

1

Parametric and Nonparametric Estimation

1. Introduction

This text is concerned with statistical estimation, in particular, with the estimation of probability densities of random variables. Typically, one considers a sample X_1, X_2, \dots, X_n of independent, identically distributed random variables with common density f , and one seeks to estimate f from the data. Both the parametric and the nonparametric settings are treated, with the emphasis on the latter. It is perhaps useful to briefly discuss how statistical estimation is interpreted in this text. In rough outline, statistical estimation deals with the following issues.

- (a) The construction of estimators.
- (b) Quantifying the notion of accuracy of the estimator.
- (c) Deriving bounds on the accuracy for a fixed “amount” n of data, and asymptotically for $n \rightarrow \infty$. The ultimate goal is to obtain the distribution of the accuracy, viewed as a random variable, again for fixed n or asymptotically.
- (d) Quantifying notions of optimality of the estimator. One would like to know the limits on the accuracy of any estimator, and whether the estimators under consideration reach these limits. For parametric estimation, this is well understood, but for nonparametric problems, this is perhaps a bit ambitious.
- (e) Settling the relevance or adequacy of the estimators for finite (small) n , given the model.
- (f) Certifying the adequacy of the indicated model. The latter goes by the name of “goodness-of-fit testing”, which falls outside the scope of this work.
- (g) Finally, the actual computation of the estimators is of concern. Typically, statistical estimators are defined implicitly as solutions to nonlinear equations or to minimization problems. Effective ways to compute and/or approximate them must be determined.

One would classify items (a) through (d) as belonging to estimation *theory*. However, in (d), the “application”, that is, the connection with the “real world”, must be kept in mind. Parts (e) through (g) definitely belong to the realm of estimation *practice*, the application of estimation to actual problems. In (g), it is more or less assumed that the implicitly defined estimators exist and are unique, and that one has a way of resolving the nonuniqueness if the need arises. So estimation *theory* creeps back in.

To illustrate what we have in mind, let f_o be a univariate probability density function (pdf), with corresponding distribution function (cdf) F_o . The simplest *density estimation* problem is where the data

$$(1.1) \quad \begin{array}{l} X_1, X_2, \dots, X_n \text{ are independent, identically distributed (iid),} \\ \text{univariate random variables, with common density } f_o(x), \end{array}$$

and we wish to estimate f_o . Here, n is commonly referred to as the sample size. It is customary to encode the data X_1, X_2, \dots, X_n in the *empirical* distribution function F_n , defined as

$$(1.2) \quad F_n(x) = \frac{\#\{X_i : X_i \leq x\}}{n}, \quad -\infty < x < \infty,$$

the fraction of the observations not exceeding x . What is lost by doing so is the order in which the observations occurred, but assuming iid observations, this is irrelevant, since F_n is a *sufficient statistic* for F_o .

How one goes about estimating f_o is influenced by the availability (or lack thereof) of information on the “model” for f_o . In the *parametric* model, the pdf f_o is assumed to belong to a parametric family

$$(1.3) \quad \mathfrak{F} \stackrel{\text{def}}{=} \{f(\cdot; \theta) : \theta \in \Theta\},$$

described by a (low-dimensional) parameter θ belonging to the set of all possible parameters Θ . Thus, there exists a $\theta_o \in \Theta$ such that

$$(1.4) \quad f_o(x) = f(x; \theta_o), \quad -\infty < x < \infty.$$

It is obligatory to mention the standard example for density estimation, viz. the family of normal densities, parametrized by $\theta = (\mu, \sigma)$,

$$(1.5) \quad f(x; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad -\infty < x < \infty.$$

In all parametric problems, the tacit assumption is that the parameter θ represents a single or perhaps a few (very few) real numbers. Thus, although the problem reduces to that of estimating these few real numbers, it should be kept in mind that the goal is still to estimate the density $f(\cdot; \theta_o)$. For now, we assume that θ represents a single real number.

The standard method for estimating θ_o , dating back all the way to FISHER (1922), is by *maximum likelihood estimation*. In this method,

the estimator (assuming it exists) is any value of θ for which the likelihood is maximal. Under the model (1.1)–(1.3), this amounts to solving

$$(1.6) \quad \begin{aligned} & \text{maximize} && \prod_{i=1}^n f(X_i; \theta) \\ & \text{subject to} && \theta \in \Theta, \end{aligned}$$

but it is more natural to consider the averaged logarithm of the likelihood, written as

$$(1.7) \quad \frac{1}{n} \sum_{i=1}^n \log f(X_i; \theta) = \int_{\mathbb{R}} \log f(x; \theta) dF_n(x).$$

The reason for this is that under reasonable conditions, the log-likelihood converges as $n \rightarrow \infty$

$$\int_{\mathbb{R}} \log f(x; \theta) dF_n(x) \longrightarrow \int_{\mathbb{R}} \log f(x; \theta) dF_o(x), \quad (\theta \text{ fixed})$$

almost surely, say, by the strong law of large numbers. Thus, the maximum likelihood estimation problem is equivalent to

$$(1.8) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} \log f(x; \theta) dF_n(x) \\ & \text{subject to} && \theta \in \Theta, \end{aligned}$$

and we refer to (1.8) as such. The questions of interest are whether (1.8) has a unique solution θ_n , how it may be computed, and what can be said about the error $\theta_n - \theta_o$ or $f(\cdot; \theta_n) - f(\cdot; \theta_o)$. One reason for the popularity of maximum likelihood estimation is that under reasonable conditions,

$$(1.9) \quad \sqrt{n}(\theta_n - \theta_o) \longrightarrow_d Y \sim N(0, \sigma^2),$$

i.e., $\sqrt{n}(\theta_n - \theta_o)$ converges in distribution to a normally distributed random variable Y , with mean 0 and variance σ^2 given via

$$(1.10) \quad \frac{1}{\sigma^2} = \int_{\mathbb{R}} \frac{|g(x; \theta_o)|^2}{f(x; \theta_o)} dx,$$

where $g(x; \theta)$ denotes the partial derivative of $f(x; \theta)$ with respect to θ . (For a precise definition of convergence in distribution, see §3.) Moreover, with some technical caveats, the variance σ^2 in (1.9) is a lower bound for the variance of all *unbiased* estimators, which usually is expressed by saying that (1.10) is the *Cramér-Rao lower bound*.

There are other methods for estimating θ_o , such as the method of moments, the method of maximal spacings, and quantile regression. Mention must be made of least-squares estimation, especially in connection with nonparametric estimation later on (even though the authors have mixed feelings about it). The idea of least-squares estimation is that assuming f_o

were known, an ideal choice of θ would minimize

$$(1.11) \quad \int_{\mathbb{R}} |f(x; \theta) - f_o(x)|^2 dx ,$$

the latter being equal to 0 if the model is correct. Because f_o is unknown, this method cannot be used. However, by rewriting (1.11) as

$$(1.12) \quad \int_{\mathbb{R}} |f(x; \theta)|^2 dx - 2 \int_{\mathbb{R}} f(x; \theta) dF_o(x) + \int_{\mathbb{R}} |f_o(x)|^2 dx ,$$

we see that the first term can be computed exactly, and that the last term is independent of θ . The second term may be estimated by

$$-2 \int_{\mathbb{R}} f(x; \theta) dF_n(x) ,$$

and so the *least-squares estimator* of the parameter θ_o is the solution to

$$(1.13) \quad \begin{array}{ll} \text{minimize} & -2 \int_{\mathbb{R}} f(x; \theta) dF_n(x) + \int_{\mathbb{R}} |f(x; \theta)|^2 dx \\ \text{subject to} & \theta \in \Theta . \end{array}$$

Of course, the questions that concerned us for maximum likelihood estimation apply here as well. In particular, under reasonable conditions, the analogue of (1.9) holds, with σ^2 given by (hold on to your hat)

$$(1.14) \quad \sigma^2 = \frac{\int_{\mathbb{R}} |g(x; \theta_o) - E|^2 f(x; \theta_o) dx}{\left\{ \int_{\mathbb{R}} |g(x; \theta_o)|^2 dx \right\}^2} ,$$

where $g(x; \theta)$ is still the partial derivative of $f(x; \theta)$ with respect to θ , and

$$E = \int_{\mathbb{R}} g(x; \theta_o) f(x; \theta_o) dx .$$

Apart from the fact that this is *ugly* compared with the maximum likelihood case, it is also worse, in that the last σ^2 is larger than the one in (1.10). (It has to be if the Cramér–Rao lower bound deserves its name.)

This concludes the introductory description of parametric estimation. It is treated in much greater detail in Part I.

If for whatever reason a parametric model for f_o is not forthcoming, we are dealing with what is customarily referred to as a *nonparametric* estimation problem. Another way of saying this is that one wishes to make as few assumptions as possible about the density f_o in (1.1). What does this actually mean? Without any assumptions, the density f_o is just a nonnegative, integrable function with integral equal to 1. Thus, it is an infinite-dimensional object, and infinitely many parameters are required to describe f_o . By way of example, an approximate way to describe f_o is by

means of the probabilities $p_i = \mathbb{P}[X_1 \in (ih, (i+1)h)]$,

$$(1.15) \quad p_i = \int_{ih}^{(i+1)h} f_o(x) dx, \quad i = 0, \pm 1, \pm 2, \dots,$$

where h is a small, positive number. It is clear that a finite amount of data, as in (1.1), will not suffice to determine the infinitely many p_i , let alone f_o . Thus, even in the nonparametric setting, some assumptions are needed, with the purpose of reducing the infinite number of parameters to (almost) finitely many. The necessity of some assumptions also surfaces when considering maximum likelihood or least-squares estimation, that is, (1.8) or (1.13). The maximum likelihood problem

$$(1.16) \quad \begin{aligned} &\text{minimize} && -\frac{1}{n} \sum_{i=1}^n \log f(X_i) \\ &\text{subject to} && f \text{ is a continuous pdf} \end{aligned}$$

has no solution. Loosely speaking, the “solution” of (1.16) would be a sum of point masses at the observations

$$(1.17) \quad f^n(x) = \frac{1}{n} \sum_{i=1}^n \delta(x - X_i), \quad -\infty < x < \infty,$$

where $\delta(x)$ is the unit mass at 0, but this is not a density. Note that the distribution corresponding to f^n is F_n , the empirical distribution function.

What can be done about (1.16) not having a solution? One approach is to replace the continuity assumption on f in (1.16) by the requirement that f be constant on the intervals $(ih, (i+1)h]$, $i = 0, \pm 1, \pm 2, \dots$. The maximum likelihood solution for f is then given by a histogram as in (1.15), see THOMPSON and TAPIA (1990), although the histogram estimator has the inconvenience of being discontinuous. An alternative is to “spread out” the point masses a bit to obtain

$$(1.18) \quad f^{nh}(x) = \frac{1}{n} \sum_{i=1}^n A_h(x - X_i), \quad -\infty < x < \infty,$$

where $A_h(x) = h^{-1}A(h^{-1}x)$ for a nice bell-shaped density A , say, with zero mean and finite variance. By choosing A as the uniform density on $[-\frac{1}{2}, \frac{1}{2}]$, the kernel estimator comes close to a histogram. Note that f^{nh} may also be written as

$$(1.19) \quad f^{nh}(x) = A_h * dF_n(x) \stackrel{\text{def}}{=} \int_{\mathbb{R}} A_h(x - y) dF_n(y).$$

We use either notation throughout. The estimator (1.18) is referred to as the kernel estimator with “kernel” A and smoothing parameter h . (The standard symbol for the kernel is K , which we wish to reserve for something else, see § 2 and Volume II.) Kernel estimators date back to AKAIKE (1954), ROSENBLATT (1956), WHITTLE (1958), PARZEN (1962), and WATSON and LEADBETTER (1963).

Figure 1.1. Kernel density estimators for the Buffalo snow fall data (divided by 1.3 times the largest observed annual snowfall), with the Gaussian kernel, and various values of the smoothing parameter h .

Based on the above motivation, one would not guess that kernel estimators can be any good. In fact, they are, *provided* the smoothing parameter is chosen properly, as we show in Part II. A preview is given in Figure 1.1, in which estimators for the density governing the annual snow fall in Buffalo, New York, are given, for various values of h , and with A the standard normal density. (For the data, see Appendix 1.) It is clear that the top two estimators in Figure 1.1 cannot be “right”. For $h = 0.01$, the density seems much too rough (the estimator is undersmoothed), and for $h = 0.1$, the density is much too smooth and all detail is washed out (the estimator is oversmoothed). The two estimators at the bottom of Figure 1.1 seem reasonable, but therein lies the problem. Which one of these two reasonable estimators is closest to the truth? There is obviously a need for *rational* procedures for selecting the smoothing parameter h , that is, the h selected should depend only on the data. We devote two chapters in Part II to this, one on the theory, and the other on the practical performance of smoothing parameter selection procedures.

Returning to the maximum likelihood problem (1.16), we ask the question of how one can incorporate the information that the solution of (1.16) should be a pdf, or a smooth pdf. One way to do this is by the method of sieves, about which more is said in §5. The oldest and the authors’ favorite

method is to consider maximum penalized likelihood estimation, that is, to add a roughness penalization term to the negative log-likelihood. Thus, the problem (1.16) is replaced by

$$(1.20) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} \log f(x) dF_n(x) + h^2 R(f) \\ & \text{subject to} && f \text{ is a continuous density ,} \end{aligned}$$

where R is the roughness penalization term and h is the smoothing parameter, analogous to its use in kernel density estimation. The authors have a soft spot for the choice of I.J. GOOD (1971)

$$(1.21) \quad R(f) = \int_{-\infty}^{\infty} \left| \frac{d}{dx} \sqrt{f(x)} \right|^2 dx ,$$

with $R(f) = +\infty$ when the derivative of \sqrt{f} is not square integrable on the line. This leads to the estimator f , implicitly defined by $\varphi = \sqrt{f}$ with

$$(1.22) \quad \varphi(x) = \frac{1}{n} \sum_{i=1}^n \frac{\mathfrak{B}_h(x - X_i)}{\varphi(X_i)} , \quad x \in \mathbb{R} ,$$

where $\mathfrak{B}_h(x) = (2h)^{-1} \exp(h^{-1}|x|)$ is the scaled, two-sided exponential kernel. This estimator turns out to be a remarkably GOOD estimator, both in theory and practice (provided h is chosen properly).

In view of (1.16)–(1.20), one may likewise define a penalized version of the least-squares problem (1.13) by seeking to

$$(1.23) \quad \begin{aligned} & \text{minimize} && - 2 \int_{\mathbb{R}} f(x) dF_n(x) + \int_{\mathbb{R}} |f(x)|^2 dx + h^2 R(f) \\ & \text{subject to} && f \text{ is a continuous pdf ,} \end{aligned}$$

with $R(f)$ the roughness penalty and h the smoothing parameter, as before. Now, the choice

$$(1.24) \quad R(f) = \int_{\mathbb{R}} |f'(x)|^2 dx ,$$

with $R(f) = +\infty$ when f' is not square integrable, is intriguing. With this penalization, the solution f of (1.23) satisfies the boundary value problem

$$(1.25) \quad \begin{aligned} & -h^2 f'' + f = dF_n(x) , \quad -\infty < x < \infty , \\ & f(x) \longrightarrow 0 \quad \text{for } |x| \rightarrow \infty , \end{aligned}$$

and is given by the pdf

$$(1.26) \quad f(x) = \mathfrak{B}_h * dF_n(x) , \quad -\infty < x < \infty .$$

See Exercise (11.3.25). In effect, \mathfrak{B}_h is the *Green's function* for the boundary value problem (1.25), see, e.g., COURANT and HILBERT (1953). Thus, in this rather natural way, we get a particular kernel estimator, and it

seems but a small leap of faith to assume that one can get other kernel estimators by appropriate penalization in (1.23). So, it seems natural to view kernel estimators as least-squares estimators. There is a way to get kernel estimators via maximum likelihood, viz. as the solution to the maximum *smoothed* likelihood problem

$$(1.27) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} [A_h * \log f](x) dF_n(x) \\ & \text{subject to} && f \text{ is a pdf ,} \end{aligned}$$

but at first glance, this seems less natural. That it is in fact very natural is explained in detail in the next section.

Why are we so concerned with interpreting kernel density estimators as solutions to minimization problems? The question is simply this. While kernel estimators are without equal for the standard density estimation problem (1.1), how does one generalize them to more complicated settings? One important example is that of indirect estimation problems, discussed in § 2 and in Volume II. Another example concerns the estimation of densities with an *a priori* known shape. Some standard shape restrictions are monotonicity and/or convexity of a density on $(0, \infty)$, and unimodality or log-concavity of densities on $(-\infty, \infty)$, or that the tails of f_o may be monotone or convex. By way of example, to estimate a smooth, unimodal density, one may consider the solution of the problem

$$(1.28) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} [A_h * \log f](x) dF_n(x) \\ & \text{subject to} && f \text{ is a unimodal density .} \end{aligned}$$

In Chapter 6, we study this problem in great detail.

In the remainder of this introductory chapter, we continue to set the tone of the text by elaborating on the topics mentioned above and discussing to what extent they are to be or not to be covered.

2. Indirect problems, EM algorithms, kernel density estimation, and roughness penalization

Here, we discuss the simplest of indirect estimation problems. It is a bit of an advertisement for Volume II, but it also sheds an unexpected light on kernel density estimation and GOOD's roughness penalization.

Let Y_1, Y_2, \dots, Y_n be iid univariate random variables with common pdf f_o . The goal is still to estimate f_o , but now we are interested in the situation in which the Y_i are contaminated by iid noise Z_1, Z_2, \dots, Z_n , independent of the Y_i . Thus, the (iid) observations are

$$(2.1) \quad X_i = Y_i + Z_i, \quad i = 1, 2, \dots, n.$$

Assuming that the common distribution K of the Z_i is known, the common distribution of the X_i has density $g = \mathcal{K}f_o$, where for any pdf φ ,

$$(2.2) \quad \mathcal{K}\varphi(x) = \varphi * dK(x) = \int_{\mathbb{R}} \varphi(x-z) dK(z), \quad -\infty < x < \infty.$$

If K has a density k , then this may be rewritten as $\mathcal{K}\varphi = k * \varphi$, the convolution of k and φ ,

$$(2.3) \quad k * \varphi(x) = \int_{\mathbb{R}} k(x-y) \varphi(y) dy, \quad -\infty < x < \infty.$$

Thus, the *nonparametric deconvolution problem* is to estimate f_o based on the iid observations X_1, X_2, \dots, X_n with common density $\mathcal{K}f_o$.

We are interested in maximum penalized likelihood estimation for the deconvolution problem, but the approach is driven by algorithmic considerations. With G_n denoting the empirical distribution function of X_1, \dots, X_n in (2.1), the maximum likelihood estimator of f_o would be any solution of

$$(2.4) \quad \begin{aligned} &\text{minimize} && - \int_{\mathbb{R}} \log \mathcal{K}f(x) dG_n(x) \\ &\text{subject to} && f \text{ is a pdf,} \end{aligned}$$

but, of course, it is not so obvious that solutions exist. Be that as it may, one may compute reasonable estimators by the following algorithm, essentially due to SHEPP and VARDI (1982). Let f_1 be an initial guess for the pdf f_o , and define for $q \geq 1$,

$$(2.5) \quad f_{q+1}(y) = f_q(y) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(X_i - y)}{k * f_q(X_i)}, \quad y \in \mathbb{R}.$$

This is an example of an EM algorithm, see (2.4.26) and Volume II, and as such has many wonderful properties, e.g., the sequence $\{f_q\}_{q \geq 1}$ converges to a solution of (2.4) provided one exists. Unfortunately, in the present case, the solution does *not* exist, and indeed, practical computations based on this algorithm show that the f_q at first get better, and then get less and less smooth as the iteration progresses. A cure was provided by SILVERMAN, JONES, WILSON, and NYCHKA (1990), who figured that if smoothness is the problem, one could solve it by adding a smoothing step to the EM algorithm. Thus, for A a smooth symmetric pdf, let $A_h(x) = h^{-1}A(h^{-1}x)$, and consider the EMS algorithm

$$(2.6) \quad \begin{aligned} \varphi_q(y) &= f_q(y) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(X_i - y)}{k * f_q(X_i)}, \quad y \in \mathbb{R}, \\ f_{q+1} &= A_h * \varphi_q, \end{aligned}$$

that is, one step of the EM algorithm, followed by a smoothing step [S]. This works quite well in practice, but the question is what is being computed. Also, the convergence of this algorithm is not clear. Order is

restored if we implement the smoothing somewhat differently. Define a nonlinear smoothing operator associated with A_h by

$$(2.7) \quad \mathcal{N}f(x) = \exp([A_h * \{\log f\}](x)) , \quad x \in \mathbb{R} ,$$

and consider the NEMS algorithm

$$(2.8) \quad \begin{aligned} \psi_q &= \mathcal{N}f_q , \\ \varphi_q(y) &= \psi_q(y) \cdot \frac{1}{n} \sum_{i=1}^n \frac{k(X_i - y)}{k * \psi_q(X_i)} , \quad y \in \mathbb{R} , \\ f_{q+1} &= A_h * \varphi_q , \end{aligned}$$

that is, one nonlinear smoothing step [N], followed by one EM step, followed by another linear smoothing step [S]. The crucial feature is still that the f_q are smooth, due to the last smoothing step. In fact, from a practical point of view, there does not appear to be much of a difference between the NEMS and EMS algorithms. The theoretical advantage of the NEMS algorithm is that it is an EM algorithm for the maximum smoothed likelihood problem

$$(2.9) \quad \begin{aligned} \text{minimize} \quad & L_{nh}(f) \stackrel{\text{def}}{=} - \int_{\mathbb{R}} \log \mathcal{K} \mathcal{N}f(x) dG_n(x) \\ \text{subject to} \quad & f \text{ is a pdf} , \end{aligned}$$

which has a unique, continuous solution f^{nh} , and the NEMS iterates converge to it in a suitable sense.

What does all of this have to do with kernel density estimation? Let us backtrack, and assume that the $Z_i = 0$ for all i . Then, G_n has the same meaning as F_n in (1.2), with X_1, X_2, \dots, X_n iid random variables with distribution function F_o . Moreover, the operator \mathcal{K} in (2.2) is just the identity operator, i.e., $\mathcal{K}\varphi = \varphi$ for all φ , and the objective function in (2.9) reads as

$$(2.10) \quad - \int_{\mathbb{R}} \log \mathcal{N}f(x) dG_n(x) = - \int_{\mathbb{R}} [A_h * \log f](x) dG_n(x) .$$

This is the plain density estimation problem, of course. Thus, the maximum smoothed likelihood density estimation problem is

$$(2.11) \quad \begin{aligned} \text{minimize} \quad & \Lambda_{nh}(f) \stackrel{\text{def}}{=} - \int_{\mathbb{R}} [A_h * \log f](x) dG_n(x) \\ \text{subject to} \quad & f \text{ is a pdf} . \end{aligned}$$

We now show that the solution of (2.11), respectively (1.27), is given by the kernel density estimator $f = A_h * dG_n$, provided A is symmetric. First, rewrite $\Lambda_{nh}(f)$ as a repeated integral, and interchange the order of inte-

gration so

$$\begin{aligned}
 \Lambda_{nh}(f) &= - \int_{\mathbb{R}} \int_{\mathbb{R}} A_h(x-y) \log f(y) dy dG_n(x) \\
 &= - \int_{\mathbb{R}} \left(\int_{\mathbb{R}} A_h(x-y) dG_n(x) \right) \log f(y) dy \\
 (2.12) \quad &= - \int_{\mathbb{R}} [A_h * dG_n](y) \log f(y) dy , \\
 &= - \int_{\mathbb{R}} f^{nh}(y) \log f(y) dy ,
 \end{aligned}$$

where in the next to last line we used the fact that A is symmetric, so that $A_h(x-y) = A_h(y-x)$ for all x, y . For reasons that may not (yet) be entirely clear, it is useful to introduce the Kullback-Leibler distance (or divergence) between two nonnegative integrable functions φ and ψ

$$(2.13) \quad \text{KL}(\varphi, \psi) = \int_{\mathbb{R}} \left\{ \varphi(y) \log \frac{\varphi(y)}{\psi(y)} + \psi(y) - \varphi(y) \right\} dy .$$

Note that for any $x \geq 0, y > 0$, and $t = x/y$,

$$(2.14) \quad x \log(x/y) + y - x = y (t \log t + 1 - t) \geq 0 ,$$

with equality if and only if $x = y$. Thus, the integrand in (2.13) is nonnegative, and $\text{KL}(\varphi, \psi)$ is well defined if we admit the value $+\infty$.

Let $f^{nh} = A_h * dG_n$, and consider $\Lambda_{nh}(f) - \Lambda_{nh}(f^{nh})$. Then, from (2.12),

$$\Lambda_{nh}(f) - \Lambda_{nh}(f^{nh}) = \int_{\mathbb{R}} f^{nh}(x) \log \frac{f^{nh}(x)}{f(x)} dx .$$

Since f and f^{nh} are pdfs, we may add $f(x) - f^{nh}(x)$ to the integrand without changing the value of the integral, so that

$$(2.15) \quad \Lambda_{nh}(f) - \Lambda_{nh}(f^{nh}) = \text{KL}(f^{nh}, f) \geq 0 ,$$

with equality if and only if $f = f^{nh}$. Thus, f^{nh} is the minimum of $\Lambda_{nh}(f)$ over all pdfs f .

This is a long road to get kernel estimators out of maximum smoothed likelihood estimation, but the authors think it is rather telling. Somewhat tongue-in-cheek one might say that kernel estimators are so good precisely because they are maximum (smoothed) likelihood estimators!

We return to the general deconvolution problem. Just as the GOOD maximum penalized likelihood estimator, see (1.22), performs quite well for plain density estimation, so does the maximum smoothed likelihood estimator for the deconvolution problem. As a matter of fact, from the point of view of L^1 error in the estimators, the NEMS estimator significantly outperforms the Fourier deconvolution kernel estimators. These kernel estimators may be thought of as penalized least-squares estimators, i.e., the

solutions to problems of the form

$$(2.16) \quad \begin{aligned} & \text{minimize} \quad \int_{\mathbb{R}} |\mathcal{K} f(x)|^2 dx - 2 \int_{\mathbb{R}} \mathcal{K} f(x) dG_n(x) + h^2 R(f) \\ & \text{subject to} \quad f \text{ is a pdf ,} \end{aligned}$$

with suitable roughness penalization functionals R and smoothing parameter h . This is analogous to the situation for plain density estimation (1.23). All of this is the main topic of Volume II. For those who cannot wait, see, e.g., STEFANSKI and CARROLL (1990), DEY, RUYMGAART and MAIR (1996) and GOLDENSHLUGER (2000) for the Fourier (L^2) estimators, and EGGERMONT and LARICCIA (1997) for more references and for some experimental comparisons.

We finish this section by relating the NEMS algorithm and its associated maximum *smoothed* likelihood problem (2.9) for the nonparametric deconvolution problem to maximum *penalized* likelihood estimation, using GOOD's roughness penalization. Suppose for the sake of argument that we replaced the negative log-likelihood $L_{nh}(f)$ for the deconvolution problem by a smoothed version

$$(2.17) \quad \widetilde{L}_{nh}(f) \stackrel{\text{def}}{=} - \int_{\mathbb{R}} A_h * dG_n(x) \log \mathcal{KN} f(x) dx .$$

Surely this would not make much of a difference. Since f and $A_h * dG_n$ are pdfs, this new negative log-likelihood may be written as

$$(2.18) \quad \widetilde{L}_{nh}(f) = \text{KL}(A_h * dG_n, \mathcal{KN} f) + R_h(f) ,$$

with

$$(2.19) \quad R_h(f) = \int_{\mathbb{R}} f(x) - \mathcal{N} f(x) dx .$$

Thus, we may view the smoothed maximum likelihood problem (2.9) as a penalized version of the original problem (2.4), with penalization $R_h(f)$. Now what kind of penalization is involved here? We determine the behavior of $R_h(f)$ for $A_h = \mathfrak{B}_h$, see (1.22). Formally, using the Green's function property of \mathfrak{B} ,

$$\mathfrak{B}_h * \log f - \log f = h^2 (\mathfrak{B}_h * \log f)'' = h^2 \mathfrak{B}_h * (\log f)'' \approx h^2 (\log f)'' .$$

It follows that for strictly positive f (dropping the argument y everywhere),

$$\begin{aligned} f - \mathcal{N} f &= f (1 - \exp(\mathfrak{B}_h * \log f - \log f)) \\ &= f (1 - \exp(h^2 \mathfrak{B}_h * (\log f)'')) \\ &\approx -h^2 f (\mathfrak{B}_h * (\log f)'') \approx -h^2 f (\log f)'' . \end{aligned}$$

Then, after integration by parts,

$$R_h(f) \approx -h^2 \int_{\mathbb{R}} f (\log f)'' = h^2 \int_{\mathbb{R}} f' (\log f)' ,$$

and thus,

$$(2.20) \quad R_h(f) \approx h^2 \int_{\mathbb{R}} \frac{|f'(y)|^2}{f(y)} dy .$$

Admittedly, this derivation is a bit suspect, but it is surprising that the GOOD penalization (1.21) pops up this way. A similar derivation leading to GOOD's roughness penalization may be done for any other reasonable smoother A .

3. Consistency of nonparametric estimators

One of the main themes of this text is the concern with consistency of nonparametric estimators and with convergence rates. Let us consider the kernel estimator f^{nh} of (1.11) for the nonparametric density estimation problem. The estimator f^{nh} is *consistent* if $f^{nh} \rightarrow f_o$ in a suitable sense, for $n \rightarrow \infty$, and $h = h(n)$ properly chosen. A more precise way of saying that $f^{nh} \rightarrow f_o$ is that

$$(3.1) \quad \text{dist}(f^{nh}, f_o) \rightarrow 0$$

for a suitable “distance” of f^{nh} to f_o . Of course, consistency being established, one would like to know how fast $\text{dist}(f^{nh}, f_o)$ tends to 0. Similar observations apply to parametric estimation. Regardless, one aspect of consistency is the choice of the distance function. Another one is the mode of convergence in (3.1). We elaborate this point first.

Let $\{a_n\}_n$ be a sequence of real-valued functions of the random variables X_1, X_2, \dots, X_n , denoted as

$$a_n = a_n(X_1, X_2, \dots, X_n) .$$

There are many ways to define the convergence of a_n , of which we need the following four.

(3.2) DEFINITION. We say that $\{a_n\}_n$ converges

- (a) in expectation to 0 if $\lim_{n \rightarrow \infty} \mathbb{E}[|a_n|] = 0$;
- (b) in probability to 0 if $\lim_{n \rightarrow \infty} \mathbb{P}[|a_n| > \varepsilon] = 0$ for all $\varepsilon > 0$;
- (c) almost surely to 0 if $\mathbb{P}[\lim_{n \rightarrow \infty} |a_n| = 0] = 1$;
- (d) in distribution to a random variable Y at rate b_n if

$$\lim_{n \rightarrow \infty} \mathbb{P}[a_n/b_n \leq y] = \mathbb{P}[Y \leq y]$$

at all continuity points y of the distribution of Y .

The notations for these four notions of convergence are

$$(3.3) \quad a_n \xrightarrow{\mathbb{E}} 0, \quad a_n \xrightarrow{\text{p}} 0, \quad a_n \xrightarrow{\text{as}} 0, \quad a_n/b_n \xrightarrow{\text{d}} Y .$$

It is also useful to have the big Oh and little oh notation available.

(3.4) DEFINITION. We say that

- (a) $a_n =_{\text{as}} \mathcal{O}(b_n)$ if $\limsup_{n \rightarrow \infty} \left| \frac{a_n}{b_n} \right| < \infty$ almost surely;
 (b) $a_n =_{\text{as}} o(b_n)$ if $\limsup_{n \rightarrow \infty} \left| \frac{a_n}{b_n} \right| = 0$ almost surely.

There are in probability versions of these, denoted as

$$(3.5) \quad a_n = \mathcal{O}_P(b_n) \quad \text{and} \quad a_n = o_P(b_n) .$$

At times, it is useful to describe the exact rate of convergence.

(3.6) DEFINITION. We say that $a_n \asymp_{\text{as}} b_n$ if

$$0 < \liminf_{n \rightarrow \infty} \frac{a_n}{b_n} \leq \limsup_{n \rightarrow \infty} \frac{a_n}{b_n} < \infty , \quad \text{almost surely} .$$

If both a_n and b_n are deterministic, we write $a_n \asymp b_n$.

We now consider the choice of the distance function in (3.1), beginning with the parametric case. Let $d \geq 1$ be a fixed integer. We assume that $\theta \in \mathbb{R}^d$, i.e., $\theta = (\theta_1, \theta_2, \dots, \theta_d)$, with the θ_i denoting scalar variables. Of the various ways to measure the distance between an estimator θ^n and the true θ_o , the Euclidean distance $\|\theta^n - \theta_o\|_2$ seems to be preferable. Here, $\|\cdot\|_2$ is the case $p = 2$ of

$$(3.7) \quad \|\theta\|_p = \left\{ \sum_{j=1}^d |\theta_j|^p \right\}^{1/p} , \quad 1 \leq p < \infty .$$

These distance measures are referred to as the ℓ^p norms on \mathbb{R}^d . However, convergence in the sense of $\|\theta^n - \theta_o\|_2$ converging in one of the previously discussed interpretations is equivalent to the same type of convergence in the other norms. This is a finite-dimensional phenomenon.

In the nonparametric case, things are decidedly more complicated, due to the infinite-dimensional setting. The question is what is the best way to measure distances between two pdfs φ and ψ . The primary choice is any one of the L^p norms $\|\varphi - \psi\|_p$, where for measurable functions φ ,

$$(3.8) \quad \|\varphi\|_p = \left\{ \int_{\mathbb{R}} |\varphi(x)|^p dx \right\}^{1/p} , \quad 1 \leq p < \infty ,$$

and for $p = \infty$,

$$(3.9) \quad \|\varphi\|_{\infty} = \inf \{ K : |\varphi(x)| \leq K \text{ almost everywhere} \} .$$

If φ is continuous, then $\|\varphi\|_{\infty} = \sup \{ |\varphi(x)| : x \in \mathbb{R} \}$. The set of all functions with $\|f\|_p < \infty$ is denoted by $L^p(\mathbb{R})$. The cases of (most) interest are $p = 1, 2$ and ∞ . The L^2 norm enjoys the most popularity in the

literature, but the authors believe that its use is a conceptual mistake and, following DEVROYE and GYÖRFI (1985), prefer the L^1 norm for the reasons outlined below. See also RACHEV (1991). In this text, the L^∞ norm is used in one significant instance when discussing unimodal density estimation (§ 6.6), but only to obtain L^1 error bounds.

Scaling invariance. One of the features of the L^1 norm is that it is the only L^p norm which is *scaling invariant*. Let $\sigma > 0$ and consider the scaled densities

$$(3.10) \quad \varphi_\sigma(x) = \sigma^{-1} \varphi(\sigma^{-1}x), \quad -\infty < x < \infty,$$

and likewise for ψ_σ . Then, it is an easy exercise to show that for $1 \leq p \leq \infty$,

$$(3.11) \quad \|\varphi_\sigma - \psi_\sigma\|_p = \sigma^{1-1/p} \|\varphi - \psi\|_p,$$

provided we take $1/p = 0$ for $p = \infty$. Thus, $p = 1$ is the only invariant case.

(3.12) EXERCISE. Verify (3.11).

Thus, if for the Buffalo snow fall data set we decide to measure snowfall in centimeters rather than inches, or in fractions of 1.3 times the largest observed annual snowfall as in Figure 1.1, and go on to estimate the pdf of the snowfall accordingly, then the L^1 distance to the true pdf does not change. The pdfs themselves do change, of course. The above scaling invariance extends to invariance under any monotone (differentiable) transformation of the random variable, see Exercise (4.1.44).

Estimating probabilities. Another feature of the L^1 distance between two pdfs is its close relationship to the *total variation* distance between two distributions. If Φ and Ψ are distributions, then the total variation distance between them is

$$(3.13) \quad \text{TV}(\Phi, \Psi) = \sup_B |\Phi(B) - \Psi(B)|,$$

where the supremum is over all (measurable) events B and for any distribution F ,

$$F(B) = \int_B dF(x)$$

is the probability assigned by F to the event Ω . Thus, the total variation distance measures the largest possible difference in the probabilities assigned to events. The relevance of the total variation distance to statistics is obvious.

It is a standard exercise to show that if Φ and Ψ have densities φ and ψ , then

$$(3.14) \quad \text{TV}(\Phi, \Psi) = 2 \|\varphi - \psi\|_1.$$

Thus, the L^1 errors give us rather sharp information on differences in probabilities.

(3.15) EXERCISE. Verify (3.14).

Cauchy sequences. A final technical observation is that when we have a sequence of densities $\{\varphi_n\}_{n \geq 1}$ which is a Cauchy sequence in L^1 , i.e.,

$$(3.16) \quad \|\varphi_n - \varphi_m\|_1 \longrightarrow 0 \quad \text{for } n, m \rightarrow \infty,$$

then the sequence has a limit, which is again a pdf. This is a useful feature when considering the existence of solutions to the various maximum penalized or smoothed likelihood estimation problems, because often one can construct Cauchy sequences of what one hopes are approximations to a solution. In contrast, if the sequence of densities $\{\varphi_n\}_n$ is a Cauchy sequence in L^2 , then the limit need not be a pdf.

(3.17) EXERCISE. Let $\varphi_n(x) = n^{-1}\psi(n^{-1}x)$ for some bounded pdf ψ . Show that the φ_n are pdfs and that $\{\varphi_n\}_n$ is a Cauchy sequence in L^2 , but that its limit is the zero function.

Of course, the L^1 norm has *some* drawbacks.

Measuring tail behavior. One of the drawbacks of the L^1 distance is that differences in the tails of the pdfs are largely ignored. If tail estimation is (more) important, then one could use the Kullback-Leibler, Hellinger, or Pearson's φ^2 distances, defined, respectively, as

$$(3.18) \quad \text{KL}(\varphi, \psi) = \int_{\mathbb{R}} \left\{ \varphi(x) \log \frac{\varphi(x)}{\psi(x)} + \psi(x) - \varphi(x) \right\} dx,$$

$$(3.19) \quad \text{H}(\varphi, \psi) = \int_{\mathbb{R}} \left| \sqrt{\varphi(x)} - \sqrt{\psi(x)} \right|^2 dx,$$

$$(3.20) \quad \text{PHI}(\varphi, \psi) = \int_{\mathbb{R}} \frac{|\varphi(x) - \psi(x)|^2}{\psi(x)} dx.$$

The following inequalities between these distances hold, for all (sub)pdfs φ , and ψ (i.e., they are nonnegative, with integral ≤ 1),

$$(3.21) \quad \frac{1}{4} \|\varphi - \psi\|_1^2 \leq \text{H}(\varphi, \psi) \leq \text{KL}(\varphi, \psi) \leq \text{PHI}(\varphi, \psi),$$

the proofs of which are delayed until Chapter 10. (Some of the individual inequalities may be improved on.) Note that $\text{H}(\varphi, \psi) \leq \|\varphi - \psi\|_1$, so that the Hellinger distance between pdfs is always bounded. This is not true for the Kullback-Leibler and Pearson's φ^2 distances.

Some comments on these distances are in order. First, only the Hellinger distance is symmetric. We already encountered the Kullback-Leibler distance in §2. The Hellinger distance is quite reasonable since square-root

densities are square integrable. Apart from this, it is also useful for theoretical reasons, such as for establishing minimal conditions for the consistency of estimators, see LE CAM (1970), or IBRAGIMOV and HAS'MINSKII (1981), but we shall not address this. We extensively use Hellinger distances when discussing the GOOD estimator (1.22) in § 5.2.

It is a nice exercise to show that these new distances are also scaling invariant.

(3.22) EXERCISE. Show that KL, H, and PHI are scaling invariant.

Ease of mathematical handling. Another drawback of the L^1 error is that it is not so easy to work with. By way of example, an analogue for the L^1 error of (1.11)–(1.12) does not exist, and that will cost us dearly. In fact, this ease of mathematical handling of the integrated squared error is an important factor in its popularity.

At this point, the question arises whether the choice of distance really matters. The answer turns out to be yes and shows a major difference between parametric and nonparametric problems: For parametric problems, all reasonable ways of measuring the difference between an estimator and the “true” parameter are equivalent (at least asymptotically), but for nonparametric problems, this is not the case. The following example and exercises give some details regarding the difference between convergence in L^∞ -norm and L^1 -norm.

(3.23) EXAMPLE. Let $\{U_j\}_j$ be a sequence of iid random variables with uniform (0,1) distribution. Define a sequence of random functions $\{Y_j\}_j$ on $(-1, 2)$ by

$$(3.24) \quad Y_j(x) = \begin{cases} j, & |x - U_j| \leq j^{-2}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $x \in (-1, 2)$ be fixed. We first show that $Y_j(x) \xrightarrow{\text{as}} 0$. If $x < 0$ or $x > 1$, this is obvious. For $0 < x < 1$, we use that the almost sure convergence is equivalent to

$$(3.25) \quad \lim_{m \rightarrow \infty} \mathbb{P}[Y_j(x) < \varepsilon \text{ for all } j \geq m] = 1 \quad \text{for all } \varepsilon > 0.$$

Since the events $Y_j(x) < \varepsilon$ are independent, we have for all $\varepsilon > 0$,

$$\mathbb{P}[Y_j(x) < \varepsilon \text{ for all } j \geq m] = \prod_{j=m}^{\infty} \mathbb{P}[Y_j(x) < \varepsilon].$$

Now, for all j large enough to avoid edge effects, $\mathbb{P}[Y_j(x) < \varepsilon] = 1 - (2/j^2)$, and $\prod_{j=1}^{\infty} \{1 - (2/j^2)\}$ is a convergent infinite product, so

$$\lim_{m \rightarrow \infty} \mathbb{P}[Y_j(x) < \varepsilon \text{ for all } i \geq m] = \lim_{m \rightarrow \infty} \prod_{j=m}^{\infty} \{1 - (2/j^2)\} = 1.$$

So (3.25) holds, and $Y_j(x) \xrightarrow{\text{as}} 0$. The same holds for $x = 0$ and $x = 1$. On the other hand,

$$\|Y_j\|_1 = 2j^{-1} \xrightarrow{\text{as}} 0, \quad \|Y_j\|_\infty = j \xrightarrow{\text{as}} \infty.$$

Summarizing, the sequence $\{Y_j\}_j$ converges a.s. to 0 in L^1 -norm, converges a.s. to ∞ in L^∞ -norm, and converges a.s. pointwise to 0.

(3.26) EXERCISE. (a) Replace (3.24) by

$$(3.24') \quad Y_j(x) = \sqrt{j}, \quad |x - U_j| \leq j^{-1},$$

and repeat the above example. (In particular, for fixed x , the sequence $\{Y_j(x)\}_j$ does not converge a.s. to 0.)

(b) For any (random) function Y on $[a, b]$, show that $\|Y\|_1 \leq (b-a) \|Y\|_\infty$.

(c) Construct a sequence of random functions $\{Y_j\}_j$ on \mathbb{R} such that

$$\|Y_j\|_1 \xrightarrow{\text{as}} \infty, \quad \|Y_j\|_\infty \xrightarrow{\text{as}} 0.$$

(3.27) EXERCISE. Construct similar examples to illustrate the difference between L^1 and L^2 .

Finally, we note that in Example (3.23), we actually proved a special case of the Borel-Cantelli lemma. Because later on we have occasion to refer to it, we state a version of it here.

(3.28) BOREL-CANTELLI LEMMA. *Let $\{Y_n\}_{n \geq 1}$ be a sequence of random variables, and let $c > 0$. Then, the following statements hold.*

(a) $\sum_{n=1}^{\infty} \mathbb{P}[|Y_n| \geq c] < \infty$ implies $\limsup_{n \rightarrow \infty} |Y_n| \leq_{\text{as}} c$;

(b) If the Y_n are mutually independent, then $\sum_{n=1}^{\infty} \mathbb{P}[|Y_n| \geq c] = \infty$ implies $\limsup_{n \rightarrow \infty} |Y_n| \geq_{\text{as}} c$.

We list the exercises as they have occurred throughout this section.

EXERCISES: (3.12), (3.15), (3.17), (3.21), (3.26), (3.27).

4. The usual nonparametric assumptions

In this section, we briefly discuss the assumptions needed to establish consistency and convergence rates in (nonparametric) kernel density estimation. Here, we only consider the kernel estimator (1.18) with $A = \mathfrak{B}$, the two-sided exponential kernel, but the conclusions apply to general kernels. The goal is to establish bounds on the L^1 error $\|\mathfrak{B}_h * dF_n - f_o\|_1$.

Since the kernel estimator $\mathfrak{B}_h * dF_n(x)$ is the mean of the iid random variables $\mathfrak{B}_h(x - X_i)$, with expected values $\mathfrak{B}_h * dF_o(x) = \mathfrak{B}_h * f_o(x)$ and variances $(\mathfrak{B}_h)^2 * f_o - (\mathfrak{B}_h * f_o)^2$, its mean is $\mathfrak{B}_h * f_o$ and its variance is

$$(4.1) \quad \mathbb{E}[|\mathfrak{B}_h * (dF_n - dF_o)(x)|^2] = (nh)^{-1} \{ (\mathfrak{B}^2)_h * dF_o(x) - (\mathfrak{B}_h * dF_o(x))^2 \} \leq (nh)^{-1} \{ (\mathfrak{B}^2)_h * dF_o(x) \} .$$

From the triangle inequality

$$|\mathfrak{B}_h * dF_n(x) - f_o(x)| \leq |\mathfrak{B}_h * dF_o(x) - f_o(x)| + |[\mathfrak{B}_h * (dF_n - dF_o)](x)| ,$$

it then follows that

$$(4.2) \quad \mathbb{E}[|\mathfrak{B}_h * dF_n(x) - f_o(x)|] \leq |\mathfrak{B}_h * dF_o(x) - f_o(x)| + (nh)^{-1/2} \sqrt{(\mathfrak{B}^2)_h * dF_o(x)} .$$

and so, upon integration,

$$(4.3) \quad \mathbb{E}[\|\mathfrak{B}_h * dF_n - f_o\|_1] \leq \|\mathfrak{B}_h * f_o(x) - f_o(x)\|_1 + (nh)^{-1/2} \|\sqrt{(\mathfrak{B}^2)_h * f_o}\|_1 .$$

At this point, it seems clear that some assumptions are needed to get convergence rates. The *usual nonparametric assumptions* are designed to deal with each term in (4.3) separately. Thus, an opaque way of phrasing the usual nonparametric assumptions is

$$(4.4) \quad \limsup_{h \rightarrow 0} h^{-2} \|\mathfrak{B}_h * f_o(x) - f_o(x)\|_1 = C ,$$

$$(4.5) \quad \limsup_{h \rightarrow 0} \|\sqrt{(\mathfrak{B}^2)_h * f_o}\|_1 = C' ,$$

for finite constants C and C' . A somewhat more meaningful way is

$$(4.6) \quad \|(f_o)''\|_1 < \infty ,$$

$$(4.7) \quad \mathbb{E}[|X|^\kappa] < \infty \quad \text{for some } \kappa > 1 .$$

These last two conditions are referred to as the *usual nonparametric assumptions*. The conditions (4.4) and (4.6) are just about equivalent, but condition (4.7) is slightly stronger than (4.5), see §4.2.

With these nonparametric assumptions, it follows from (4.3) that

$$(4.8) \quad \mathbb{E}[\|\mathfrak{B}_h * dF_n - f_o\|_1] \leq c \{ h^2 + (nh)^{-1/2} \} ,$$

for a suitable constant c . Asymptotically, $c = \min(C, C')$. The conclusion is that

$$(4.9) \quad \|f_o - \mathfrak{B}_h * dF_n\|_1 \longrightarrow_{\mathbb{E}} 0 , \quad \text{provided } h \longrightarrow 0, \quad nh \longrightarrow \infty .$$

Moreover, the right-hand side of (4.8) is minimized (asymptotically) for $h \asymp n^{-1/5}$, which gives

$$(4.10) \quad \mathbb{E}[\|\mathfrak{B}_h * dF_n - f_o\|_1] = \mathcal{O}(n^{-2/5}) .$$

There is of course no indication that this is a sharp bound, but in fact it is, see DEVROYE (1987). What is surprising is that the same rate applies to the almost sure convergence of $\|\mathfrak{B}_h * dF_n - f_o\|_1$, even for data-driven choices of h . We come back to all of this in Chapters 4 and 7.

5. Parametric vs nonparametric rates

We have seen that the difference between parametric and nonparametric estimation lies in the dimensionality of the parameter to be estimated. In this section, we show that this results in differences between the convergence rates of some natural estimators. The assumption is that these natural estimators achieve the optimal convergence rate, or at least cannot be drastically improved on.

To elaborate on this, and to make it more precise, suppose that the density f_o to be estimated belongs to some class \mathcal{F} of densities, e.g., the class of all densities on $(-\infty, \infty)$, the class of all decreasing densities on $(0, \infty)$, or the class $PDF(C, C')$ of all densities f for which

$$(5.1) \quad \limsup_{h \rightarrow 0} h^{-2} \|f - \mathfrak{B}_h * f\|_1 \leq C ,$$

$$(5.2) \quad \limsup_{h \rightarrow 0} \|\sqrt{(\mathfrak{B}^2)_h} * f\|_1 \leq C' ,$$

for fixed, known constants C, C' .

Let f^n be an estimator of f_o , that is, f^n is a function of the data

$$(5.3) \quad f^n(x) = f^n(x; X_1, X_2, \dots, X_n) , \quad -\infty < x < \infty .$$

Thus, as in kernel estimation, f^n may incorporate a smoothing parameter h , as long as the choice of h is data driven, $h = h_n(X_1, X_2, \dots, X_n)$. The best possible (expected) estimation error is then

$$(5.4) \quad \mathfrak{R}_{\mathcal{F}}(n) = \inf_{f^n} \sup_{f_o \in \mathcal{F}} \mathbb{E}[\|f^n - f_o\|_1] ,$$

where we picked the L^1 error for reasons discussed in §3. The above quantity is known as the (expected) minimax error for the class \mathcal{F} . In this text, we are not overly concerned with determining the rate at which the minimax error tends to 0, although we assume that we know what it is. For the generic parametric estimation problem, the maximum likelihood estimator has expected minimax error $\mathcal{O}(n^{-1/2})$, so that $\mathfrak{R}_{\mathcal{F}}(n) = \mathcal{O}(n^{-1/2})$. This rate goes by the name of the parametric rate, even though there are exceptional parametric problems in which one achieves better rates. We come

back to this in Part I. For nonparametric problems, one usually achieves a lower convergence rate. By way of example, for the class $PDF(C, C')$, one has

$$(5.5) \quad 0 < \liminf_{n \rightarrow \infty} n^{2/5} \mathfrak{R}_{PDF(C, C')}(n) \leq \limsup_{n \rightarrow \infty} n^{2/5} \mathfrak{R}_{PDF(C, C')}(n) < \infty .$$

In § 4, we showed that this was the upper bound. Thus, it would be nice if the parametric convergence rate characterized parametric problems, as in the following “theorem”.

(5.6) “THEOREM”. *If $\mathfrak{R}_{\mathcal{F}}(n) = \mathcal{O}(n^{-1/2})$, then \mathcal{F} is a parametric family with a finite- dimensional parameter.*

Unfortunately, the theorem fails, a counter example being provided by the class \mathcal{F} of pdfs whose Fourier transforms (characteristic functions) are known to vanish outside of the interval $(-1, 1)$, see § 4.7. Thus, the dimensionality of the parameter to be estimated is not an indicator of the achievable convergence rate.

The following distinction appears to come closer. Although for many parametric problems there exists unbiased estimators with finite variance, this does not happen for problems with a nonparametric convergence rate, as shown by DEVROYE and LUGOSI (2000).

(5.7) THEOREM. [DEVROYE and LUGOSI (2000)] *The minimax error satisfies $\limsup_{n \rightarrow \infty} \sqrt{n} \mathfrak{R}_{\mathcal{F}}(n) = \infty$ if and only if for every $n \geq 1$ and for every estimator f^n at least one of the following two statements holds.*

(a) f^n is biased

$$\sup_{f_o \in \mathcal{F}} \|f_o - \mathbb{E}[f^n]\|_1 > 0;$$

(b) f^n has infinite variance

$$\sup_{f_o \in \mathcal{F}} \|\sqrt{\mathbb{E}[(f^n)^2]} - f_o\|_1 = +\infty .$$

It should be noted that if (a) fails, i.e., $\mathbb{E}[f^n] = f_o$, and (b) holds, then

$$\|\sqrt{\mathbb{E}[(f^n)^2]} - f_o\|_1 = +\infty ,$$

and this resembles the variance of the estimator being ∞ . The proof of the theorem is simple and ingenious.

PROOF. \Rightarrow : Let $f_o \in \mathcal{F}$, and suppose that f^n is an unbiased estimator of f_o . Let n and p be natural numbers. Later, we take p fixed and let $n \rightarrow \infty$. Consider the following estimator of f_o :

$$(5.8) \quad f_{pn}(x) = \frac{1}{n} \sum_{i=1}^n f^p(x; X_{pi+1}, X_{pi+2}, \dots, X_{pi+p}) ,$$

$-\infty < x < \infty$. Obviously, f_{pn} is unbiased. Since f_{pn} is the sum of n independent, identically distributed random variables, then for each x , the variance of $f_{pn}(x)$ satisfies

$$\text{Var}[f_{pn}(x)] = \frac{1}{n} \text{Var}[f^p(x)] .$$

It follows that

$$(5.9) \quad \mathbb{E}[\|f_{pn} - f_o\|_1] = \mathbb{E}[\|f_{pn} - \mathbb{E}[f_{pn}]\|_1] \leq n^{-1/2} \|\sqrt{\text{Var}[f^p]}\|_1 .$$

This holds for any $f_o \in \mathcal{F}$. Now, choose f_o , depending possibly on the sample size pn , so as to be within a factor 2 of the minimax error, i.e.,

$$\mathfrak{R}_{\mathcal{F}}(pn) \leq 2 \mathbb{E}[\|f_{pn} - f_o\|] ,$$

so from (5.9),

$$(pn)^{1/2} \mathfrak{R}_{\mathcal{F}}(pn) \leq 2 p^{1/2} \|\sqrt{\text{Var}[f^p]}\|_1 .$$

Now, keep p fixed, and let $n \rightarrow \infty$. Then, the left-hand side tends to ∞ , by assumption. Since p is fixed, this implies that

$$\|\sqrt{\text{Var}[f^p]}\|_1 \rightarrow \infty \quad \text{as } n \rightarrow \infty .$$

In other words, the variance equals ∞ .

Q.e.d.

(5.10) EXERCISE. Prove the only if part of the theorem.

EXERCISES : (5.10).

6. Sieves and convexity

Roughness penalization in its various forms is a standard way of making sense out of the nonparametric maximum likelihood problems (1.16) and (2.4). A second method for obtaining smooth estimators was proposed by GRENANDER (1981), and its idea is to restrict the minimization in the maximum likelihood problem to classes of smooth densities, given the colorful name of *sieves*. In this section, we discuss sieves and make the point that the method of sieves is similar to roughness penalization.

The identifying feature of a sieve is that it is a *small* set of densities, as opposed to the *big* class of all densities. There are essentially two kinds of sieves. One kind is obtained by considering finite-dimensional subspaces of $L^1(\mathbb{R})$, and then a sieve consists of all pdfs in a particular subspace. Since each subspace is finite-dimensional, the pdfs it contains cannot be arbitrarily rough. The other kind of sieve is obtained by considering compact subsets of $L^1(\mathbb{R})$. For the present purpose, compactness of a set means that the set can be approximated well by finite-dimensional subspaces, in a sense to be made precise below. However, things get more interesting

for particular kinds of compact sets. We discuss the two kinds of sieves in some detail.

The simplest example of sieves is when we have a nested sequence of finite-dimensional subspaces of continuous functions

$$(6.1) \quad V_1 \subset V_2 \subset \cdots \subset V_m \subset \cdots \subset L^1(\mathbb{R}) ,$$

which is dense in $L^1(\mathbb{R})$, i.e., for all $\varphi \in L^1(\mathbb{R})$,

$$(6.2) \quad \lim_{m \rightarrow \infty} \inf \{ \|v - \varphi\|_1 : v \in V_m \} = 0 .$$

(6.3) *EXAMPLE.* A useful and simple example of such subspaces is when V_m consists of all step functions that vanish outside of the interval $[-m, m]$, and with jumps at the points j/m , $j = 0, \pm 1, \pm 2, \dots, \pm m^2$. One verifies that (6.2) does indeed hold, and V_{2m} , $m = 1, 2, \dots$, is a nested sequence as in (6.1). Moreover, $\dim(V_m) = 2m^2 + 1$. The pdfs in V_m are histograms.

With these subspaces in hand, the method of sieves determines estimators in the deconvolution problem (2.4) as solutions to

$$(6.4) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} \log \mathcal{K}f(x) dG_n(x) \\ & \text{subject to} && f \in V_m, f \text{ is a pdf} , \end{aligned}$$

for a suitably chosen value of m , similar to choosing the smoothing parameter in penalization problems. Variations on the above theme abound, of which we mention two.

For the first one, recall that if f is a pdf, then $\sqrt{f} \in L^2(\mathbb{R})$. Alternatively, if $\varphi \in L^2(\mathbb{R})$, then $\varphi^2 \in L^1(\mathbb{R})$, and φ^2 is nonnegative. So it makes sense to choose a sequence of finite-dimensional subspaces of $L^2(\mathbb{R})$

$$(6.5) \quad W_1 \subset W_2 \subset \cdots \subset W_m \subset \cdots \subset L^2(\mathbb{R}) ,$$

which is dense in $L^2(\mathbb{R})$, analogous to (6.2), and to consider the problem

$$(6.6) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} \log[\mathcal{K}(\varphi^2)](x) dG_n(x) \\ & \text{subject to} && \varphi \in W_m, \|\varphi\|_2 = 1 . \end{aligned}$$

If $\varphi^{n,m}$ denotes the solution to (6.6), then $f = (\varphi^{n,m})^2$ is the estimator of the pdf. Subspaces of $L^2(\mathbb{R})$ satisfying (6.5) may be constructed using any convenient orthonormal basis for $L^2(\mathbb{R})$, see, e.g., SANSONE (1991).

In the second variation, one defines exponential families

$$(6.7) \quad f(x) = \varphi_o(x) \exp(w(x)) , \quad -\infty < x < \infty ,$$

where φ_o is some fixed pdf and $w \in W_m$, say. The minimization problem may then be formulated accordingly. We note the special case of density

estimation, where $\mathcal{K}f = f$ for all f ,

$$(6.8) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} w(x) dG_n(x) \\ & \text{subject to} && w \in W_m, \text{ and} \\ & && \int_{\mathbb{R}} \varphi_o(x) \exp(w(x)) dx = 1. \end{aligned}$$

Here, the objective function is simple when compared with (6.6), say, but the pdf constraint is more complicated. A related problem is considered in § 5.3.

In the above, it should be observed that the sieves serve the second purpose of *discretizing* the estimation problem. The original problems being nonparametric, that is, infinite-dimensional, a crucial step in any computation is the approximation of the infinite-dimensional problem by a finite-dimensional one. It is the task of numerical analysis to investigate discretization errors, but in this text, we assume that the discretization effects are negligible compared with the effects of noisy data. Of course, one had better make sure, experimentally say, that this is indeed the case.

A cursory inspection of estimation by the method of sieves reveals that the dimension of the subspaces V_m plays the role of the smoothing parameter, and the question arises of how it should be chosen. On the one hand, we want the subspaces V_m or W_m to be big, so that our unknown density f_o may be approximated well by elements from V_m or W_m . On the other hand, bigger subspaces contain rougher pdfs, and that means that the estimators will be rougher. This is the analogue of the bias-variance trade off for kernel density estimation.

On to the second kind of sieves, obtained by considering nested sequences of compact subsets of continuous functions

$$(6.9) \quad C_1 \subset C_2 \subset \cdots \subset C_m \subset \cdots \subset L^1(\mathbb{R}),$$

which are dense in $L^1(\mathbb{R})$. We do not go into the details, except to note that compactness here may be interpreted to mean *almost finite dimensional* in the sense that if $C \subset L^1(\mathbb{R})$ is compact, then for any nested sequence of finite-dimensional subspaces V_m which is dense in $L^1(\mathbb{R})$, as in (6.1)–(6.2), one has

$$(6.10) \quad \sup_{\varphi \in C} \inf_{v \in V_m} \|v - \varphi\|_1 \longrightarrow 0 \quad \text{as } m \rightarrow \infty.$$

For related material, see Appendix 3.

So in the above interpretation, admitting compact sets as sieves does not drastically alter the state of affairs, especially keeping in mind the need for discretization. We arrive at a different interpretation when considering nested compact subsets of the form

$$(6.11) \quad C_M = \{ f \text{ is a pdf} : R(f) \leq M \},$$

for a suitable roughness functional $R(f)$. In the actual application of these sets, the constant M must be chosen appropriately. We restrict attention to convex roughness functionals, that is, functionals satisfying

$$(6.12) \quad R(\lambda f + (1 - \lambda)g) \leq \lambda R(f) + (1 - \lambda)R(g) ,$$

for all f, g , and all $0 \leq \lambda \leq 1$. Now, the sieved version of (6.4) reads as

$$(6.13) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} \log \mathcal{K}f(x) dG_n(x) \\ & \text{subject to} && R(f) \leq M , \ f \text{ is a pdf} . \end{aligned}$$

It is important to note that the objective function in (6.13) is likewise convex. It follows from the theory of convex minimization problems, studied in detail in Part III, that (6.13) is equivalent to

$$(6.14) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} \log \mathcal{K}f(x) dG_n(x) + h^2 R(f) \\ & \text{subject to} && f \text{ is a pdf} , \end{aligned}$$

where the *Lagrange multiplier* h^2 is chosen such that the solution f^{nh} of (6.14) satisfies

$$(6.15) \quad R(f^{nh}) = M .$$

(In exceptional cases, the solution f of (6.13) satisfies $R(f) < M$. This possibility is ignored here.) It is now apparent that this version of the method of sieves and maximum penalized likelihood estimation are just about equivalent, especially upon noting that typically the constant M is unknown. In view of (6.15), this means that one may omit all references to M , and consider h^2 as the primary unknown. However, one sees that choosing M in the sieve context is the same as choosing h in the penalization framework.

Another interesting kind of compact sieves are the *convolution sieves*, although the name *kernel sieves* comes to mind also. These sieves are indexed by a smoothing parameter h and take the form

$$(6.16) \quad C_h = \{ A_h * dP : P \text{ is a probability distribution} \} .$$

So in this case, the solutions to the (sieved) maximum likelihood problem are restricted to being kernel estimators. The resulting problem may be phrased as a problem to estimate a distribution function

$$(6.17) \quad \begin{aligned} & \text{minimize} && - \int_{\mathbb{R}} \log(A_h * dP(x)) dF_n(x) \\ & \text{subject to} && P \text{ is a probability distribution} . \end{aligned}$$

Denote the solution by $P^{n,h}$. Then, the estimator for f_o is $f^{nh} = A_h * dP^{n,h}$. For the plain density estimation problem ($\mathcal{K}f = f$ for all f), WALTER and BLUM (1984) give explicit solutions for the case in which A is the two-sided

exponential kernel. The optimal P consists of discrete point masses at the X_i , but the weights are not necessarily equal.

This description of the method of sieves must suffice, and serves as the authors' justification for studying estimation from the penalization point of view only. The above also hints at the important role played in this text by convexity.

7. Additional notes and comments

Ad § 1: There are numerous papers on the failure and optimality of parametric maximum likelihood estimation. We stay optimistic, and only quote the optimality paper, YATRACOS (1998). The derivation of the least-squares method (1.13), and (1.23) leading to kernel density estimators is very well known in nonparametric density estimation, see STONE (1984). For the method of maximum spacings, see PYKE (1965), CHENG and AMIN (1983), RANNEYBY (1984), and the recent GOSH and RAO JAMMALA-MADAKA (2001). See also § 2.8.

Ad § 2: GOOD (1971) uses *Bayesian* arguments to justify his roughness penalization functional (1.21). That it pops up in a rather unexpected way in § 2, via nonlinear smoothing of EM algorithms, lends further support to his insights.

Ad § 3: The L^1 setting for density estimation was originally advocated by DEVROYE and GYÖRFI (1985), and the authors wholeheartedly embrace their point of view. However, there is a huge literature on the L^2 setting, to which the reader is welcome.

Ad § 5: This section is largely based on the last section of DEVROYE and LUGOSI (2000).

Ad § 6: The model (6.8) comes back in disguised form in §§ 5.3 and 4, under the name of log splines. See the references there.

Maximum Penalized Likelihood Estimation

Volume I: Density Estimation

Eggermont, P.P.B.; LaRiccia, V.N.

2001, XVIII, 512 p., Hardcover

ISBN: 978-0-387-95268-0