run nine separate times, each with a different random split of the data into ten parts. The figure shows the nine slightly different CV error curves.

k-Fold Cross-Validation

- k-**fold CV** is an alternative to LOOCV
- Procedure (See Figure 5.5)
  1. Randomly divide the set of observations into k groups, or **folds**, of approximately equal size
  2. The first fold is treated as a validation set, and the method is fit on the remaining $k-1$ folds. Compute the mean squared error, MSE_1, on the observations in the held-out fold
  3. Repeat this procedure k times to compute $\text{MSE}_1, \text{MSE}_2, \cdots, \text{MSE}_k$
  4. Compute the k-fold CV estimate for the test MSE by averaging these values

$$\text{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i \tag{5.3}$$

k-Fold Cross-Validation

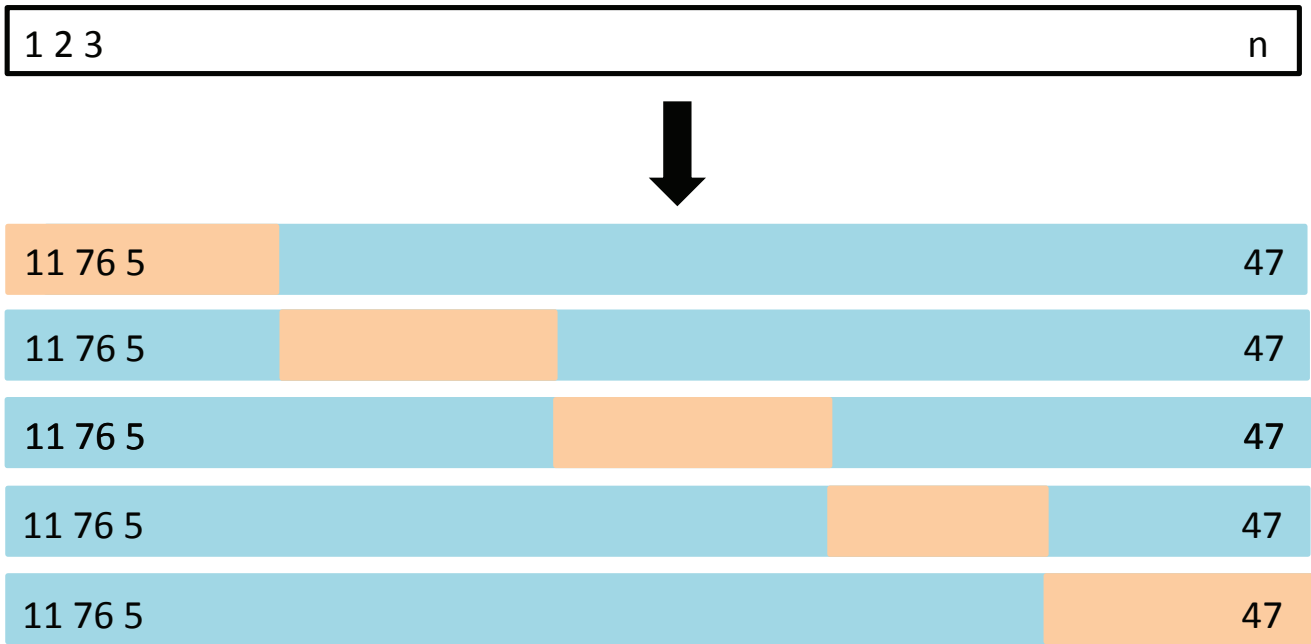

Figure 5.5: A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.

k-Fold Cross-Validation

- LOOCV is a special case of k-fold CV with $k = n$
- In practice, one typically performs k-fold CV with $k = 5$ or $k = 10$
- Advantage of k-fold CV
  - k-fold CV is less computational expensive than LOOCV
  - k-fold CV gives more accurate estimate of the test error rate or test MSE than LOOCV from a bias-variance trade-off

k-Fold Cross-Validation

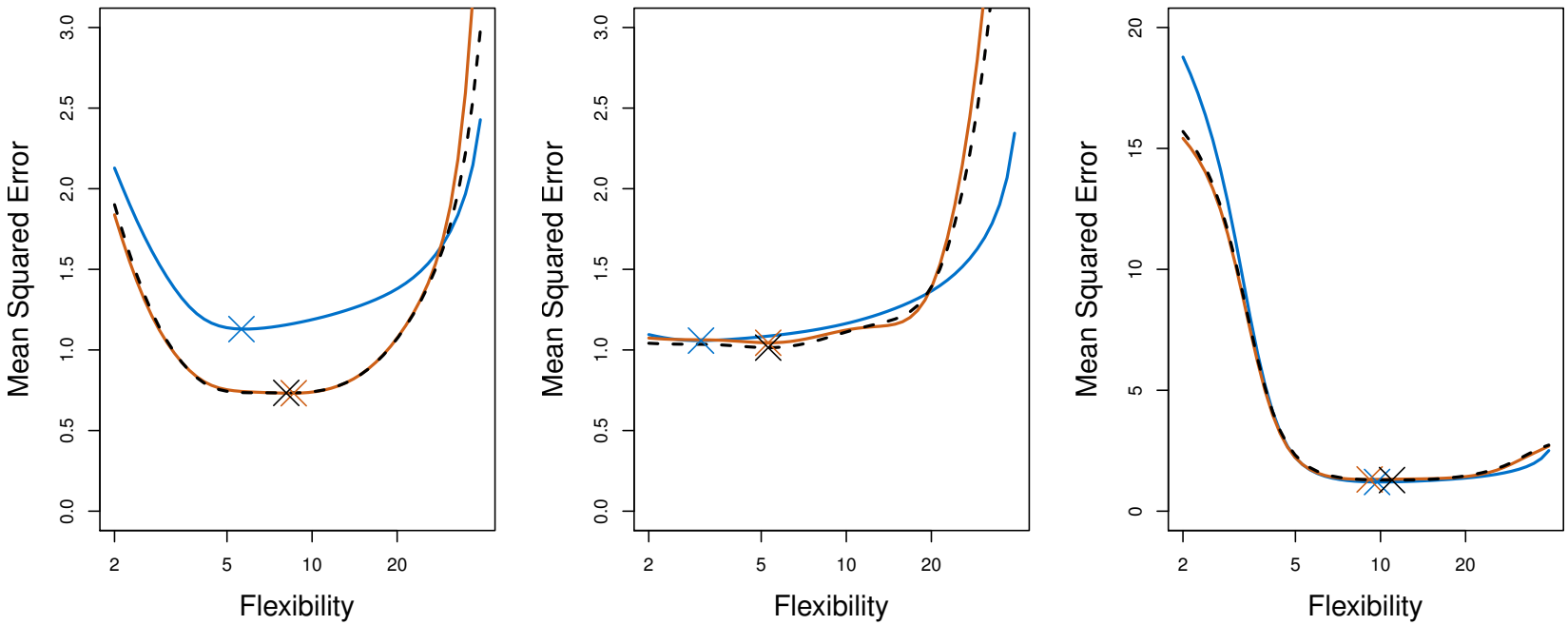

Figure 5.6: True and estimated test MSE for the simulated data set in Figure 2.9 (left), 2.10 (center), and 2.11 (right). The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the 10-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.

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

- Bias in computing test error rate
  - The validation set approach can lead to overestimates of the test error rate, since it uses only a half observations for training data
  - LOOCV will give approximately unbiased estimates of the test error rate
  - k-fold CV will lead to an intermediate level of bias
- Variance in computing test error rate
  - Bias is not the only source for concern in an estimating procedure; we must also consider the procedure's variance
  - In LOOCV, the outputs from n fitted models are highly correlated since each model is trained on an almost identical set of observations. The average of highly correlated quantities has high variance
  - In k-CV, outputs of k fitted models are somewhat less correlated with each other. Therefore the test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV

Cross-Validation on Classification Problems

- For a quantitative response, we use MSE to quantify test error
- For a qualitative response, we instead use the number of misclassified observations. For instance, the LOOCV error rate takes the form

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i \tag{5.4}$$

  where $\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$

- The k-fold CV error rate and validation set error rates are defined analogously

Cross-Validation on Classification Problems

- For the simulated data displayed in Figure 2.13
  - Bayes error rate : 0.133
  - Degree 1: linear logistic regression (error rate = 0.201)
  - Degree 2: quadratic logistic regression (error rate = 0.197)

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 \qquad (5.5)$$

  - Degree 3: cubic logistic regression (error rate = 0.160)
  - Degree 4: quartic logistic regression (error rate = 0.162)

- The test error usually displays a characteristic of U-shape
  - The optimal model corresponds to the model which achieves the minimum of the test error

Cross-Validation on Classification Problems

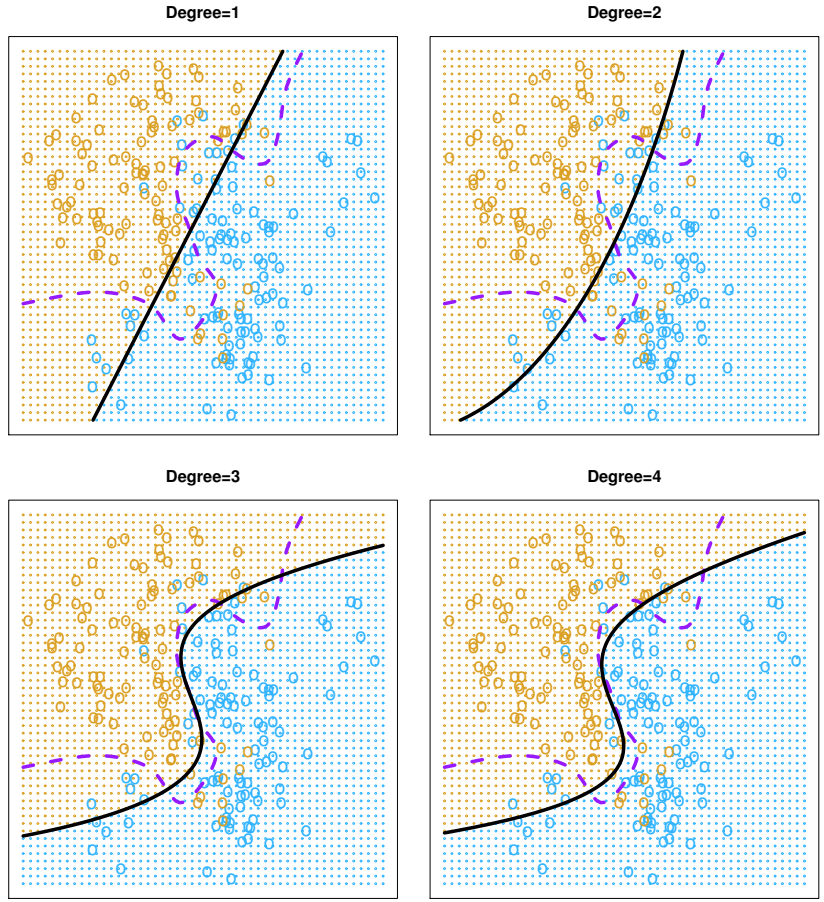

Figure 5.7: Logistic regression fits on the two-dimensional classification data displayed in Figure 2.13. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1–4) logistic regressions are displayed in black. The test error rates for the four logistic regression fits are respectively 0.201, 0.197, 0.160, and 0.162, while the Bayes error rate is 0.133.

Cross-Validation on Classification Problems

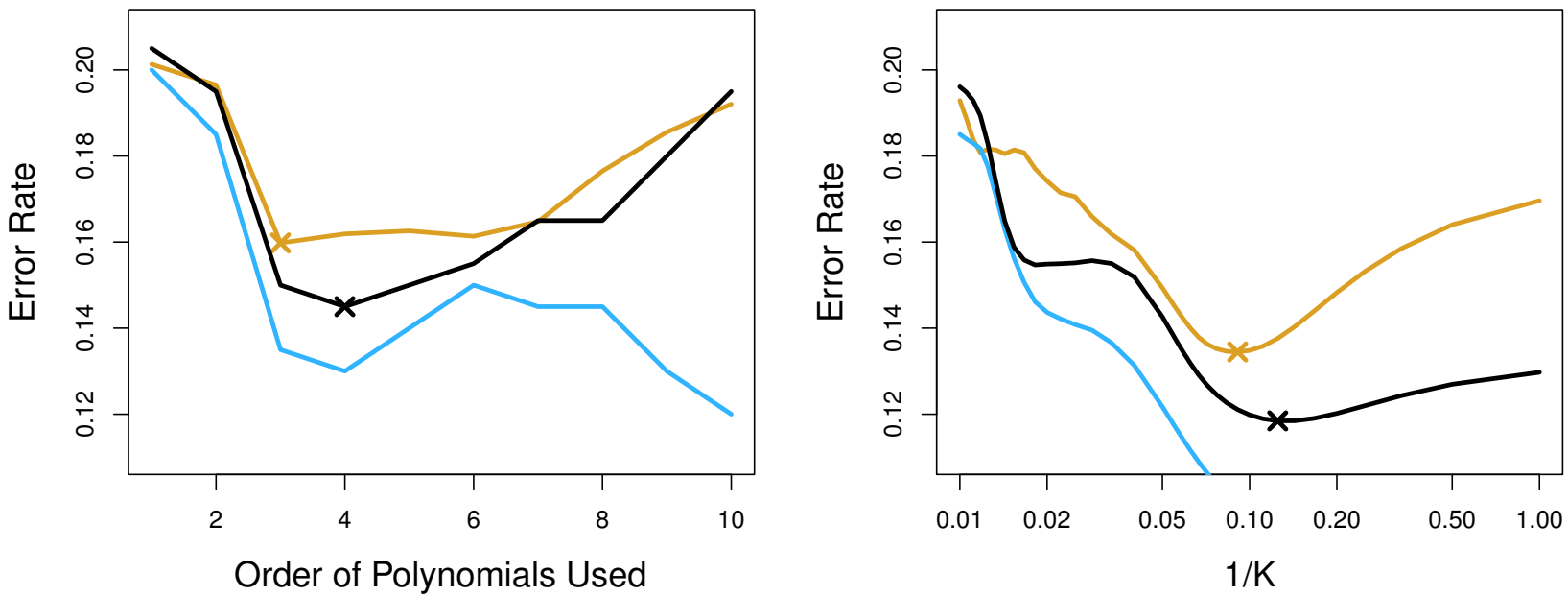

Figure 5.8: Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7. Left: Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis. Right: The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.

The Bootstrap

- The **bootstrap** is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical machine learning method
  - The bootstrap can be used to estimate the standard errors of the coefficients from a linear regression fit
  - The power of the bootstrap lies in the fact that it can be easily applied to a wide range of statistical learning methods, including some for which a measure of variability is otherwise difficult to obtain and is not automatically output by statistical software

The Bootstrap

- Best investment allocation example
  - Suppose we wish to invest a fixed sum of money in two financial assets that yields 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
  - We wish to choose α to minimize the total risk, or variance, of our investment, $\text{Var}(\alpha X + (1-\alpha)Y)$
  - One can show that the value that minimizes the risk is given by

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \tag{5.6}$$

  - and the estimator is

$$\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}} \tag{5.7}$$

  - From simulations, the value of $\hat{\alpha}$ resulting from each simulated data set ranges from 0.532 to 0.657 with $\sigma_X^2 = 1$, $\sigma_Y^2 = 1.25$ and $\sigma_{XY} = 0.5$ (true value is $\alpha = 0.6$)

The Bootstrap

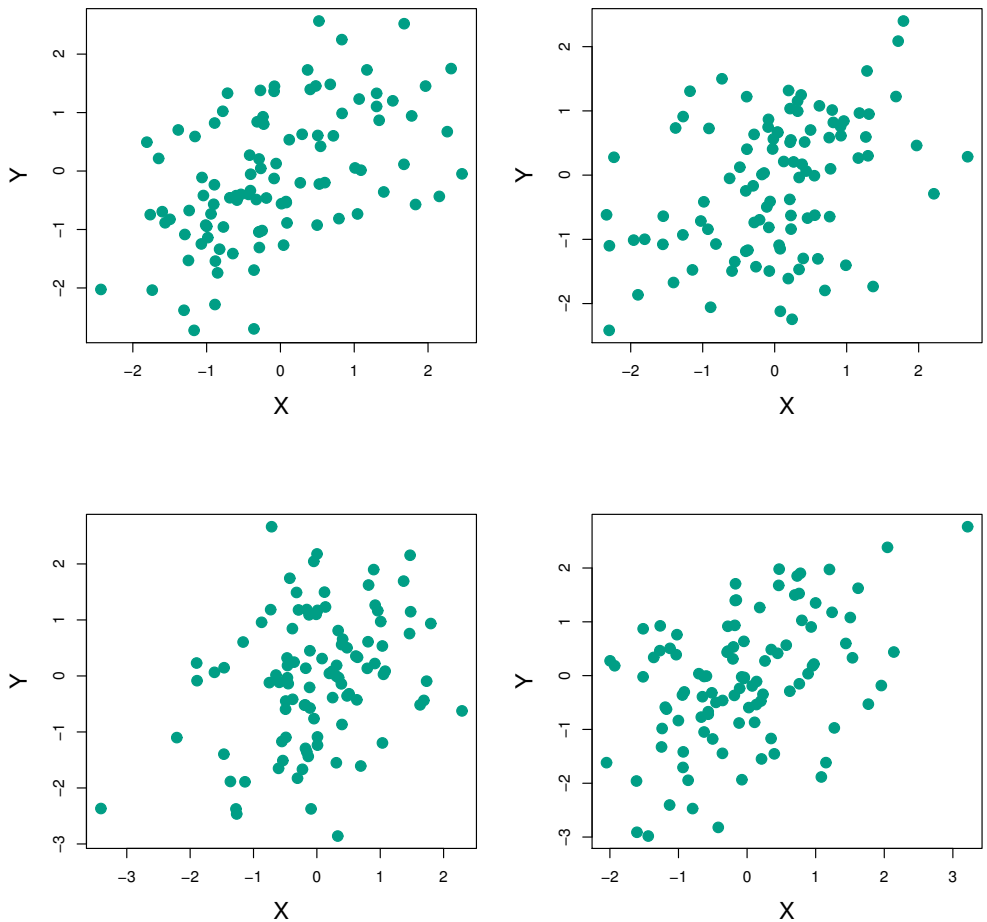

Figure 5.9: Each panel displays 100 simulated returns for investments X and Y. From left to right and top to bottom, the resulting estimates for α are 0.576, 0.532, 0.657, and 0.651.

The Bootstrap

- Best investment allocation example
  - We wish to quantify the accuracy of our estimate of α, or estimate the standard deviation of $\hat{\alpha}$
  - To do so, we repeated the process of simulating 100 paired observations of X and Y, and estimating α using (5.7), 1,000 times
  - We thereby obtained 1,000 estimates of α, which we call $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1,000}$
  - The mean over all 1,000 estimates for α is

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

  - The standard deviation of the estimates is

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

  - This gives a very good idea of the accuracy of $\hat{\alpha}$: $\mathrm{SE}(\hat{\alpha}) \approx 0.083$
  - How to get this in practice?

The Bootstrap

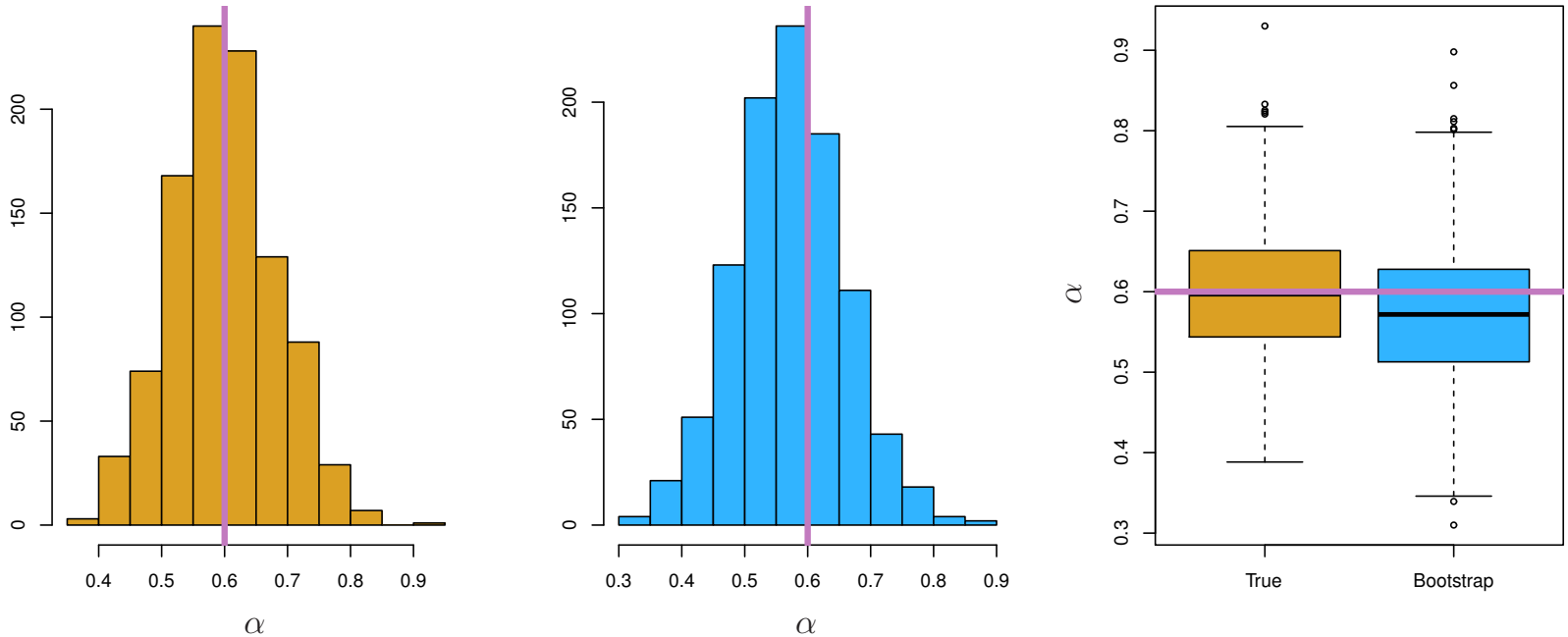

Figure 5.10: Left: A histogram of the estimates of α obtained by generating 1,000 simulated data sets from the true population. Center: A histogram of the estimates of α obtained from 1,000 bootstrap samples from a single data set. Right: The estimates of α displayed in the left and center panels are shown as box plots. In each panel, the pink line indicates the true value of α.

The Bootstrap

- In practice, we cannot use the above procedure because for real data we cannot generate new samples from the original population
- Idea of bootstrap: Rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations *from the original data set*.
- Figure 5.11 for a data set with $n = 3$ observations
  - We randomly select n observation with **replacement** from the data set and produce a bootstrap data set Z^{*1}
  - From Z^{*1}, we compute a new bootstrap estimate for α, which we call $\hat{\alpha}^{*1}$
  - This procedure is repeated B times for some large value of B, and produce B different bootstrap data sets, $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$, and B corresponding α estimates, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$
  - Now we obtain an estimate of the standard error of $\hat{\alpha}$ as the standard error of these bootstrap estimates using the formula

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2} \tag{5.8}$$

The Bootstrap

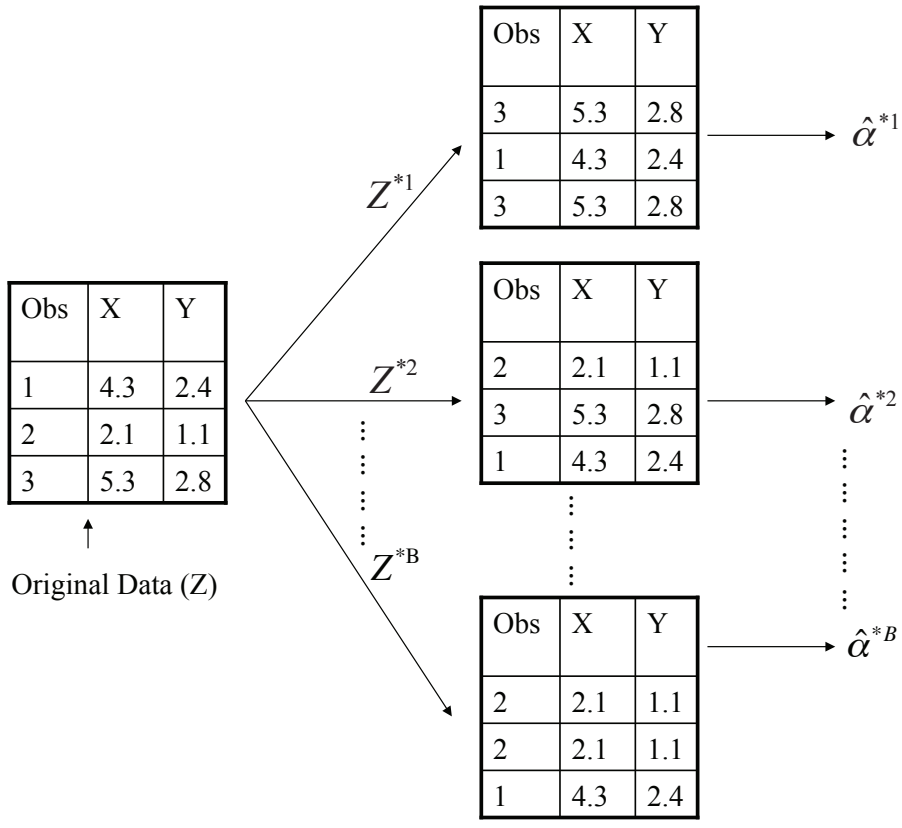

Figure 5.11: A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.

Lab: Cross-Validation and the Bootstrap