5 Resampling Methods

- Goal

  1. To choose the model with the most suitable flexibility level (model selection).
  2. The uncertainty (Standard error) in the parameter estimate can be measured by computing the sampling distribution of the estimator.

5.1 Cross-Validation

5.1.1 The Validation Set Approach

- **Implementation**:

  1. Randomly divide observations in half, so $n/2$ as validation set and the other half as training set
  2. Compute the MSE for the validation set
  3. Repeat the steps stated above so we can get the results as Figure 1

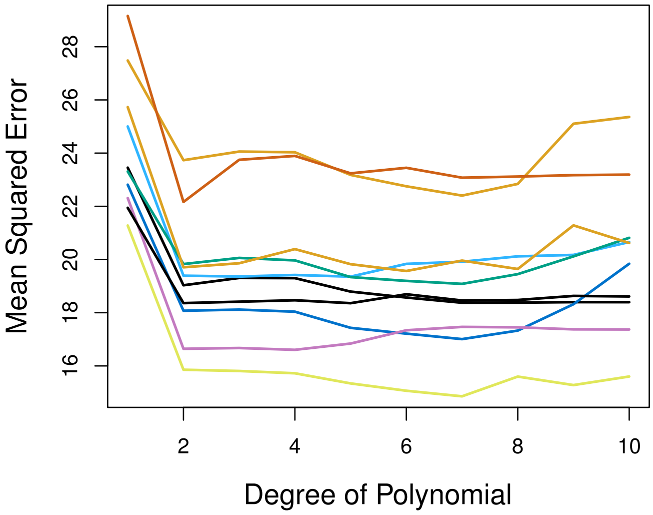

Figure 1

- Advantages

  1. Before, we only choose part of the data as training set. Now, we should train our models on different set of data to overcome overfitting.
  2. From Figure 1 we observed 10 curves indicate that the model with a quadratic term has a smaller validation set MSE then the model with only a linear term. This helps us to choose the degree of polynomial when training a model.

- Drawbacks

  1. The validation estimate of the test error can be highly variable, depending on which observations are included in the training set and in the validation set.
  2. In the validation approach only half of the data are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations. This implies the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

5.1.2 Leave-One-Out Cross-Validation (LOOCV)

Implementation:

Validation Set	Training Set	MSE
(x_1, y_1)	$\{(x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\}$	$\text{MSE}_1 = (y_1 - \hat{y}_1)^2$
(x_2, y_2)	$\{(x_1, y_1), (x_3, y_3), \ldots, (x_n, y_n)\}$	$\text{MSE}_2 = (y_2 - \hat{y}_2)^2$
(x_3, y_3)	$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$	$\text{MSE}_3 = (y_3 - \hat{y}_3)^2$
$\vdots$	$\vdots$	$\vdots$
(x_n, y_n)	$\{(x_1, y_1), (x_2, y_2), \ldots, (x_{n-1}, y_{n-1})\}$	$\text{MSE}_n = (y_n - \hat{y}_n)^2$
		$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \text{MSE}_i$

- Advantages

1. LOOCV create less bias. Since we repeatedly fit the method using training sets that contain $n - 1$ observations.
2. LOOCV does not overestimate the test error rate as much as the validation set approach does.
3. LOOCV will always yield the same results compared to validation set approach.

- Drawbacks

1. Computational expensive

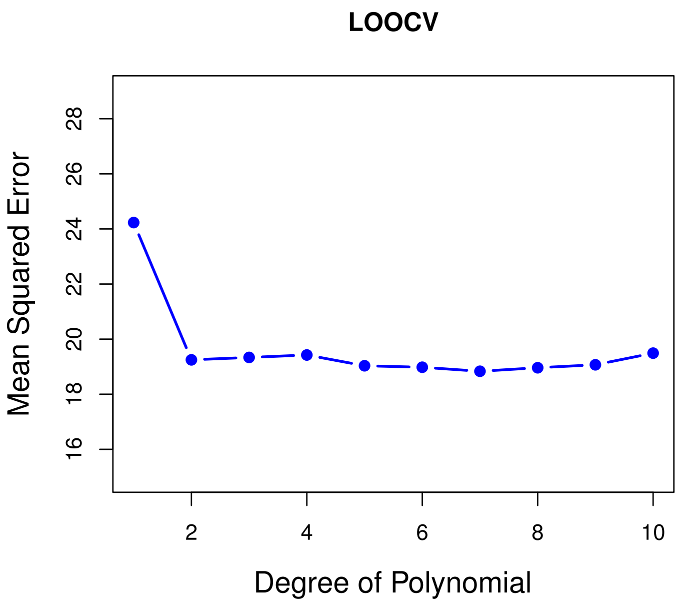

Figure 2: The LOOCV error curve

5.1.3 k-Fold Cross-Validation

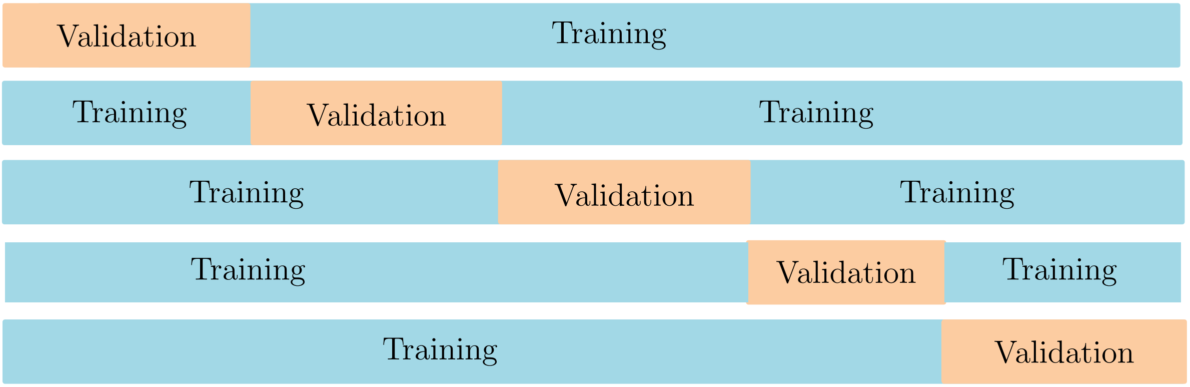

Figure 3: 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. If we set $k = n$, then we get LOOCV.

- **Implementation**:

  1. $k = (1, \ldots, n/k)$

2. Randomly divide the data into k groups, each group contains n/k observations.
3. Choose the kth fold as validation set, the other observations as training set.
4. Compute the MSE for the kth fold.
5. Repeat the process for k times.

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

When we perform cross-validation, our goal might be to determine how well a method can be expected to perform on independent data; in this case, the actual estimate of the test MSE is of interest. But at other times we are interested in finding the right flexibility level that fit the real data. That means, we are looking for the flexibility level location of the minimum point in the estimated test MSE curve.

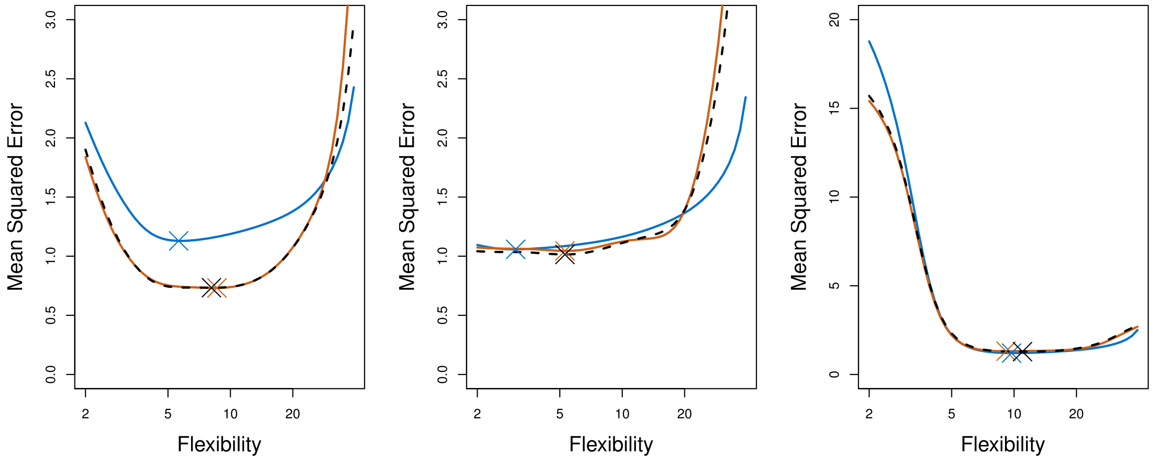

Figure 4: All of the curves come close to identifying the correct levels of flexibility (the flexibility level corresponding to the smallest MSE)

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

- **LOOCV**:

  - Bias: approximately unbiased estimates of the test error, since each training set contains $n-1$ observations.
  - Variance: higher variance, the MSE is averaging the outputs of n fitted models and each iteration we are training on an almost identical set of observations; therefore, these outputs are highly correlated with each other.

- **k-Fold CV**:
  - Bias: intermediate level of bias, since each training set contains $n-\frac{n}{k}$ observations
  - Variance: lower variance, the MSE is the averaging the outputs of k fitted model (less fitted model than LOOCV) that are less correlated with each other, since the overlap between the training sets in each model is smaller.

5.1.5 Cross-Validation on Classification Problems

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

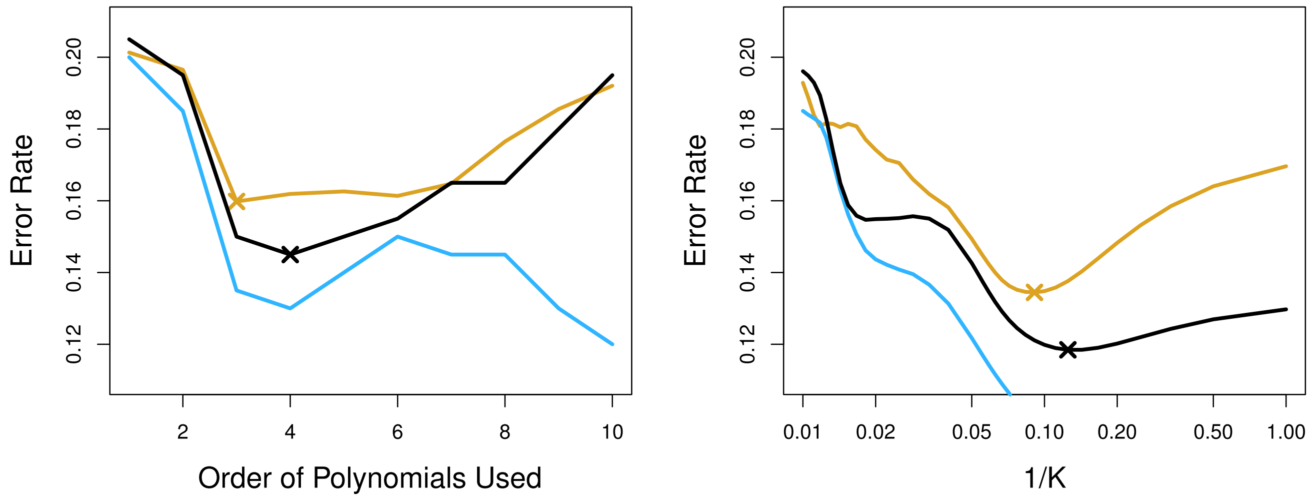

Figure 5: Test error(brown), training error(blue) and 10-fold CV error(black) Left: Logistic Regression using polynomial functions of the predictore. Right: The KNN classifier with different values of K

From Figure 5 we find that the test error displays a characteristic of U-shape. The 10-fold CV error rate provides a good approximation to the test error rate. The 10-fold CV reaches a minimum when fourth-order polynomials are used, which is very close to the minimum of the test error curve, which occurs when third-order polynomials are used.

5.2 The Bootstrap

Bootstrap is a simple Monte Carlo technique to approximate the sampling distribution. Through the sampling distribution, we can evaluate how good our data fit to the model.

Suppose we have a model fit to a set of training data. We denote the training set by $\boldsymbol{Z} = (z_1, z_1, \ldots, z_N)$ where $z_i = (x_i, y_i)$.

- **Implementation**:

1. Randomly draw data sets with replacement from the training data. Each sets $\boldsymbol{Z}^{*b}$ contains N observations.
2. By plugging the data sets $\boldsymbol{Z}^{*b}$ to the estimator $S(\cdot)$ we defined, we get our estimates $S(\boldsymbol{Z}^{*b})$.
3. We can estimate any aspect of the distribution of $S(\boldsymbol{Z})$, say, its variance, $\widehat{\mathrm{Var}}[S(\boldsymbol{Z})]$. where $\bar{S}^* = \sum_b S(\boldsymbol{Z}^{*b})/B$

One approach to estimate prediction error is to fit the model on a set of bootstrap samples, and then keep track of how well it predicts the original training set. If $\hat{f}^{*b}(x_i)$ is the predicted value at x_i, from the model fitted to the bth bootstrap dataset, our estimate is

$$\widehat{\mathrm{Err}_{\mathrm{boot}}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N}(y_i - \hat{f}^{*b}(x_i))^2$$

The bootstrap implementation can be illustrated as below Figure,

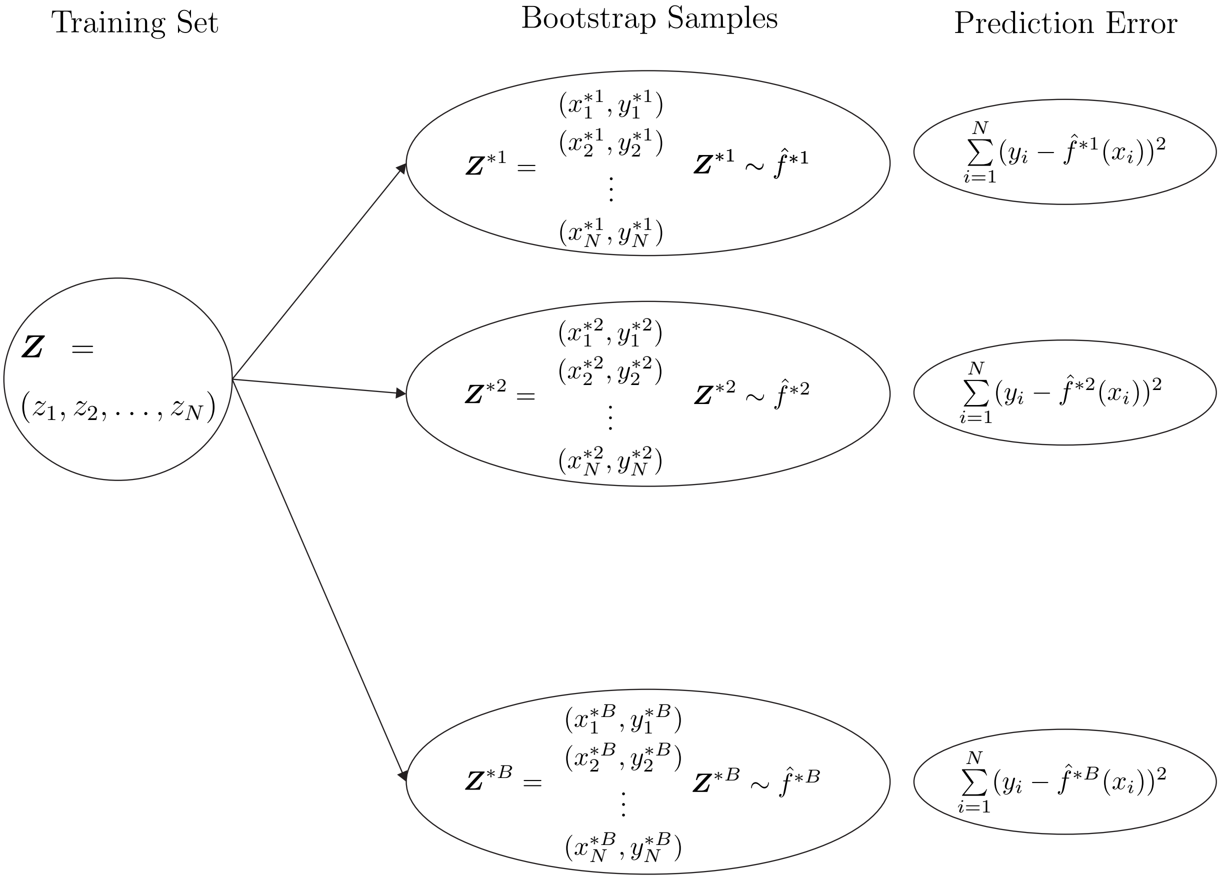
Training Set
Bootstrap Samples
Prediction Error
$\boldsymbol{Z} = (z_1, z_2, \ldots, z_N)$
$\boldsymbol{Z}^{*1} = (x_1^{*1}, y_1^{*1}), (x_2^{*1}, y_2^{*1}), \vdots, (x_N^{*1}, y_N^{*1})$
$\boldsymbol{Z}^{*1} \sim \hat{f}^{*1}$
$\sum_{i=1}^{N}(y_i - \hat{f}^{*1}(x_i))^2$
$\boldsymbol{Z}^{*2} = (x_1^{*2}, y_1^{*2}), (x_2^{*2}, y_2^{*2}), \vdots, (x_N^{*2}, y_N^{*2})$
$\boldsymbol{Z}^{*2} \sim \hat{f}^{*2}$
$\sum_{i=1}^{N}(y_i - \hat{f}^{*2}(x_i))^2$
$\boldsymbol{Z}^{*B} = (x_1^{*B}, y_1^{*B}), (x_2^{*B}, y_2^{*B}), \vdots, (x_N^{*B}, y_N^{*B})$
$\boldsymbol{Z}^{*B} \sim \hat{f}^{*B}$
$\sum_{i=1}^{N}(y_i - \hat{f}^{*B}(x_i))^2$

Figure 6

However, $\widehat{\mathrm{Err}_{\mathrm{boot}}}$ is not a good way of estimating prediction error. The reason is that the bootstrap datasets are acting as the training samples, while the original traiing set is acting as the test sample, and these two samples have observations in common. This overlap can make overfit predictions look unrealistically good.

In order to solve this issue, we apply the leave-one-out bootstrap estimate of prediction error, this is,

$$\widehat{\mathrm{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}(y_i - \hat{f}^{*b}(x_i))^2$$

- C^{-i} : The set of indices of the bootstrap samples b that do not contain observation i
- $|C^{-i}|$: the number of such bootstrap samples

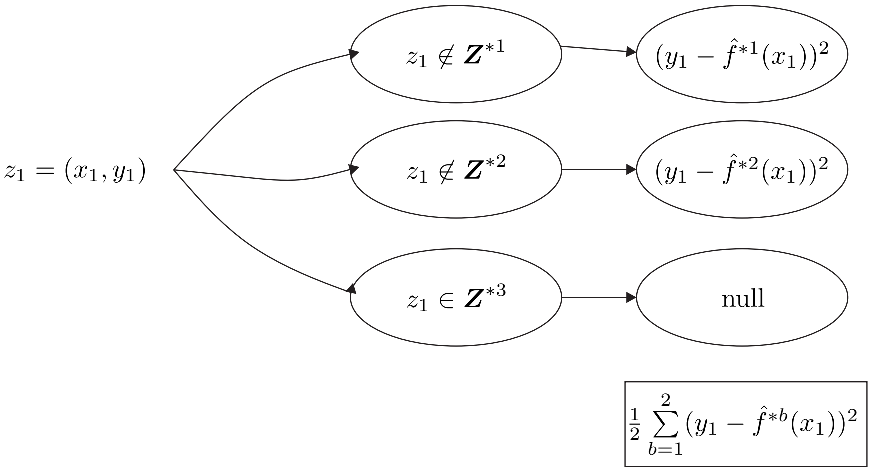
$z_1 = (x_1, y_1)$
$z_1 \notin \boldsymbol{Z}^{*1}$
$z_1 \notin \boldsymbol{Z}^{*2}$
$z_1 \in \boldsymbol{Z}^{*3}$
$(y_1 - \hat{f}^{*1}(x_1))^2$
$(y_1 - \hat{f}^{*2}(x_1))^2$
null
$\frac{1}{2}\sum_{b=1}^{2}(y_1 - \hat{f}^{*b}(x_1))^2$

Figure 7: The figure gives an example on $z_1 = (x_1, y_1)$. After repeating the process through all the observations $(z_2, \ldots, z_N)$ and summarize them, we can get $\widehat{\mathrm{Err}}^{(1)}$.

The ".632 estimator" is designed to alleviate this bias. It is defined by,

$$\widehat{Err}^{(.632)} = .368 \cdot \overline{\mathrm{err}} + .632 \cdot \widehat{\mathrm{Err}}^{(1)}$$

$\overline{\mathrm{err}}$ is the training error. The derivation of the .632 estimator is complex; intuitively it pulls the leave-one out bootstrap estimate($\widehat{\mathrm{Err}}^{(1)}$) down toward the training error rate, and hence reduces its upward bias.

5.3 Exercises

2. We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

(c) Argue that the probability that the jth observation is not in the bootstrap sample is $(1 - 1/n)^n$.

(d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

(e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

(f) When $n = 10000$, what is the probability that the jth observation is in the bootstrap sample?

(g) Create a plot that displays, for each integer value of n from 1 to 100000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

(h) We will now investigate numerically the probability that a bootstrap sample of size $n = 100$ contains the jth observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

Answer:

A : The event that the bootstrap observation is the jth observation from the original data set.

A' : The event that the bootstrap observation is not the jth observation from the original data set.

$P(A)$: The probability that the bootstrap observation is the jth observation from the original data set.

$P(A') = 1 - P(A)$: The probability that the bootstrap observation is not the jth observation from the original data set.

(a) $P(A') = 1 - \frac{1}{n}$

(b) $P(A') = 1 - \frac{1}{n}$

(c) We draw n times from the original sample with replacement, so the probability that for n repeat drawings, the jth observation is not include in the bootstrap is,

$$P(A')^n = (1 - \frac{1}{n})^n$$

(d) Probability of the jth obs. in 5 observations bootstrap sample :
$1 - P(A')^5 = 1 - (1 - \frac{1}{5})^5 = 0.67232$

(e) Probability of the jth obs. in 100 observations bootstrap sample :
$1 - P(A')^{100} = 1 - (1 - \frac{1}{100})^{100} = 0.63397$

(f) Probability of the jth obs. in 10000 observations bootstrap sample:
$1 - P(A')^{10000} = 1 - (1 - \frac{1}{10000})^{10000} = 0.6321$

(g) $P(\text{observation } i \in \text{bootstrap sample } b) = 1 - (1 - \frac{1}{n})^n = 1 - e^{-1} = 0.632$.
This implies the average number of distinct observations in each bootstrap sample is about $0.632 \cdot n$, so its bias will roughly behave like that of twofold cross-validation. More detail in The Elements of Statistical Learning Section 7.11.

Reference

張翔 (2012),《 提綱挈領學統計 》, 四版, 鼎茂圖書

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference.* Springer Science & Business Media.

J Li, Linear Discriminant Analysis, `http://sites.stat.psu.edu/~jiali/course/stat597e/notes2/lda.pdf`