

CHAPTER 1

POINT ESTIMATION ALGORITHMS

Huadong Liu

*Department of Computer Science, University of Tennessee
203 Claxton Complex, Knoxville, TN 37996-3450, USA
hliu@cs.utk.edu*

Overview

Point estimation can be used in both predictive and descriptive data mining tasks. Three classical point estimation methods — the method of moments, maximum likelihood estimation, and the Expectation-Maximization algorithm — are discussed in this chapter, followed by a review of measurements of estimation performance. This chapter intends to introduce basic concepts and methods of point estimation. These concepts and methods are the basis for more advanced estimation techniques.

Keywords: Bias, EM algorithm, maximum likelihood estimation, mean squared error, method of moments, point estimation, standard error.

1. Introduction

Statistics is the science of collecting, analyzing and presenting data. Many statistical techniques are used to perform data mining tasks. These techniques include point estimation, interval estimation, regression and many others. For a population whose distribution is known but depends on one or more unknown parameters, point estimation predicts the value of the unknown parameter and interval estimation determines the range of the unknown parameter. Point estimation techniques and algorithms will be discussed in this chapter. These classical techniques and algorithms are illustrated with examples and are not meant to reflect the state of the art

in this area. Many other useful techniques such as robust estimation methods [152] and re-sampling methods [105] have been developed and ongoing research continues to advance estimation techniques.

2. Motivation

Point estimation is a well-known and computationally tractable tool for learning the parameters of a data mining model. It can be used for many data mining tasks such as *summarization* and *time-series prediction*. Summarization is the process of extracting or deriving representative information about the data. Point estimation is used to estimate mean, variance, standard deviation, or any other statistical parameter for describing the data. In time-series prediction, point estimation is used to predict one or more values appearing later in a sequence by calculating parameters for a sample.

In this chapter, Sec. 3 discusses methods of point estimation, including the method of moments, maximum likelihood estimation, and the EM algorithm. Criteria to measure the performance of estimation methods, including bias, mean squared error, standard error, efficiency, and consistency are reviewed in Sec. 4. Finally, the summarization of the chapter is provided in Sec. 5.

3. Methods of Point Estimation

Several methods exist for obtaining point estimates, including least squares, the method of moments, maximum likelihood estimation, Bayes estimators, and robust estimation. The method of moments and maximum likelihood estimation for deriving estimates for parameters will be discussed in this section with simple examples. The EM algorithm for finding maximum-likelihood estimates will also be described.

A few formal definitions are needed before discussing methods for point estimation. Let X_1, X_2, \dots, X_n be a random sample, and let $\Theta = \{\theta_1, \dots, \theta_k\}$ be the set of population parameters. An *estimator* is a function that maps a random sample X_1, \dots, X_n to a set of parameter values $\hat{\Theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_k\}$, where $\hat{\theta}_j$ is the *estimate* of parameter θ_j .

3.1. The Method of Moments

The *method of moments*, introduced by Karl Pearson circa 1894, is one of the oldest methods of determining estimates [99]. In [149], the method of

moments was defined as follows: let X_1, X_2, \dots, X_n be a random sample from a population whose density function depends on a set of unknown parameters $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$. Assume that the first k population moments exist as functions $\phi_r(\Theta)$ of the unknown parameters, where $r = 1, 2, \dots, k$. Let

$$\hat{\phi}_r = \frac{1}{n} \sum_{i=1}^n X_i^r \quad (1)$$

be the r th sample moment. By equating $\hat{\phi}_r$ to ϕ_r , where $r = 1, \dots, k$, k equations in k unknown parameters can be obtained.

Therefore, if there are k population parameters to be estimated, the method of moments consists of the following two steps:

- (i) Express the first k population moments in terms of the k population parameters $\theta_1, \theta_2, \dots, \theta_k$;
- (ii) Equate the population moments obtained from step (i) to the corresponding sample moments calculated using Eq. (1) and solve $\theta_1, \theta_2, \dots, \theta_k$ as the estimates of parameters.

Example 1: This is an example adapted from [105]. Suppose one wanted to find estimates for parameters of the gamma distribution using the method of moments. The gamma probability density function

$$f(x; \lambda, t) = \frac{\lambda e^{-\lambda x} (\lambda x)^{t-1}}{\int_0^\infty e^{-y} y^{t-1} dy}, \quad x \geq 0,$$

has two parameters, the shape parameter t and the scale parameter λ .

Since two parameters are unknown, the first step is to express $E(X)$ and $E(X^2)$ in terms of t and λ . Though the probability density function of the gamma distribution looks complicated, the mean and variance of a gamma random variable are quite simple. The mean and variance are

$$E[X] = \frac{t}{\lambda}, \quad (2)$$

and

$$V(X) = E[X^2] - (E[X])^2 = \frac{t}{\lambda^2}, \quad (3)$$

respectively.

The next step is to solve the above two equations for t and λ in terms of $E(X)$ and $E(X^2)$. Substituting t in Eq. (3) with $\lambda E[X]$, which can be

derived from Eq. (2), yields

$$E[X^2] - (E[X])^2 = \frac{\lambda E[X]}{\lambda^2}. \quad (4)$$

Rearranging Eq. (4) gives the following expression for λ

$$\lambda = \frac{E[X]}{E[X^2] - (E[X])^2}. \quad (5)$$

By substituting Eq. (5) for λ in Eq. (2), the parameter t is obtained in terms of $E(X)$ and $E(X^2)$:

$$t = \frac{(E[X])^2}{E[X^2] - (E[X])^2}. \quad (6)$$

To get the estimates for λ and t , just substitute $E[X]$ and $E[X^2]$ with sample moments in Eqs. (5) and (6). This yields

$$\hat{t} = \frac{\bar{X}^2}{\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2},$$

and

$$\hat{\lambda} = \frac{\bar{X}}{\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2}.$$

3.2. Maximum Likelihood Estimation

Sir Ronald A. Fisher circa 1920 introduced the method of maximization of likelihood functions [82]. Given a random sample X_1, \dots, X_n distributed with the density (mass) function $f(x; \Theta)$, the *likelihood function* of the random sample is the joint probability density function, denoted by

$$L(\Theta; X_1, \dots, X_n) = f(X_1, \dots, X_n; \Theta). \quad (7)$$

In Eq. (7), Θ is the set of unknown population parameters $\{\theta_1, \dots, \theta_k\}$. If the random sample consists of random variables that are independent and identically distributed with a common density function $f(x; \Theta)$, the likelihood function can be reduced to

$$L(\Theta; X_1, \dots, X_n) = f(X_1; \Theta) \times \dots \times f(X_n; \Theta),$$

which is the product of individual density functions evaluated at each sample point.

A *maximum likelihood estimate*, therefore, is a set of parameter values $\hat{\Theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_k\}$ that maximizes the likelihood function of the sample. A

well-known approach to find $\hat{\Theta}$ is to take the derivative of L , set it equal to zero and solve for Θ . Thus, $\hat{\Theta}$ can be obtained by solving the *likelihood equation*

$$\frac{\partial}{\partial \Theta} L(\Theta) = 0.$$

It is important to note that a solution to the likelihood equation is not necessarily a maximum; it could also be a minimum or a stationary point (in the case of $L(\Theta) = \Theta^3$, for example). One should ensure that the solution is a maximum before using it as a maximum likelihood estimate.

It is sometimes easier, especially when working with an exponential function, to solve the logarithm of the likelihood function, $\log L(\Theta)$, that is,

$$\frac{\partial}{\partial \Theta} \log L(\Theta) = 0, \quad \text{where } \log L(\Theta) = \sum_{i=1}^n \log f(X_i; \Theta).$$

Since the logarithm function is monotonically increasing, which means that if $x_1 < x_2$, $\log(x_1) < \log(x_2)$, the likelihood function $L(\Theta)$ and its logarithm $\log L(\Theta)$ are maximized by the same Θ .

Example 2: Consider a population of balls with colors {red(r), blue(b), green(g)}. Assume the color of a ball occurs with the following probabilities as a function of the parameter θ ($0 < \theta < 1$):

$$\begin{aligned} f(r; \theta) &= \theta^2, \\ f(b; \theta) &= 2\theta(1 - \theta), \\ f(g; \theta) &= (1 - \theta)^2. \end{aligned}$$

If a sample of three balls $X_1 = r, X_2 = b, X_3 = r$ is observed, then

$$L(\theta; X_1, X_2, X_3) = f(r, b, r; \theta) = f(r, \theta)f(b, \theta)f(r, \theta) = 2\theta^5(1 - \theta).$$

Taking the derivative of the logarithm of $L(\theta; X_1, X_2, X_3)$ and setting it to zero, the likelihood equation is obtained

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{5}{\theta} - \frac{1}{1 - \theta} = 0,$$

which has the unique solution $\theta = \frac{5}{6}$. Because

$$\frac{\partial^2 \log L(\theta)}{\partial \theta^2} = -\frac{5}{\theta^2} - \frac{1}{(1 - \theta)^2} < 0$$

for all $\theta \in (0, 1)$, $\theta = \frac{5}{6}$ maximizes $L(\theta)$.

Example 3: This is an example taken from [99]. Suppose one wanted to find estimates of a normal distribution with unknown mean μ and unknown variance v . The likelihood function for a random sample of size n is

$$L(\Theta) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi v}} e^{-\frac{(X_i - \mu)^2}{2v}} = \left(\frac{1}{2\pi v} \right)^{\frac{n}{2}} e^{-\frac{1}{2v} \sum_{i=1}^n (X_i - \mu)^2}. \quad (8)$$

Since Eq. (8) has an exponential expression, the logarithm can be used to obtain

$$\log L(\Theta) = -\frac{n}{2} \log(2\pi v) - \frac{1}{2v} \sum_{i=1}^n (X_i - \mu)^2. \quad (9)$$

By taking the partial derivative of Eq. (9) with respect to μ and v , the following two likelihood equations can be obtained:

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{v} \sum_{i=1}^n (X_i - \mu), \quad (10)$$

and

$$\frac{\partial \log L}{\partial v} = -\frac{n}{2v} + \frac{1}{2v^2} \sum_{i=1}^n (X_i - \mu)^2. \quad (11)$$

By setting Eqs. (10) and (11) to zero and solving them for μ and v respectively, the maximum likelihood estimates are obtained with

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i,$$

and

$$\hat{v} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

3.3. The Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is a method for finding maximum-likelihood estimates of population parameters of an underlying distribution from a given incomplete data set. It provides an iterative scheme for obtaining maximum likelihood estimates, converting a hard problem into a sequence of simpler problems. The EM algorithm obtains the initial estimates for population parameters either by random guess or previous knowledge of the data. Then it iteratively uses the estimates for

the missing data to obtain new estimates and continues until estimates converge.

The Basic EM algorithm was defined in [16] as follows: let \mathcal{X} be an incomplete data set observed or generated by some distribution with unknown parameters $\theta_1, \theta_2, \dots, \theta_k$ and \mathcal{Y} be the unknown data set. To simplify the notation, Θ is used to represent these unknown parameters. Assume that a complete data set $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ exists and assume a joint density function,

$$p(z; \Theta) = p(x, y; \Theta) = p(y; x, \Theta)p(x; \Theta),$$

where $x \in \mathcal{X}$, $y \in \mathcal{Y}$ and $z \in \mathcal{Z}$. With this joint density function, the complete-data likelihood function can be defined as

$$L(\Theta; \mathcal{Z}) = L(\Theta; \mathcal{X}, \mathcal{Y}) = p(\mathcal{X}, \mathcal{Y}; \Theta).$$

The EM algorithm first finds the expected value of the complete-data log-likelihood $\log p(\mathcal{X}, \mathcal{Y}; \Theta)$ with respect to the unknown data \mathcal{Y} , given the observed data \mathcal{X} and current parameter estimates.

Define

$$Q(\Theta, \hat{\Theta}^{i-1}) = E(\log p(\mathcal{X}, \mathcal{Y}; \Theta) | \mathcal{X}, \hat{\Theta}^{i-1}), \quad (12)$$

where $\hat{\Theta}^{i-1}$ is the current parameter estimates used to evaluate the expectation and Θ are the new parameters optimized to increase the value of Q . On the right-hand side of Eq. (12), \mathcal{X} and $\hat{\Theta}^{i-1}$ are constants, Θ is a normal variable to be adjusted, and \mathcal{Y} is a random variable governed by the distribution of the data. So, the right-hand side of Eq. (12) can be re-written as:

$$E(\log p(\mathcal{X}, \mathcal{Y}; \Theta) | \mathcal{X}, \hat{\Theta}^{i-1}) = \int_{y \in \Upsilon} \log p(\mathcal{X}, y; \Theta) f(y; \mathcal{X}, \hat{\Theta}^{i-1}) dy,$$

where f is the marginal distribution of the unknown data \mathcal{Y} that depends on both the observed data \mathcal{X} and the current parameter estimates $\hat{\Theta}^{i-1}$, and Υ is the space of values y can take on. The evaluation of $E(\log p(\mathcal{X}, \mathcal{Y}; \Theta) | \mathcal{X}, \hat{\Theta}^{i-1})$ is called the E-step of the EM algorithm.

The second step of the EM algorithm maximizes $E(\log p(\mathcal{X}, \mathcal{Y}; \Theta) | \mathcal{X}, \hat{\Theta}^{i-1})$ and sets

$$\hat{\Theta}^i = \underset{\Theta}{\operatorname{argmax}} Q(\Theta, \hat{\Theta}^{i-1}).$$

This is called the M-step. The E-step and M-step steps are repeated as necessary. Each iteration is guaranteed to increase the log-likelihood of the

complete-data and the algorithm is guaranteed to converge to a local maximum of the likelihood function [16]. The EM algorithm can be summarized as follows:

Algorithm 1 EM Algorithm

Input:

Data set distribution with unknown parameters $\Theta = \{\theta_1, \dots, \theta_k\}$;
 Incomplete data set \mathcal{X} ;
 Convergence threshold ϵ ;

Output:

Parameter estimates $\hat{\Theta} = \{\hat{\theta}_1, \dots, \hat{\theta}_k\}$

Initialize $\hat{\Theta}^0$

$n \leftarrow 0$;

repeat

$n \leftarrow n + 1$;

$Q(\Theta, \hat{\Theta}^{n-1}) \leftarrow E(\log p(\mathcal{X}, \mathcal{Y}; \Theta) | \mathcal{X}, \hat{\Theta}^{n-1})$;

Compute the maximum likelihood estimates of Θ to maximize

$Q(\Theta, \hat{\Theta}^{n-1})$;

$\hat{\Theta}^n \leftarrow \operatorname{argmax}_{\Theta} Q(\Theta, \hat{\Theta}^{n-1})$;

until $|\hat{\Theta}^n - \hat{\Theta}^{n-1}| < \epsilon$

$\hat{\Theta} \leftarrow \hat{\Theta}^n$;

The EM algorithm is useful in computational pattern recognition [16], image retrieval [159], computer vision [101], and many other fields. In data mining, the EM algorithm can be used when the data set has missing values due to limitations of the observation process. It is especially useful when maximizing the likelihood function directly is analytically intractable. In that case, the likelihood function can be simplified by assuming that the hidden parameters are known.

4. Measures of Performance

As discussed in Sec. 3, there are several different ways (estimators) to estimate unknown parameters. In order to assess the usefulness of estimators, some criteria are necessary to measure the performance of estimators. In this section, five criteria used to assess estimators — bias, mean squared error, standard error, efficiency, and consistency will be discussed. At the end of this section, the Jackknife method will be introduced to estimate

bias and standard error of an estimator. As discussed before, an estimator is a function that maps a random sample to a set of parameter estimates. Furthermore, if the sample obtained is a random sample, an estimator is also a random variable since the estimates are calculated using the sample. In the following discussion, $\hat{\theta}$ is denoted as the estimator (random variable) of an unknown parameter θ .

4.1. Bias

The bias of an estimator provides a measure of the average error in the estimator $\hat{\theta}$ of a parameter θ . The *bias* of an estimator is defined as the difference between the expected value of the estimator and the actual value

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta. \quad (13)$$

An estimator is *unbiased* if the expected value of the estimator equals the true parameter value, i.e., $E[\hat{\theta}] = \theta$. Otherwise, the estimator is *biased*. For example, the maximum likelihood estimate of the mean for a normal distribution is unbiased, since $E[\hat{\mu}] = \mu$ [109]. However, this it is not the case for the maximum likelihood estimate of the variance \hat{v} [68]. It can be shown that $E[\hat{v}] = \frac{(n-1)}{n}v$, where n is the sample size.

To determine the expected value in Eq. (13), the distribution of the statistic $\hat{\theta}$ must be known to analytically calculate the bias. If the distribution of the statistic is not known, then some methods such as the Jackknife (see Sec. 4.6) can be used to estimate the bias of $\hat{\theta}$.

4.2. Mean Squared Error

The *mean squared error* (MSE) is the expected value of the squared error. Let θ be a parameter and $\hat{\theta}$ be an estimator of the parameter, the mean squared error of the estimator is defined as

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]. \quad (14)$$

It is sometimes more useful to rewrite the MSE equation in terms of the bias and the variance [109]. The first step of the rewriting is to expand the expected value on the right-hand side of Eq. (14) to get

$$MSE(\hat{\theta}) = E[(\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2)] = E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2. \quad (15)$$

The next step of the rewriting is to add to and subtract $(E[\hat{\theta}])^2$ from the right-hand side of Eq. (15) so that

$$MSE(\hat{\theta}) = E[\hat{\theta}^2] - (E[\hat{\theta}])^2 + (E[\hat{\theta}])^2 - 2\theta E[\hat{\theta}] + \theta^2. \quad (16)$$

By simplifying Eq. (16), the mean squared error can be written as

$$MSE(\hat{\theta}) = E[\hat{\theta}^2] - (E[\hat{\theta}])^2 + (E[\hat{\theta}] - \theta)^2 = V(\hat{\theta}) + [bias(\hat{\theta})]^2. \quad (17)$$

Equation (17) shows how the mean squared error, variance and bias of an estimator are related. Since the mean squared error is the sum of the variance and the squared bias, two non-negative quantities, the error will be small when the variance and the absolute value of the bias are both small. When $\hat{\theta}$ is unbiased, the mean squared error is equal to the variance.

4.3. Standard Error

The standard error gives a measure of the precision of the estimators. The *standard error* of an estimator $\hat{\theta}$ is defined as the standard deviation of its sampling distribution

$$SE(\hat{\theta}) = \sqrt{V(\hat{\theta})} = \sigma_{\hat{\theta}}.$$

The sample mean can be used as an example to illustrate the concept of standard error. Let $f(x)$ represent a probability density function with finite variance σ^2 and mean μ . Let \bar{X} be the sample mean for a random sample of size n drawn from this distribution. By the Central Limit Theorem [105], the distribution of \bar{X} is approximately normally distributed with mean μ and variance $\frac{\sigma^2}{n}$. So the standard error is given by

$$SE(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

When the standard deviation σ for the underlying population is unknown, then an estimate S for the parameter can be used as a substitute for it and leads to the estimated standard error

$$\widehat{SE}(\bar{X}) = \hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}.$$

4.4. Efficiency

Another measure used to compare estimators is called efficiency. Suppose there are two estimators $\hat{\theta}$ and $\hat{\theta}'$ for a parameter θ based on the sample X_1, \dots, X_n . If the MSE of one estimator is less than the MSE of the other,

i.e., $MSE(\hat{\theta}) < MSE(\hat{\theta}')$, then the estimator $\hat{\theta}$ is said to be more *efficient* than $\hat{\theta}'$. The *relative efficiency* of $\hat{\theta}$ with respect to $\hat{\theta}'$ is defined as the ratio

$$eff(\hat{\theta}, \hat{\theta}') = \frac{MSE(\hat{\theta}')}{MSE(\hat{\theta})}.$$

If this ratio is greater than one, then $\hat{\theta}$ is a more efficient estimator of the parameter θ . When the estimator is unbiased, the ratio is just the ratio of their variance, and the most efficient estimator would be the one with minimum variance.

4.5. Consistency

Unlike the four measures defined in previous subsections, consistency is defined for increasing sample sizes, not a fixed sample sizes. Like the efficiency, consistency is also defined using the MSE. Let $\hat{\theta}_n$ be the estimator of a parameter based on a sample of size n , then an estimator is said to be *consistent* if

$$\lim_{n \rightarrow \infty} MSE(\hat{\theta}_n) = 0, \quad (18)$$

or

$$\lim_{n \rightarrow \infty} [V(\hat{\theta}_n) + (bias(\hat{\theta}_n))^2] = 0 \quad (19)$$

when MSE is written in terms of bias and variance. Thus, Eq. (18) or Eq. (19) holds if and only if both variance and bias of $\hat{\theta}_n$ tend to zero as n approaches infinite.

4.6. The Jackknife Method

Given a random sample and a parameter θ , its estimate is also a random variable and has some error associated with it. Estimates of bias and standard error of the estimator $\hat{\theta}$ can assess the accuracy of the results. The Jackknife method is a technique for estimating bias and standard error of statistics [44].

The Jackknife obtains the estimate of a parameter from a set of observed data by generating that statistic repeatedly on the data set excluding a single data value during each iteration. The Jackknife method consists of taking repeated sub-samples of the original sample of n independent observations by omitting a single observation at a time. Thus, each sub-sample

Algorithm 2 Jackknife

Input:

An estimator $\hat{\theta} = \tau(X_1, \dots, X_n)$; a random sample X_1, \dots, X_n ;

Output:

Jackknife estimate of bias of $\hat{\theta}$; Jackknife estimate of standard error of $\hat{\theta}$;

for $i = 1$ to n

Leave out the sample point X_i ;

Calculate the value of the statistic using remaining sample points to obtain $\hat{\theta}_{(i)}$;

end for

Calculate the overall Jackknife estimate using Eq. (20), the Jackknife estimate of bias of $\hat{\theta}$ using Eq. (21), and the Jackknife estimate of standard error of $\hat{\theta}$ using Eq. (22).

consists of $n - 1$ observations formed by deleting a different observation from the sample. The Jackknife estimate and its standard error are then calculated from these truncated sub-samples.

Suppose that there is a set of n values X_1, \dots, X_n , the i th Jackknife estimate is calculated by omitting the i th value

$$\hat{\theta}_{(i)} = \tau(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

For example, the i th Jackknife estimate for the mean μ would be

$$\hat{\mu}_{(i)} = (1/n - 1) \sum_{j=1}^{i-1} X_j + (1/n - 1) \sum_{j=i+1}^n X_j.$$

Given a set of Jackknife estimates, $\hat{\theta}_{(i)}$, $i = 1, 2, \dots, n$, an overall estimate, $\hat{\theta}_{(\cdot)}$ can be obtained by

$$\hat{\theta}_{(\cdot)} = (1/n) \sum_{i=1}^n \hat{\theta}_{(i)}. \quad (20)$$

The estimate of the bias of $\hat{\theta}$ obtained by the Jackknife method is given by [44]

$$\widehat{Bias}_{jack}(\hat{\theta}) = (n - 1)(\hat{\theta}_{(\cdot)} - \hat{\theta}). \quad (21)$$

The estimated standard error of $\hat{\theta}$ using the Jackknife method is defined as

$$\widehat{SE}_{jack}(\hat{\theta}) = \left[\frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{(i)} - \hat{\theta}_{(\cdot)})^2 \right]^{\frac{1}{2}}. \quad (22)$$

5. Summary

In this chapter, the method of moments, maximum likelihood estimation, and the EM algorithm have been discussed with simple examples. Even though classical point estimation is a useful theoretical topic, it requires some knowledge about the data involved and violates an important principle of data mining — avoid making any assumptions about the data. Also, point estimation is too simple for data mining applications that have huge data sets and complex processing models. The need to solve real problems has driven the evolution of estimation techniques and algorithms. It has progressed from least squares to the method of moments, to maximum likelihood, to Bayes and empirical Bayes procedures, to risk-reduction approaches, to robustness, and to re-sampling techniques [82]. Readers interested in further details of these advanced topics will benefit from reading [95].