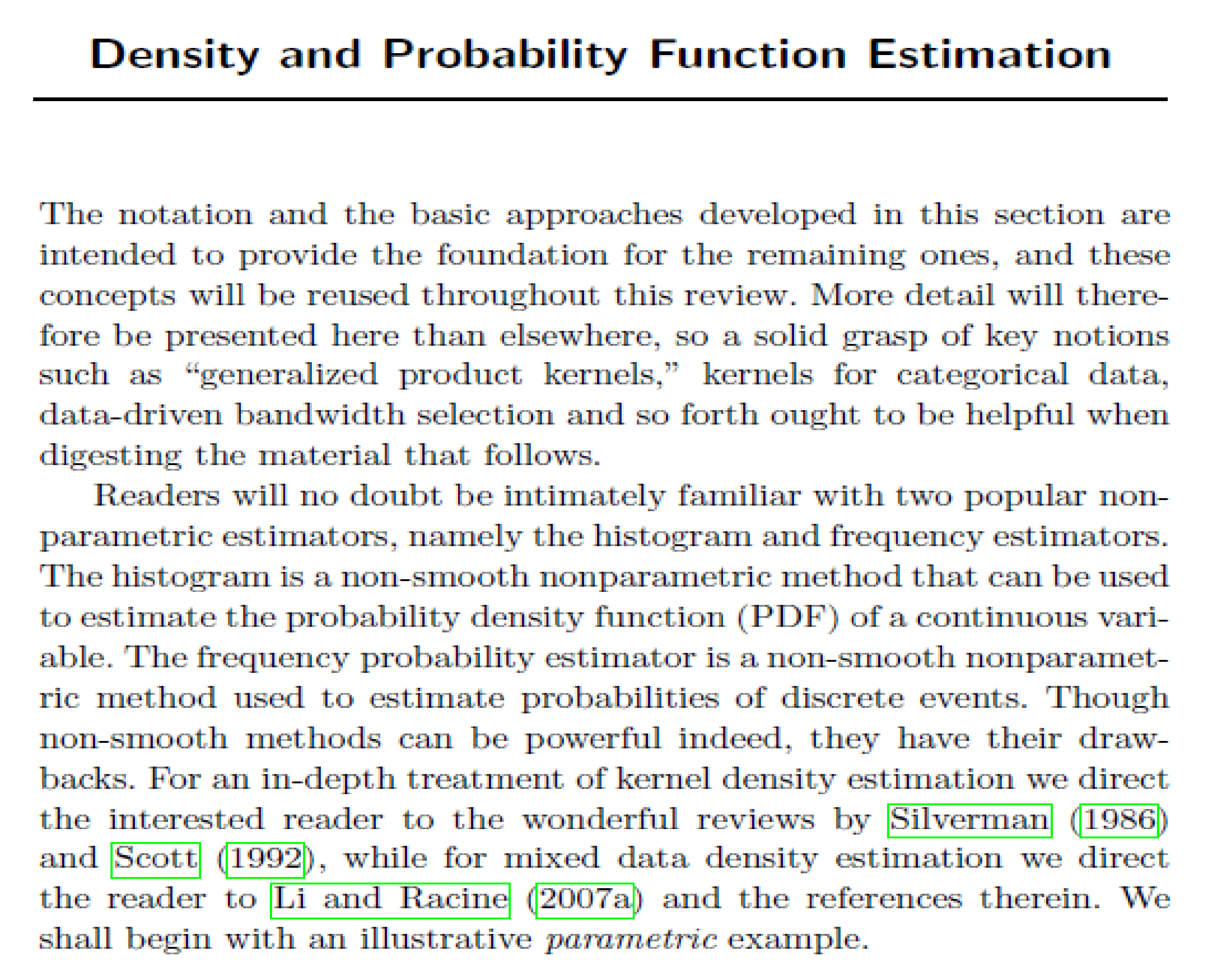
Lect. No. 1 Smoothing Techniques

Prof. Munaf Yousif Hmood

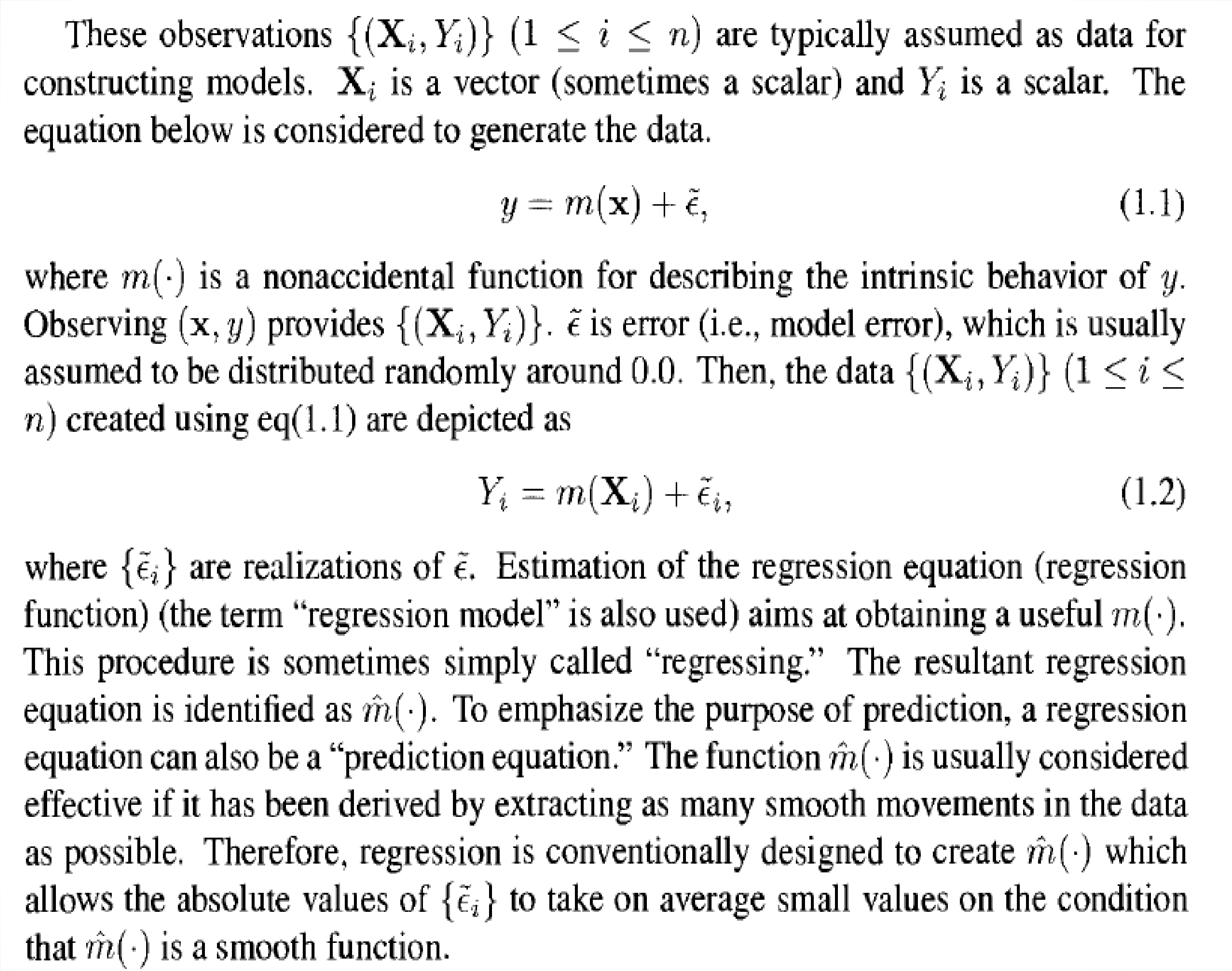
1-11-2020



Density and Probability Function Estimation

The notation and the basic approaches developed in this section are intended to provide the foundation for the remaining ones, and these concepts will be reused throughout this review. More detail will therefore be presented here than elsewhere, so a solid grasp of key notions such as "generalized product kernels," kernels for categorical data, data-driven bandwidth selection and so forth ought to be helpful when digesting the material that follows.

Readers will no doubt be intimately familiar with two popular nonparametric estimators, namely the histogram and frequency estimators. The histogram is a non-smooth nonparametric method that can be used to estimate the probability density function (PDF) of a continuous variable. The frequency probability estimator is a non-smooth nonparametric method used to estimate probabilities of discrete events. Though non-smooth methods can be powerful indeed, they have their drawbacks. For an in-depth treatment of kernel density estimation we direct the interested reader to the wonderful reviews by Silverman (1986) and Scott (1992), while for mixed data density estimation we direct the reader to Li and Racine (2007a) and the references therein. We shall begin with an illustrative *parametric* example.



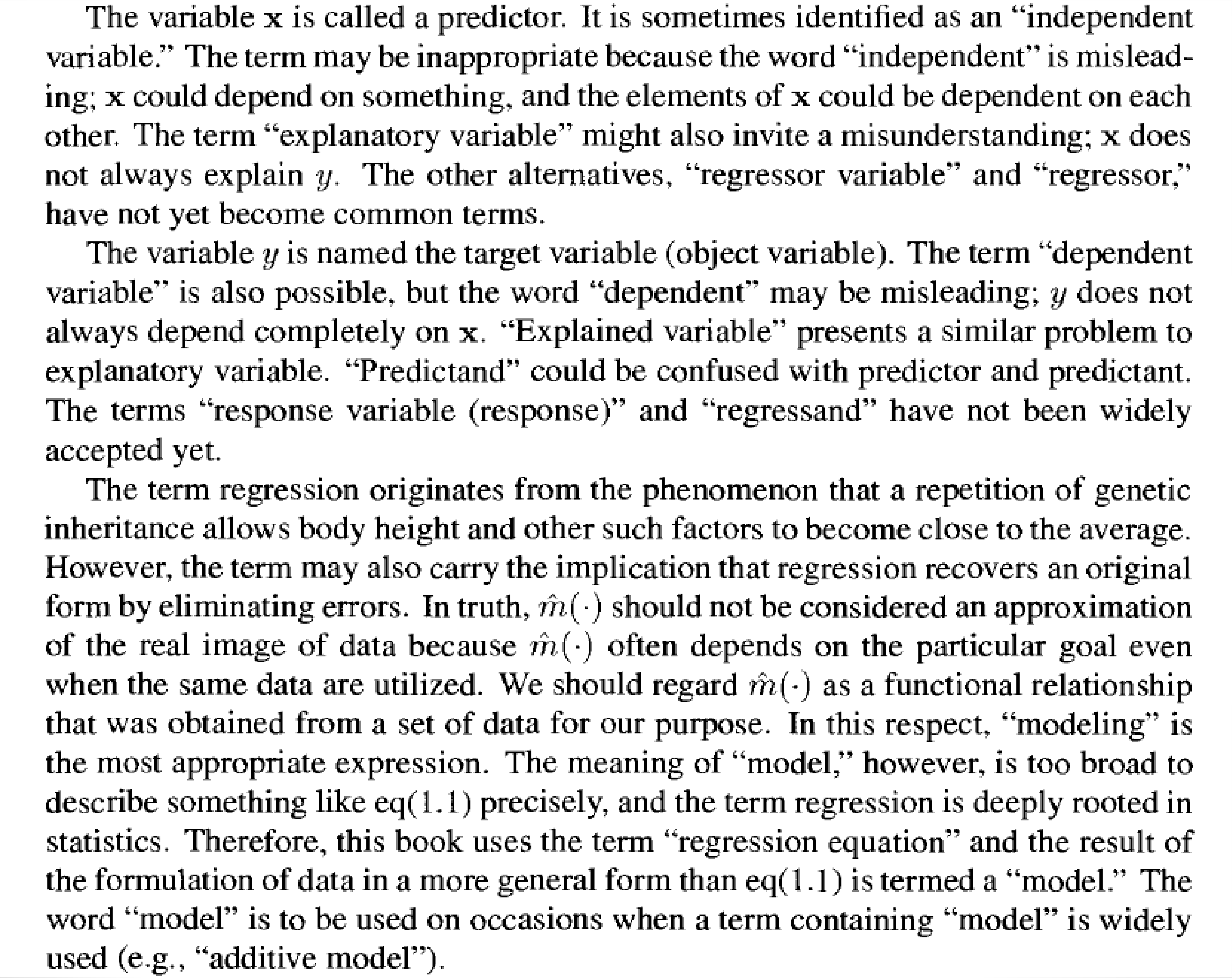
These observations $\{(\mathbf{X}_i, Y_i)\}$ $(1 \le i \le n)$ are typically assumed as data for constructing models. $\mathbf{X}_i$ is a vector (sometimes a scalar) and Y_i is a scalar. The equation below is considered to generate the data.

$$y = m(\mathbf{x}) + \tilde{\epsilon}, \tag{1.1}$$

where $m(\cdot)$ is a nonaccidental function for describing the intrinsic behavior of y. Observing $(\mathbf{x}, y)$ provides $\{(\mathbf{X}_i, Y_i)\}$. $\tilde{\epsilon}$ is error (i.e., model error), which is usually assumed to be distributed randomly around 0.0. Then, the data $\{(\mathbf{X}_i, Y_i)\}$ $(1 \le i \le n)$ created using eq(1.1) are depicted as

$$Y_i = m(\mathbf{X}_i) + \tilde{\epsilon}_i, \tag{1.2}$$

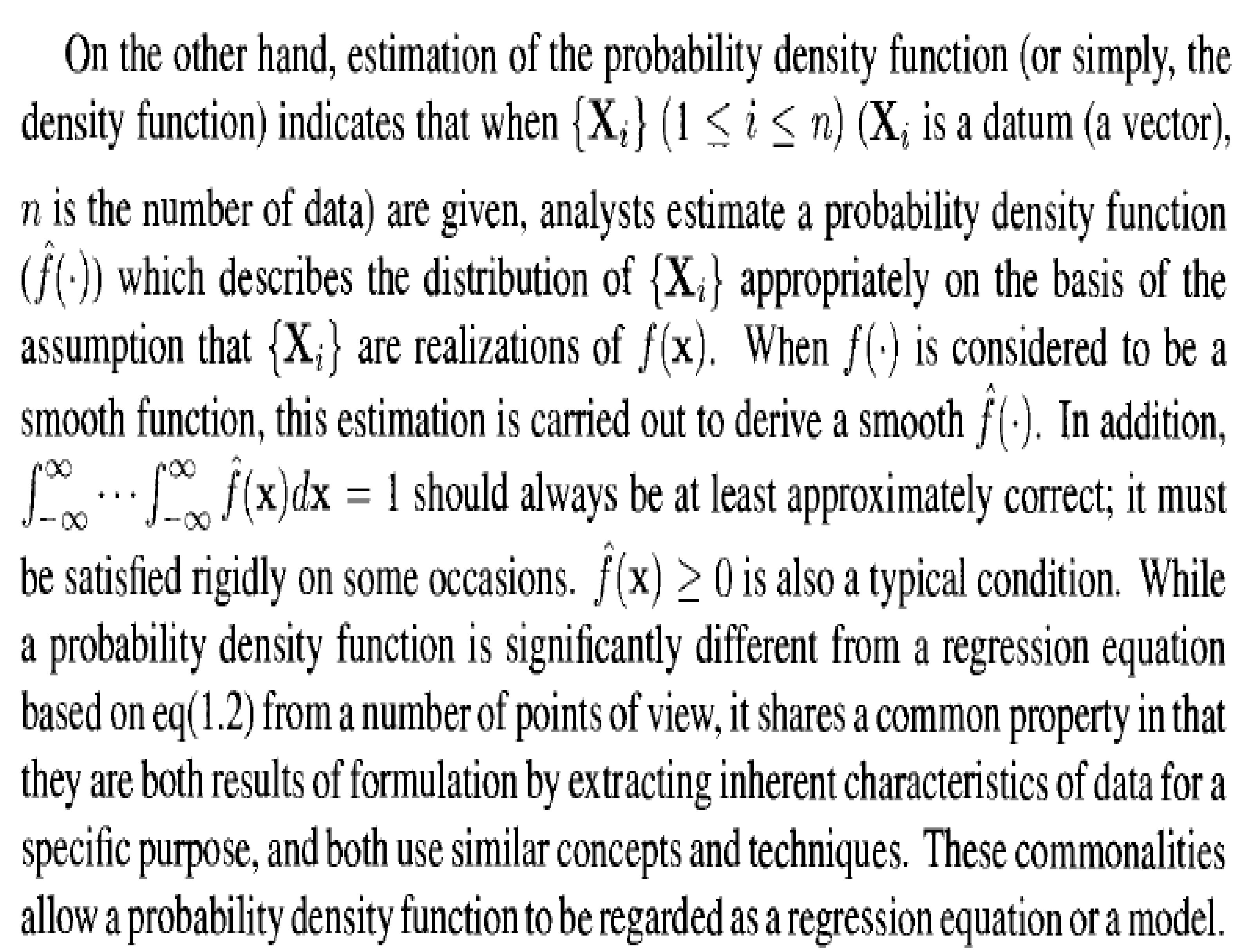
where $\{\tilde{\epsilon}_i\}$ are realizations of $\tilde{\epsilon}$. Estimation of the regression equation (regression function) (the term "regression model" is also used) aims at obtaining a useful $m(\cdot)$. This procedure is sometimes simply called "regressing." The resultant regression equation is identified as $\hat{m}(\cdot)$. To emphasize the purpose of prediction, a regression equation can also be a "prediction equation." The function $\hat{m}(\cdot)$ is usually considered effective if it has been derived by extracting as many smooth movements in the data as possible. Therefore, regression is conventionally designed to create $\hat{m}(\cdot)$ which allows the absolute values of $\{\tilde{\epsilon}_i\}$ to take on average small values on the condition that $\hat{m}(\cdot)$ is a smooth function.



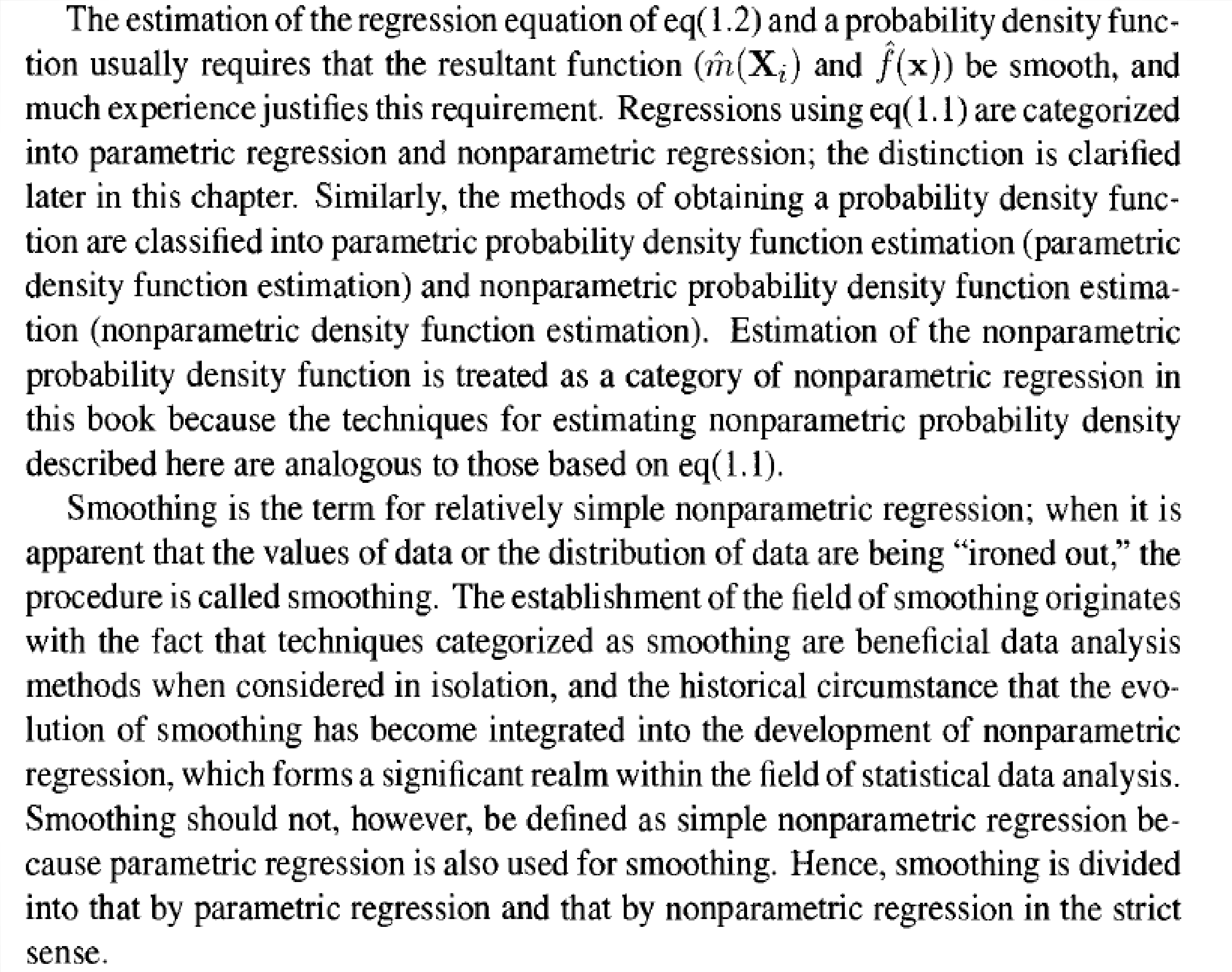
The variable **x** is called a predictor. It is sometimes identified as an "independent variable." The term may be inappropriate because the word "independent" is misleading; **x** could depend on something, and the elements of **x** could be dependent on each other. The term "explanatory variable" might also invite a misunderstanding; **x** does not always explain y. The other alternatives, "regressor variable" and "regressor," have not yet become common terms.

The variable y is named the target variable (object variable). The term "dependent variable" is also possible, but the word "dependent" may be misleading; y does not always depend completely on **x**. "Explained variable" presents a similar problem to explanatory variable. "Predictand" could be confused with predictor and predictant. The terms "response variable (response)" and "regressand" have not been widely accepted yet.

The term regression originates from the phenomenon that a repetition of genetic inheritance allows body height and other such factors to become close to the average. However, the term may also carry the implication that regression recovers an original form by eliminating errors. In truth, $\hat{m}(\cdot)$ should not be considered an approximation of the real image of data because $\hat{m}(\cdot)$ often depends on the particular goal even when the same data are utilized. We should regard $\hat{m}(\cdot)$ as a functional relationship that was obtained from a set of data for our purpose. In this respect, "modeling" is the most appropriate expression. The meaning of "model," however, is too broad to describe something like eq(1.1) precisely, and the term regression is deeply rooted in statistics. Therefore, this book uses the term "regression equation" and the result of the formulation of data in a more general form than eq(1.1) is termed a "model." The word "model" is to be used on occasions when a term containing "model" is widely used (e.g., "additive model").



On the other hand, estimation of the probability density function (or simply, the density function) indicates that when $\{\mathbf{X}_i\}$ $(1 \leq i \leq n)$ ($\mathbf{X}_i$ is a datum (a vector), n is the number of data) are given, analysts estimate a probability density function $(\hat{f}(\cdot))$ which describes the distribution of $\{\mathbf{X}_i\}$ appropriately on the basis of the assumption that $\{\mathbf{X}_i\}$ are realizations of $f(\mathbf{x})$. When $f(\cdot)$ is considered to be a smooth function, this estimation is carried out to derive a smooth $\hat{f}(\cdot)$. In addition, $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \hat{f}(\mathbf{x}) d\mathbf{x} = 1$ should always be at least approximately correct; it must be satisfied rigidly on some occasions. $\hat{f}(\mathbf{x}) \geq 0$ is also a typical condition. While a probability density function is significantly different from a regression equation based on eq(1.2) from a number of points of view, it shares a common property in that they are both results of formulation by extracting inherent characteristics of data for a specific purpose, and both use similar concepts and techniques. These commonalities allow a probability density function to be regarded as a regression equation or a model.



The estimation of the regression equation of eq(1.2) and a probability density function usually requires that the resultant function ($\hat{m}(\mathbf{X}_i)$ and $\hat{f}(\mathbf{x})$) be smooth, and much experience justifies this requirement. Regressions using eq(1.1) are categorized into parametric regression and nonparametric regression; the distinction is clarified later in this chapter. Similarly, the methods of obtaining a probability density function are classified into parametric probability density function estimation (parametric density function estimation) and nonparametric probability density function estimation (nonparametric density function estimation). Estimation of the nonparametric probability density function is treated as a category of nonparametric regression in this book because the techniques for estimating nonparametric probability density described here are analogous to those based on eq(1.1).

Smoothing is the term for relatively simple nonparametric regression; when it is apparent that the values of data or the distribution of data are being "ironed out," the procedure is called smoothing. The establishment of the field of smoothing originates with the fact that techniques categorized as smoothing are beneficial data analysis methods when considered in isolation, and the historical circumstance that the evolution of smoothing has become integrated into the development of nonparametric regression, which forms a significant realm within the field of statistical data analysis. Smoothing should not, however, be defined as simple nonparametric regression because parametric regression is also used for smoothing. Hence, smoothing is divided into that by parametric regression and that by nonparametric regression in the strict sense.

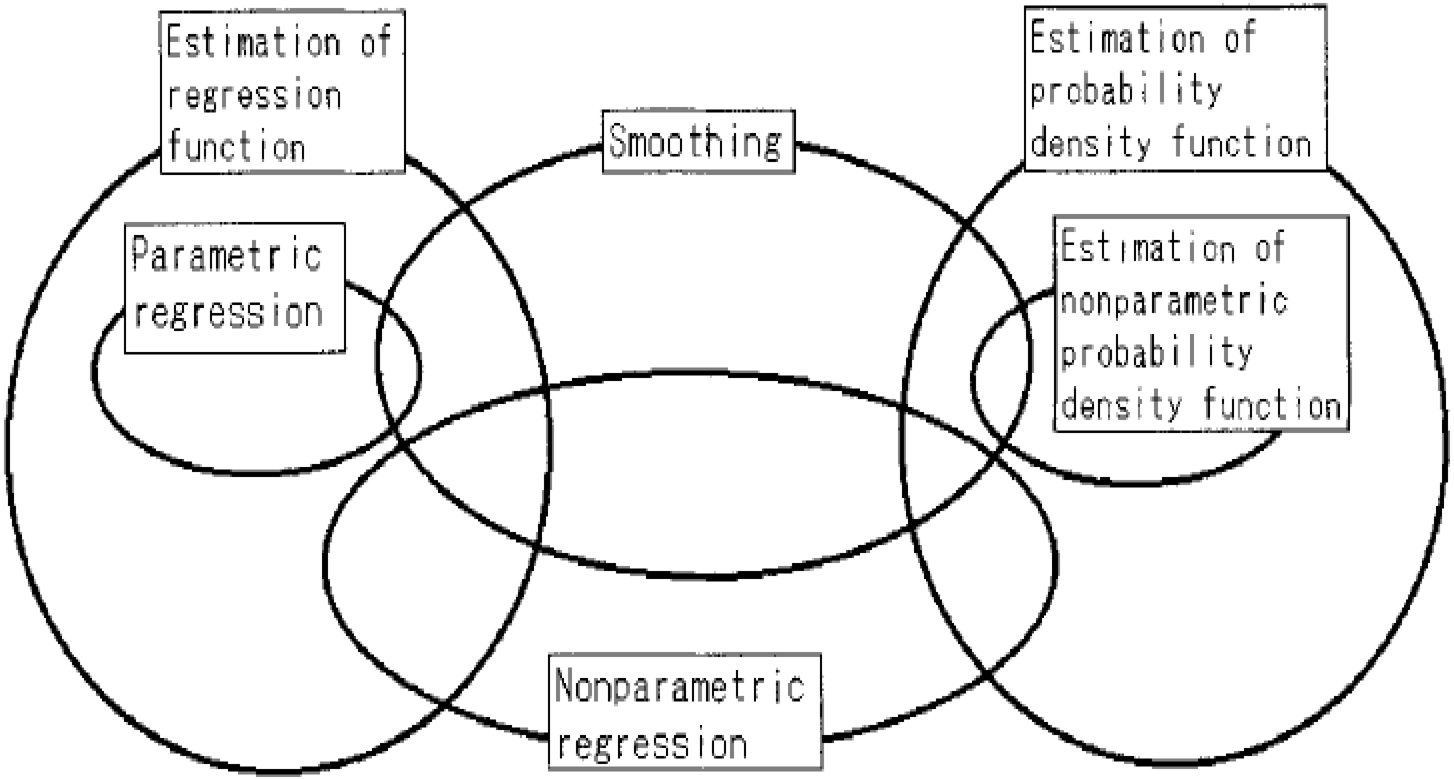

Estimation of regression function
Smoothing
Estimation of probability density function
Parametric regression
Estimation of nonparametric probability density function
Nonparametric regression

Figure 1.1 Venn diagram to show the relationship of parametric regression and nonparametric regression. This relationship is not necessarily a conclusively established one, and therefore other mappings may be reported in the literature.

1.2 ARE THE MOVING AVERAGE AND FOURIER SERIES SUFFICIENTLY USEFUL?

Figure 1.2 illustrates the monthly average exchange rate of US dollar and yen from February 1987 through May 1999. A long term trend with a superimposed fine oscillation forms the shape; the fine oscillation is the short term variation. The subtraction of fine oscillation (i.e., the short term variation) from the longitudinal data gives the long term variation. The two types of variations reflect different mechanisms in economic change and hence the separation of the data into these two trends allows a clear understanding. It is reasonable that a long term trend should be considered as a gentle curve extracted from data, and the subtraction of this trend from the data gives the short term variation. Such an extraction of long term variation is one role of smoothing.

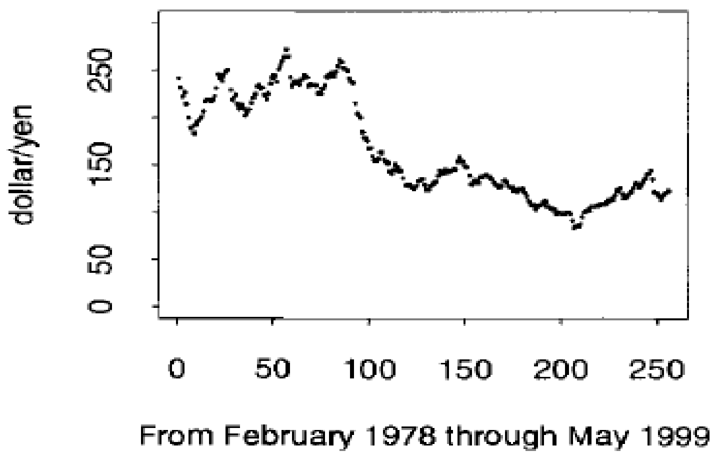
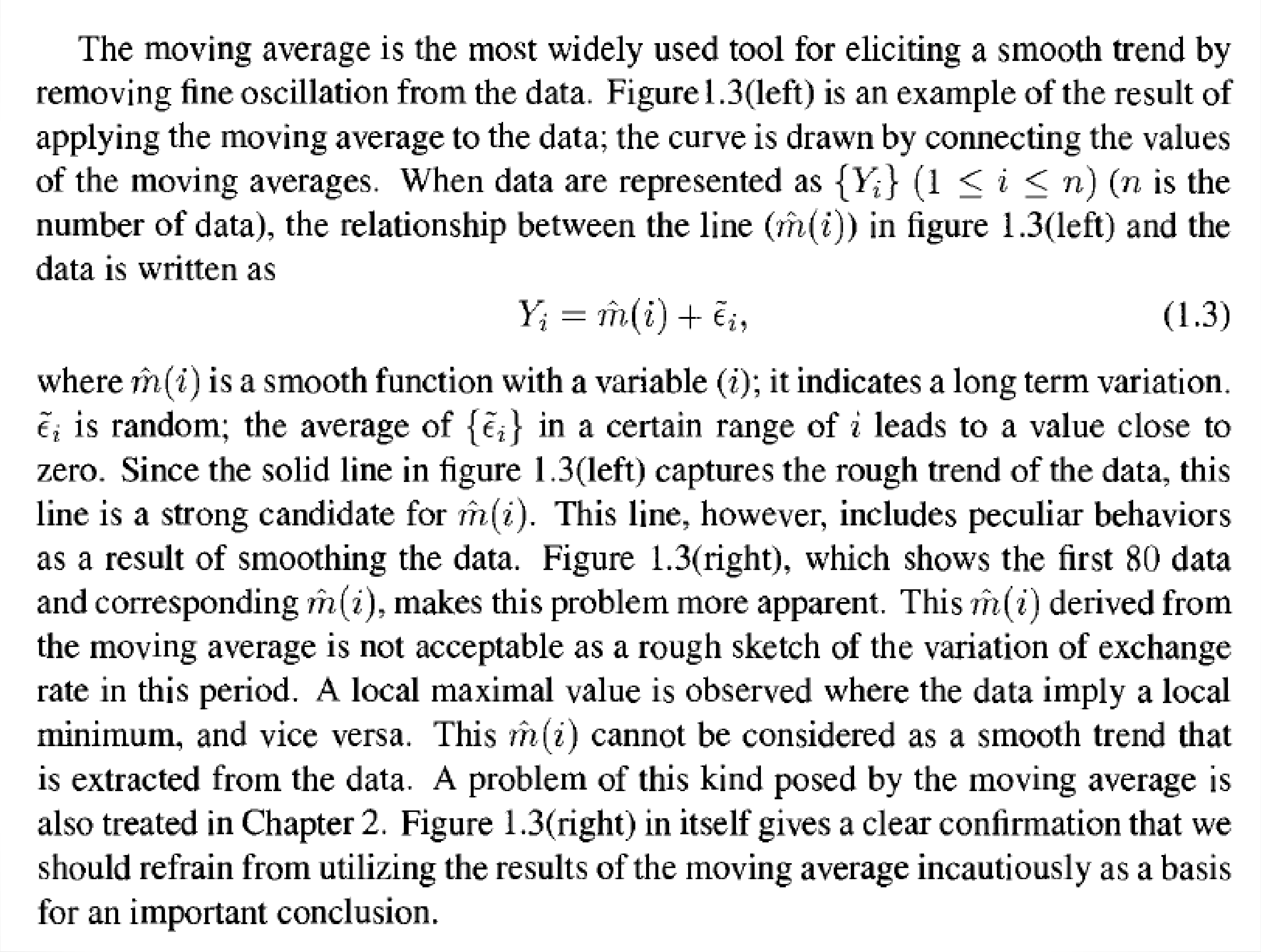


Figure 1.2 Monthly average exchange rate of US dollar and yen from February 1987 through May 1999. Numbers on the x-axis indicate months beginning at February 1987.



The moving average is the most widely used tool for eliciting a smooth trend by removing fine oscillation from the data. Figure1.3(left) is an example of the result of applying the moving average to the data; the curve is drawn by connecting the values of the moving averages. When data are represented as $\{Y_i\}$ $(1 \leq i \leq n)$ (n is the number of data), the relationship between the line ($\hat{m}(i)$) in figure 1.3(left) and the data is written as

$$Y_i = \hat{m}(i) + \tilde{\epsilon}_i, \tag{1.3}$$

where $\hat{m}(i)$ is a smooth function with a variable (i); it indicates a long term variation. $\tilde{\epsilon}_i$ is random; the average of $\{\tilde{\epsilon}_i\}$ in a certain range of i leads to a value close to zero. Since the solid line in figure 1.3(left) captures the rough trend of the data, this line is a strong candidate for $\hat{m}(i)$. This line, however, includes peculiar behaviors as a result of smoothing the data. Figure 1.3(right), which shows the first 80 data and corresponding $\hat{m}(i)$, makes this problem more apparent. This $\hat{m}(i)$ derived from the moving average is not acceptable as a rough sketch of the variation of exchange rate in this period. A local maximal value is observed where the data imply a local minimum, and vice versa. This $\hat{m}(i)$ cannot be considered as a smooth trend that is extracted from the data. A problem of this kind posed by the moving average is also treated in Chapter 2. Figure 1.3(right) in itself gives a clear confirmation that we should refrain from utilizing the results of the moving average incautiously as a basis for an important conclusion.

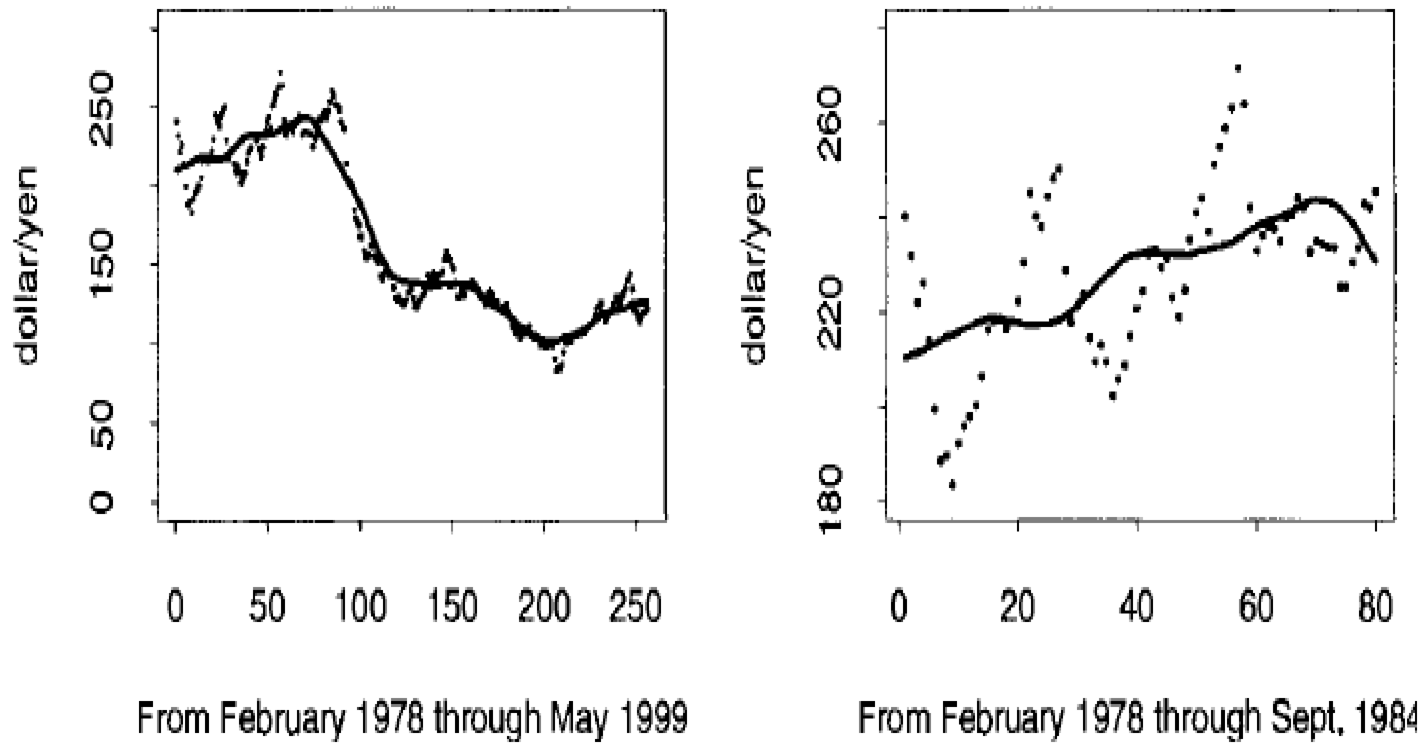


Figure 1.3 Example of the result of moving average using the data in figure 1.2. The result of extraction of the first 80 data from figure 1.3 (left) and corresponding $\hat{m}(i)$ (right).

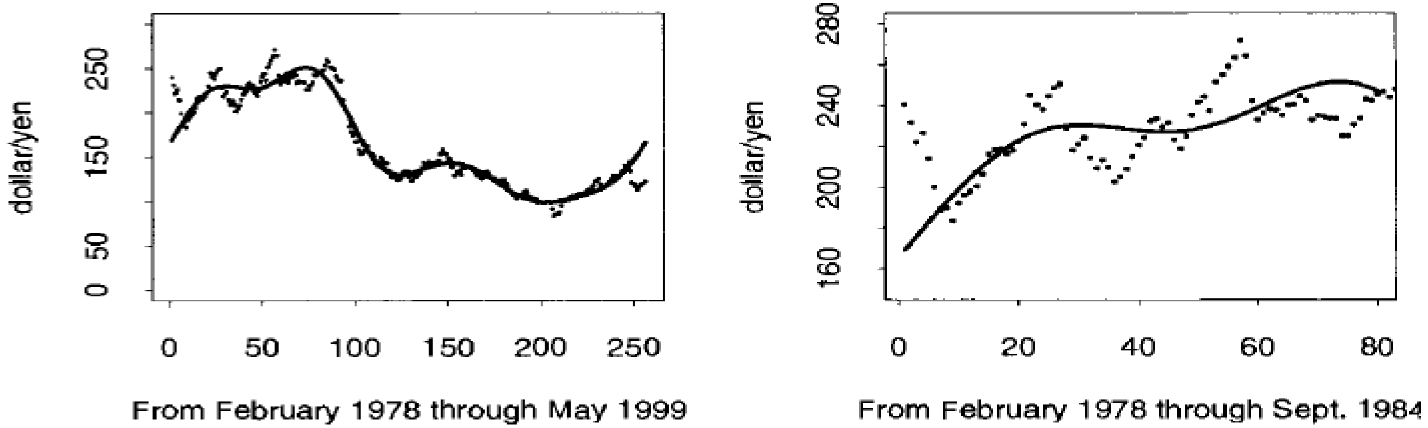


Figure 1.4 A smooth line ($\hat{m}(i)$) that is the sum of longer waves. Waves are obtained by the decomposition into sine waves and cosine waves (left). The result of extraction of the first 80 data from those in figure 1.4(left) and corresponding $\hat{m}(i)$ (right).

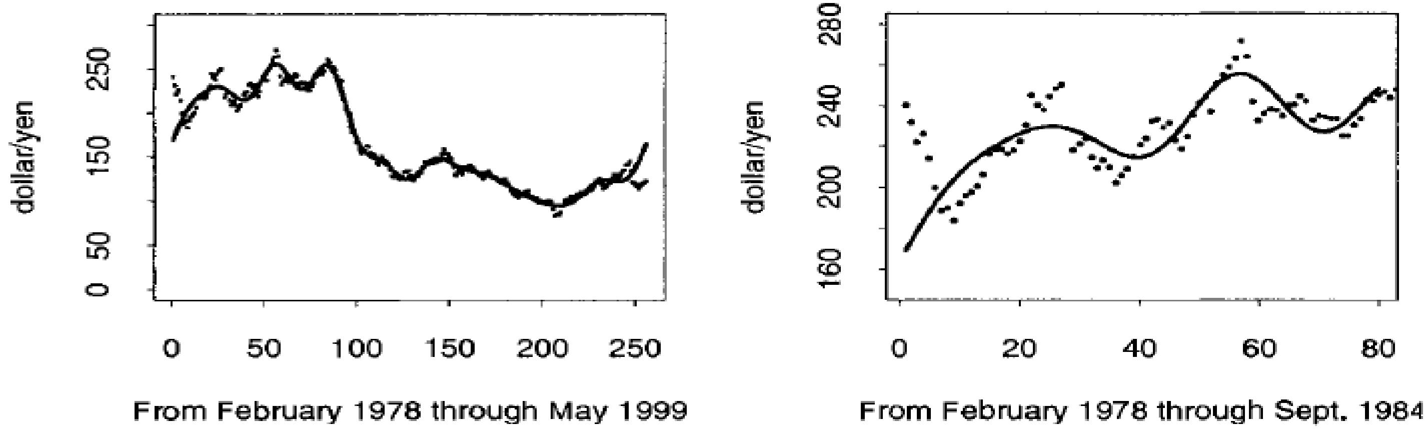


Figure 1.5 A smooth line ($\hat{m}(i)$) that is the sum of longer waves. Waves are obtained by the decomposition into sine waves and cosine waves (the number of waves is greater than that in figure 1.4(right)), (left). The result of extraction of the first 80 data from those in figure 1.4(left) and corresponding $\hat{m}(i)$ (right).

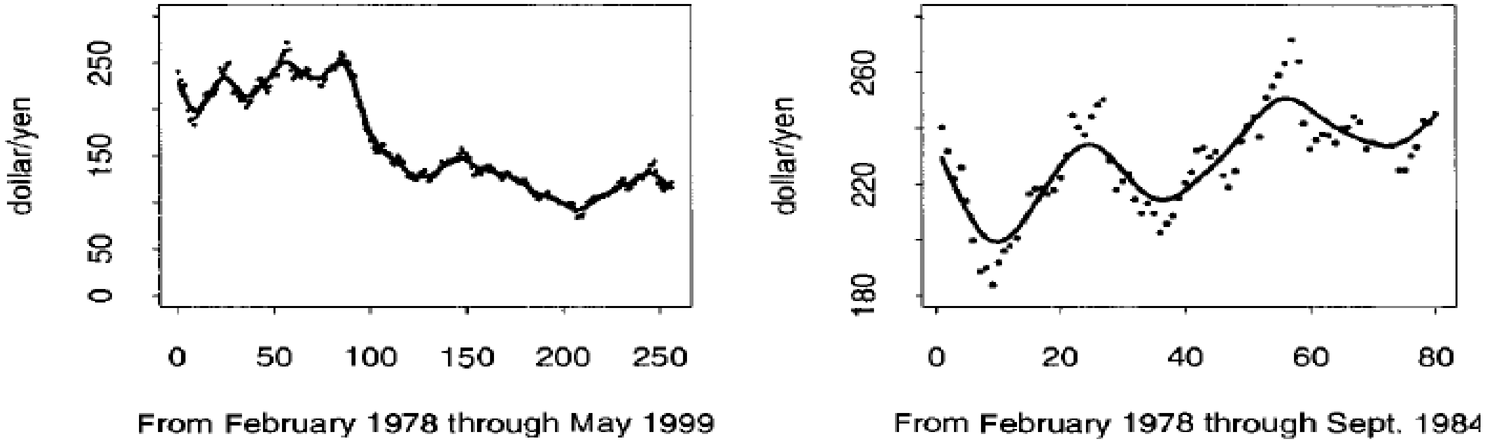


Figure 1.6 Curve obtained using smoothing spline with the data shown in figure 1.2 (left). First 80 data of those in figure 1.6(left) and the corresponding $\hat{m}(i)$ (right).

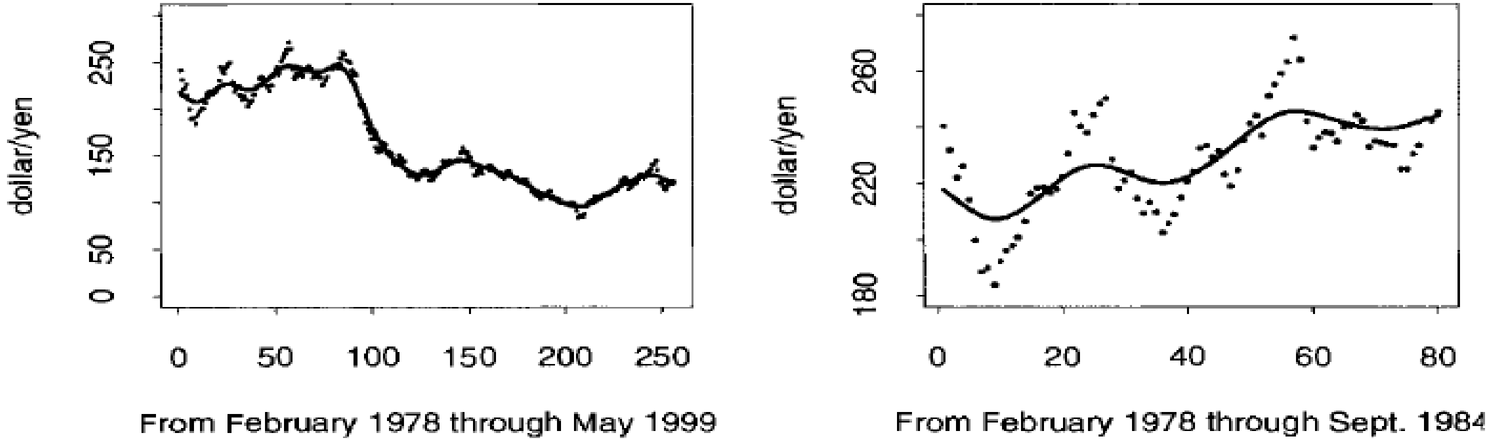
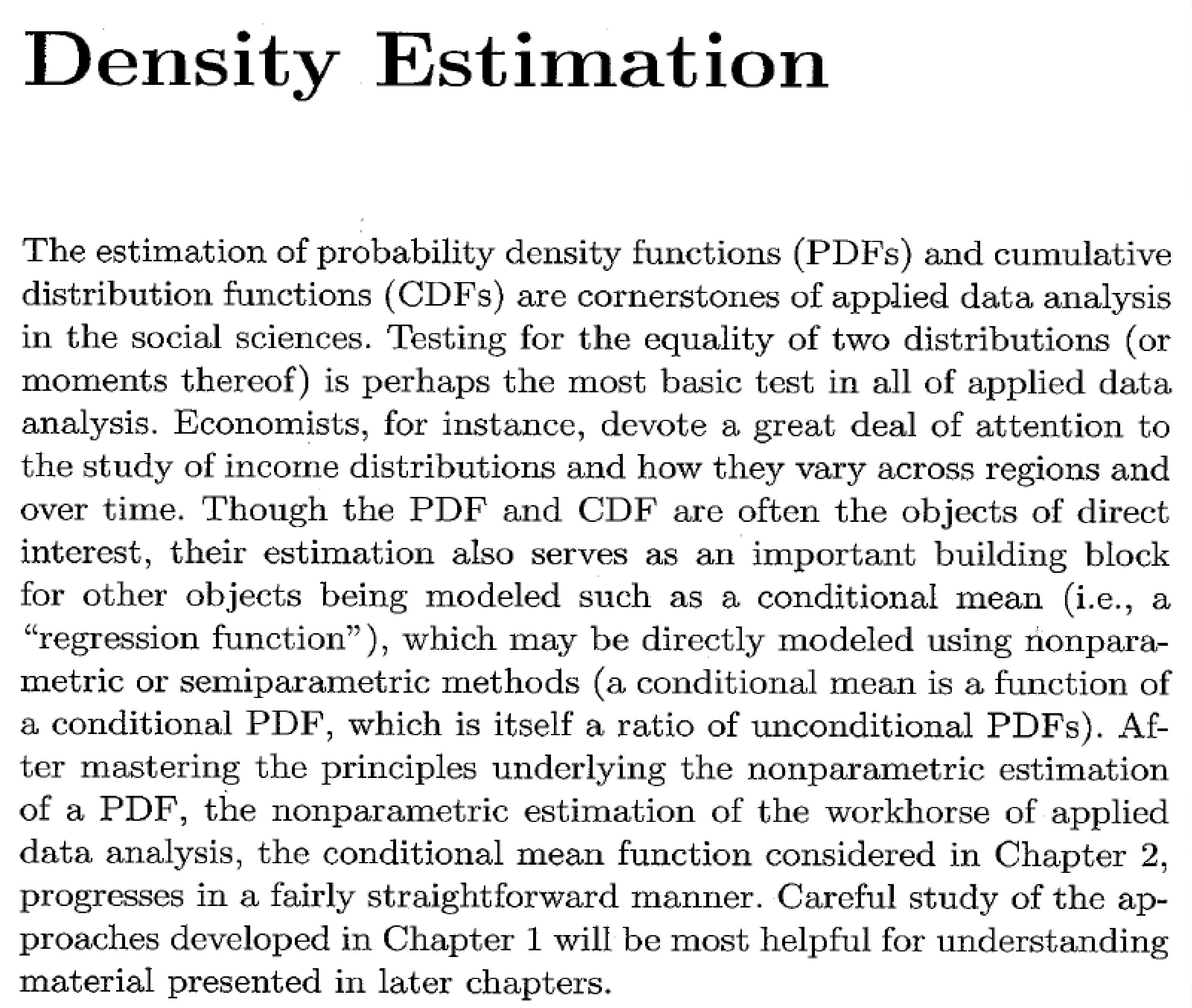
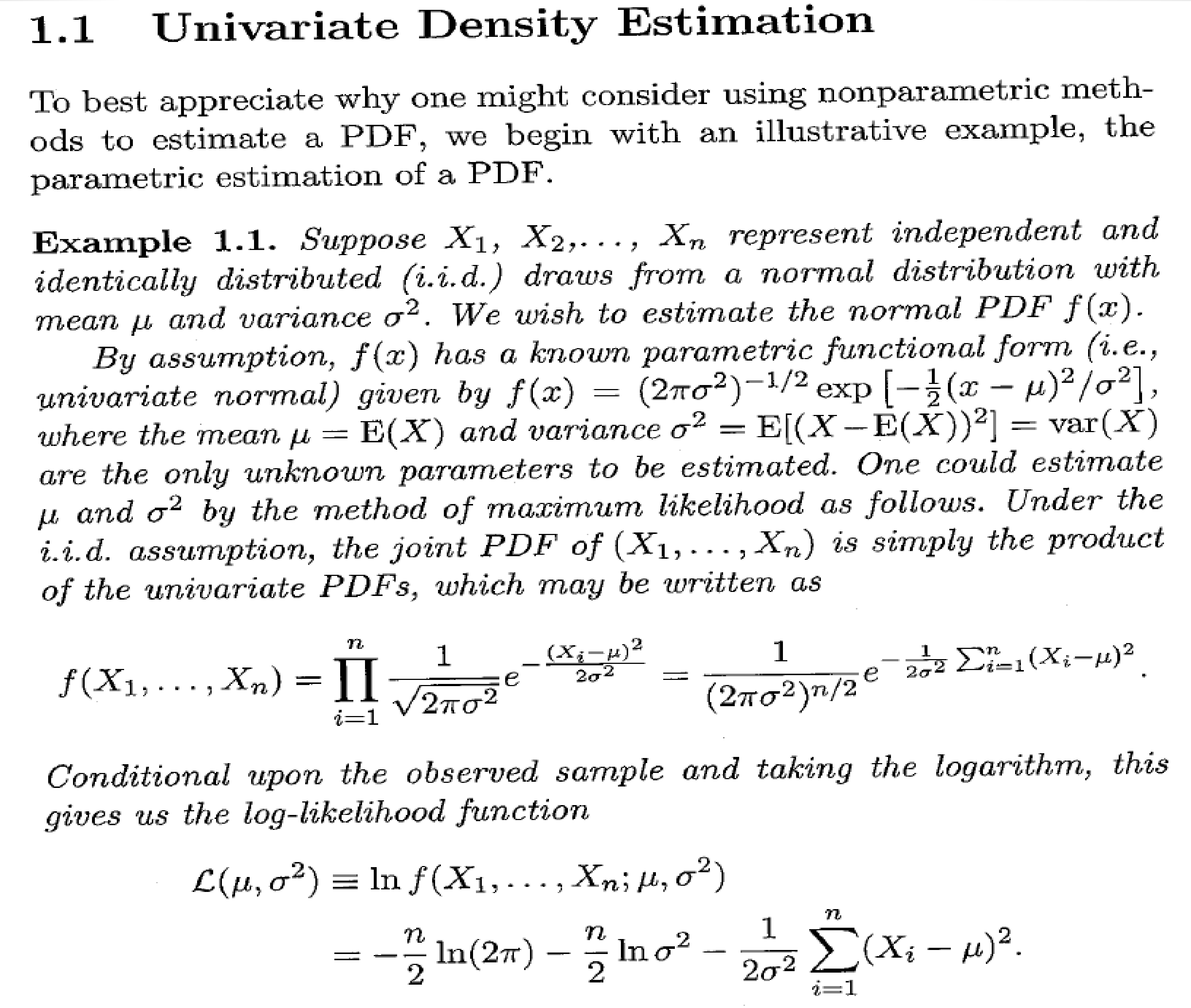


Figure 1.7 Curve obtained using the smoothing spline with the data shown in figure 1.2 (the degree of smoothness is larger than that in figure 1.6(left)), (left). First 80 data of those in figure 1.7(left) and the corresponding $\hat{m}(i)$ (right).



Density Estimation

The estimation of probability density functions (PDFs) and cumulative distribution functions (CDFs) are cornerstones of applied data analysis in the social sciences. Testing for the equality of two distributions (or moments thereof) is perhaps the most basic test in all of applied data analysis. Economists, for instance, devote a great deal of attention to the study of income distributions and how they vary across regions and over time. Though the PDF and CDF are often the objects of direct interest, their estimation also serves as an important building block for other objects being modeled such as a conditional mean (i.e., a "regression function"), which may be directly modeled using nonparametric or semiparametric methods (a conditional mean is a function of a conditional PDF, which is itself a ratio of unconditional PDFs). After mastering the principles underlying the nonparametric estimation of a PDF, the nonparametric estimation of the workhorse of applied data analysis, the conditional mean function considered in Chapter 2, progresses in a fairly straightforward manner. Careful study of the approaches developed in Chapter 1 will be most helpful for understanding material presented in later chapters.



1.1 Univariate Density Estimation

To best appreciate why one might consider using nonparametric methods to estimate a PDF, we begin with an illustrative example, the parametric estimation of a PDF.

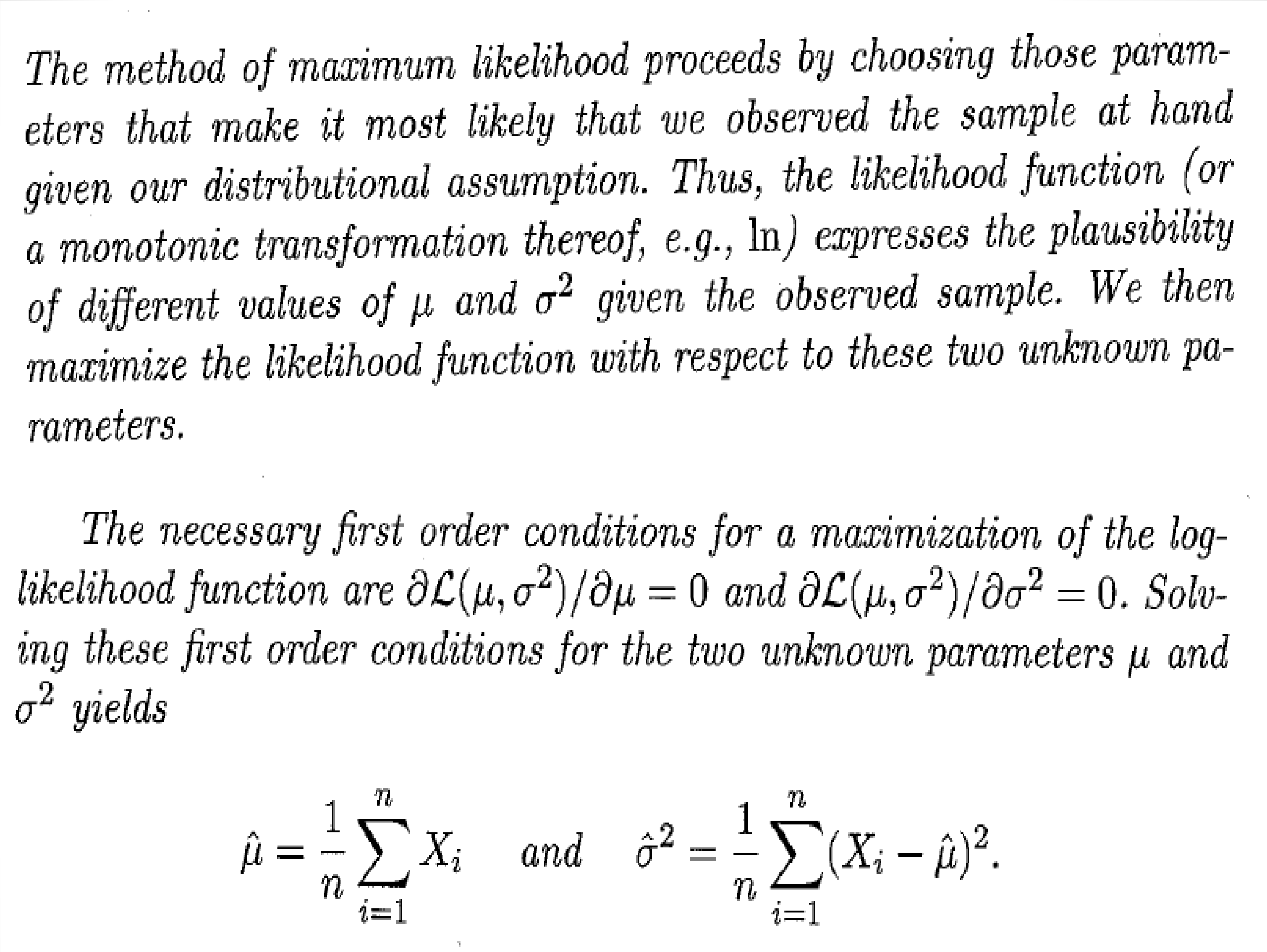
Example 1.1. *Suppose X_1, $X_2, \ldots,$ X_n represent independent and identically distributed (i.i.d.) draws from a normal distribution with mean μ and variance σ^2. We wish to estimate the normal PDF $f(x)$.*

By assumption, $f(x)$ has a known parametric functional form (i.e., univariate normal) given by $f(x) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{1}{2}(x-\mu)^2/\sigma^2\right]$, where the mean $\mu = \mathrm{E}(X)$ and variance $\sigma^2 = \mathrm{E}[(X - \mathrm{E}(X))^2] = \mathrm{var}(X)$ are the only unknown parameters to be estimated. One could estimate μ and σ^2 by the method of maximum likelihood as follows. Under the i.i.d. assumption, the joint PDF of $(X_1, \ldots, X_n)$ is simply the product of the univariate PDFs, which may be written as

$$f(X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}.$$

Conditional upon the observed sample and taking the logarithm, this gives us the log-likelihood function

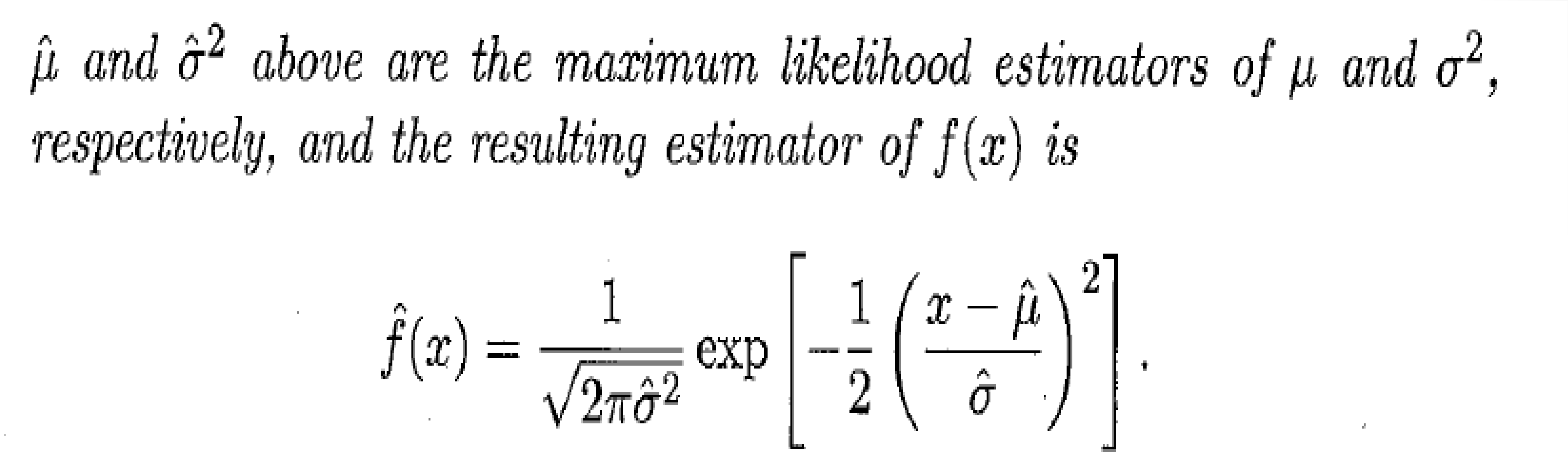
$$\begin{aligned}\mathcal{L}(\mu, \sigma^2) &\equiv \ln f(X_1, \ldots, X_n; \mu, \sigma^2)\\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.\end{aligned}$$



The method of maximum likelihood proceeds by choosing those parameters that make it most likely that we observed the sample at hand given our distributional assumption. Thus, the likelihood function (or a monotonic transformation thereof, e.g., ln*) expresses the plausibility of different values of* μ *and* σ^2 *given the observed sample. We then maximize the likelihood function with respect to these two unknown parameters.*

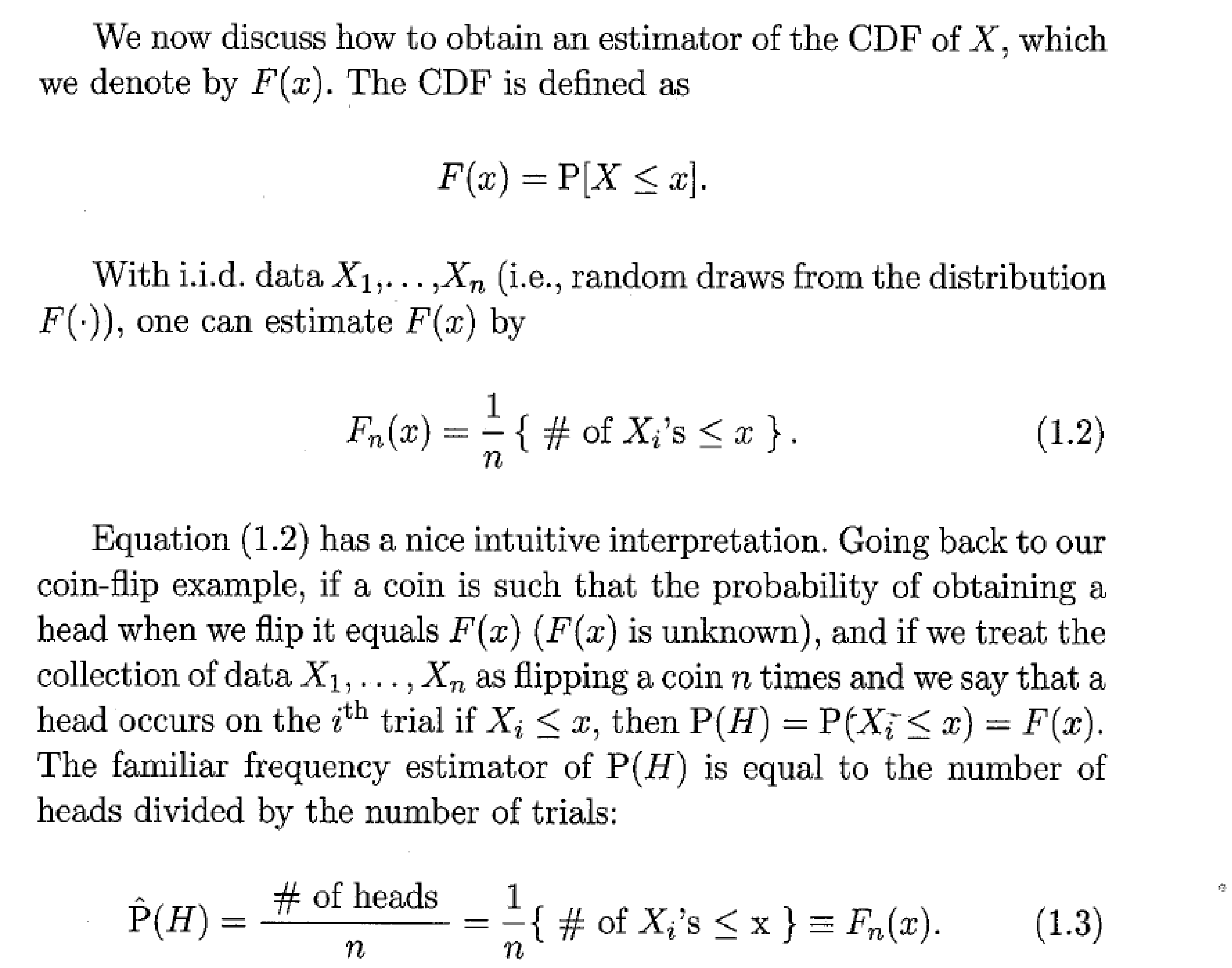
The necessary first order conditions for a maximization of the log-likelihood function are $\partial\mathcal{L}(\mu,\sigma^2)/\partial\mu = 0$ *and* $\partial\mathcal{L}(\mu,\sigma^2)/\partial\sigma^2 = 0$*. Solving these first order conditions for the two unknown parameters* μ *and* σ^2 *yields*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$



$\hat{\mu}$ and $\hat{\sigma}^2$ above are the maximum likelihood estimators of μ and σ^2, respectively, and the resulting estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right].$$



We now discuss how to obtain an estimator of the CDF of X, which we denote by $F(x)$. The CDF is defined as

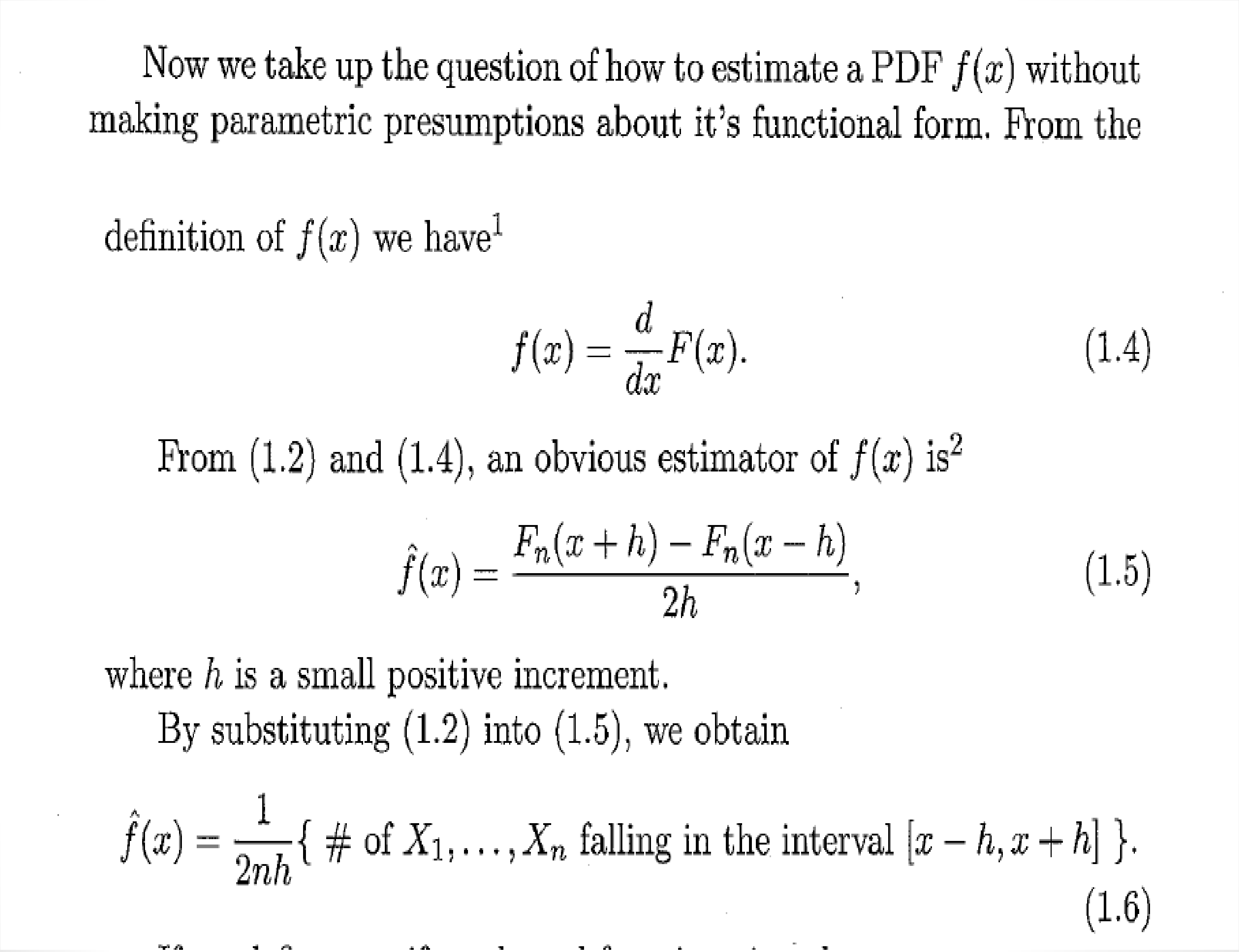
$$F(x) = \mathrm{P}[X \leq x].$$

With i.i.d. data X_1,. . . ,X_n (i.e., random draws from the distribution $F(\cdot)$), one can estimate $F(x)$ by

$$F_n(x) = \frac{1}{n}\{\ \#\text{ of } X_i\text{'s} \leq x\ \}. \tag{1.2}$$

Equation (1.2) has a nice intuitive interpretation. Going back to our coin-flip example, if a coin is such that the probability of obtaining a head when we flip it equals $F(x)$ ($F(x)$ is unknown), and if we treat the collection of data $X_1, \ldots, X_n$ as flipping a coin n times and we say that a head occurs on the i^{th} trial if $X_i \leq x$, then $\mathrm{P}(H) = \mathrm{P}(X_i \leq x) = F(x)$. The familiar frequency estimator of $\mathrm{P}(H)$ is equal to the number of heads divided by the number of trials:

$$\hat{\mathrm{P}}(H) = \frac{\#\text{ of heads}}{n} = \frac{1}{n}\{\ \#\text{ of } X_i\text{'s} \leq \mathrm{x}\ \} \equiv F_n(x). \tag{1.3}$$



Now we take up the question of how to estimate a PDF $f(x)$ without making parametric presumptions about it's functional form. From the definition of $f(x)$ we have[1]

$$f(x) = \frac{d}{dx}F(x). \tag{1.4}$$

From (1.2) and (1.4), an obvious estimator of $f(x)$ is[2]

$$\hat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}, \tag{1.5}$$

where h is a small positive increment.

By substituting (1.2) into (1.5), we obtain

$$\hat{f}(x) = \frac{1}{2nh}\{\text{ \# of } X_1, \ldots, X_n \text{ falling in the interval } [x-h, x+h]\ \}. \tag{1.6}$$

Thanks for Listening