Distributional inference

Albers, Casper Johannes

Part II

Statistical inference, distributional inference in particular

Chapter 3

Estimating a density by adapting an initial guess

'In fact, one of the few points of agreement between the antagonists R.A. Fisher and J. Neyman was that one would seldom have enough information about the unknown state of nature to assume a prior distribution for it and that lacking such prior possibilities the formula of Bayes would not be applicable.'

T. Ferguson[1]

Most of classical statistics is based on a statistical model $\mathcal{P}$ and the idea that nothing is known a priori about the location of the true distribution P in the family $\mathcal{P}$. A priori information about the location of P in $\mathcal{P}$ is almost always a factual reality. We are interested in situations where this information is expressed in the form of an initial guess P_0 of P. It goes without saying that this initial guess may be difficult to compose and that it should not dominate our inferences.

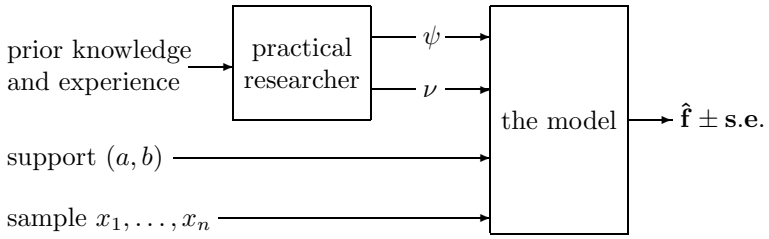
For a large variety of reasons, it is of interest to pay attention to nonparametric density estimation methods and to go beyond the familiar kernel methods. See, e.g., Silverman (1986) and De Bruin et al. (1999) for more specific motivation.

When initial knowledge is present, it should be incorporated by the model that is used for generating inferences. De Bruin et al. (1999) provide a unique method for density estimation, where such initial knowledge is taken into account. The method will be defined in Section 3.2 and discussed in the sections thereafter. These discussions show that (some) improvement is needed, amongst other reasons because the rate of convergence is too low to be satisfactory. Improvement is found in a U-statistic symmetrization, see Sections 3.6 and 3.7. Theoretical and simulation results suggest that our improved method is a good alternative to existing density estimation methods. In Chapter 5 an extensive case study is performed.

The current chapter is based on the article by Albers and Schaafsma with the same title that will appear in Computational Statistics and Data Analysis (2003a). Most results in Sections 3.3 and 3.4 were previously published as a technical report (Albers and Schaafsma, 2001a).

[1] Mathematical Statistics – a decision-theoretic approach, Academic Press, New York, 1967, p. 32.

Figure 3.1: Visualization of our method. The researcher's outputs, an initial distribution ψ and ν, a 'measure of belief in ψ', are inputs to the model.



3.1 Introduction

Many statisticians, epistemologists, and other scientists have discussed the fundamental problem of practical statistics, which is as follows. *Given* is the number m of successes in n independent and identical Bernoulli trials with unknown probability of success p. *Required* is a numerical assessment of the probability of s successes in a future series of r experiments. Of course, if n is large then the $\text{Bin}(r, \frac{m}{n})$-distribution will approximately yield the required posterior probabilities. The 'solution' for non-large n, provided by Bayes, Laplace, Pearson, Fisher and others was that in the absence of further information about p one should use the Bayesian approach with uniform prior. The De Finetti Representation Theorem (see, e.g., BILLINGSLEY, 1995) states that (probabilistic) coherency is impossible, unless the Bayesian approach is adopted. In this Bayesian approach one needs to express a priori knowledge about the statistical model in the form of a (subjective) prior distribution. This is a hazardous task, and serious 'misspecification' of the prior will lead to bad inferences. 'The Bayesian approach to statistical problems, though fruitful in many ways, has been rather unsuccessful in treating nonparametric problems. This is due primarily to the difficulty in finding workable prior distributions on the parameter space, which in nonparametric problems is taken to be a set of probability distributions on a given sample space.' (FERGUSON, 1973). Note that Ferguson is referring to nonparametric problems for which he introduced Dirichlet-process priors. These are convenient if the distribution function F has to be estimated, but not if a nonparametric density estimate is required. We concentrate the attention of such continuous analogue of the above, with $r = 1$, is as follows.

Given are the outcomes $x_{[1]} < x_{[2]} < \ldots < x_{[n]}$ of an independent random sample $X_1, \ldots, X_n$ from a probability distribution on $\mathbb{R}$ with a density f which is 'smooth' (at least continuous) and strictly positive on a given support (a, b), e.g. $(0, \infty)$, but further (largely) unknown. *Required* is an estimate $f_n(x)$ of $f(x)$ in a given point x or, more generally, an estimate f_n of f on (a, b), including a statement about the uncertainty involved, see Figure 3.1 for a visualization. Such $f_n(x)$ should be regarded as distributional inference for the next observation $x_{[n+1]}$.

The performance of density estimates is measured by specifying a difference between the true density and the estimate. Many dissimilarity coefficients can be considered (such as L_1 or L_2-distance between the densities, maximal absolute difference, L_1-distance between the quantile functions, or, equivalently, L_1-distance between the distribution functions, etcetera). We restrict the attention to the L_1-distance or total variation distance

$$||f_n - f||_1 = \int_a^b |f_n(x) - f(x)| \, \mathrm{d}x.$$

Note: various authors, e.g., DEVROYE (1987), define the total variation as

$$d_{TV} = \sup_A \left| \int_A f_n - \int_A f \right|,$$

which is half the L_1-distance. Although this definition has the advantage that the metric is 'normalized' (has values between 0 and 1), we use the first mentioned definition, taken from the field of functional analysis (see, e.g.DUNFORD AND SCHWARTZ, 1957).

We prefer the L_1-distance (and with that the total variation distance) above other dissimilarity coefficients because it is invariant under all piecewise continuous (differentiable) bijections and because it provides the metric which was preferred by a variety of other authors (see, e.g., DEVROYE AND GYŐRFI, 1985, DEVROYE, 1987, WAND AND DEVROYE, 1993, and DE BRUIN ET AL., 1999).

Working with the distribution function F has the disadvantage that the best unbiased (hence natural) estimate, the empirical distribution function, violates the differentiability requirement. Bayesian analogues derived by FERGUSON (1973), as indicated earlier, on the basis of Dirichlet-process priors, require the specification of a distribution function Ψ, but provide Bayes estimates of F which are not differentiable either. The total variation between these estimates and the true distribution is equal to 2 and consistency is precluded (for a review of consistency and convergence, see GHOSAL, 1997, and GHOSAL ET AL., 2000). A plethora of methods exists for specifying estimates of f such that the total variation converges to 0 in probability if $n \to \infty$ and f is sufficiently smooth (see, e.g., GHOSAL ET AL., 1999). In Ferguson's theory, a Dirichlet prior must be specified by providing a distribution function. The method of DE BRUIN ET AL. (1999) is also based on the specification of a distribution function Ψ, now as an initial guess of the true distribution function F. The derivative ψ of Ψ is an initial guess of f. The theory behind this method requires that ψ is chosen a priori. It is assumed that the 'support' $\{x : \ \psi(x) > 0\}$ is specified as the true interval $\{x : \ f(x) > 0\}$ (usually $(0, \infty)$ or $(-\infty, \infty)$). The method provides a special estimate f_n of f, as well as a statement about the standard errors involved.

The underlying rationale is that there are many ways to characterize a probability distribution. One can use the distribution function F, the density function $f = F'$, the quantile function $H = F^{-1}$, the characteristic function, the moment-generating

function, etcetera. Each of these characterizations entails its own estimation approach (see SILVERMAN, 1986). The empirical distribution function is not appropriate if one is interested in the estimation of $f = F'$. For that purpose some smoothing operation has to be applied. It is, of course, an interesting question whether the empirical distribution function F_n, the empirical quantile function $H_n = F_n^{-1}$, or some analogue characteristic has to be subjected to the smoothing operation.

We concentrate the attention on the estimation of H, which will turn out to be very convenient to estimate the quantile density $h = H'$ and the probability density $f = F' = (H^{-1})'$. Our estimation will be performed by fitting a Bernstein polynomial to the data (next section), and applying an additional smoothing by using U-statistic symmetrization (Section 3.6). Other perspectives of density estimation have their own advantages and disadvantages. Of particular interest in this context are:

1. The estimation of f directly through kernel estimation, see e.g. SILVERMAN (1986) and VAN ES (1988);
2. The estimation of f by means of a Bernstein polynomial (instead of estimating $H = F^{-1}$ like we do), see e.g. VITALE (1975), and GHOSAL (2001);
3. Estimation through orthogonal expansion schemes, wavelets, splines, etcetera.

The comparison between the density estimate introduced in the next section and the ones mentioned above will be discussed in Section 3.5.

3.2 The original density estimation method

The method is based on the characterization by means of the quantile function (see for extensive literature in this context GILCHRIST, 2000). Before discussing the arguments behind the method, we give the precise definition, which is as follows.

Definition 3.1 (De Bruin et al.) *Replace the ordered sample $x_{[1]}, \ldots, x_{[n]}$ by*

$$y_{[0]} = 0, \qquad y_{[i]} = \Psi(x_{[i]}) \quad (i = 1, \ldots, n), \qquad y_{[n+1]} = 1,$$

where $\psi = \Psi'$ is an a priori guess for $f = F'$. Then the estimator B_n of the quantile function $B = (F \circ \Psi^{-1})^{-1} = \Psi \circ F^{-1}$ of the distribution of $Y_1 = \Psi(X_1)$ is

$$B_n(p) = \sum_{i=0}^{n+1} y_{[i]} \binom{n+1}{i} p^i (1-p)^{n+1-i}.$$

The estimation of f is performed via back-transformation: the estimate B_n is used to define the estimate $H_n = \Psi^{-1} \circ B_n$. Next $F_n = H_n^{-1} = B_n^{-1} \circ \Psi$ is used to estimate $F = H^{-1}$ and, finally, the estimate $f_n = F_n'$ of f is obtained by numerical differentiation. The notation $B_n(p)$ is preferred over the usual notation $B_{n+1}(p)$ because the underlying sample size is n.

The most convincing argument in favor of $B_n(p)$ is as follows. Involving the Ψ-transformed quantile distribution function $B(p) = \Psi(F^{-1}(p))$ of Y, we have the physical probability

$$\begin{aligned} \mathrm{P}(B(p) \leq Y_{[i]}) &= \mathrm{P}(p \leq G(Y_{[i]}) = U_{[i]}) \\ &= \sum_{k=0}^{i-1} \binom{n}{k} p^k (1-p)^{n-k}, \end{aligned}$$

where $G = B^{-1}$ is the probability distribution function of Y, and $U_{[i]}$ is the i-th order statistic of a sample of size n from the standard uniform distribution $\mathcal{U}(0,1)$. This result was first derived by THOMPSON (1936). This physical probability is used to postulate the epistemic probability[2]

$$\mathrm{P}(B(p) \leq y_{[i]}) = \sum_{k=0}^{i-1} \binom{n}{k} p^k (1-k)^{n-k}.$$

Plugging in a sample statistic for a population statistic to obtain a probability statement is, of course, not new. It is quite similar to Fisher's Fiducial Argument (see, e.g., FISHER, 1930 and SALOMÉ, 1998). Thompson's rule easily provides

$$\mathrm{P}(Y_{[i]} < B(p) \leq Y_{[i+1]}) = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n,$$

or, epistemologically equivalent, a distributional inference about $B(p)$ such that these binomial probabilities are assigned to the intervals $(y_{[i]}, y_{[i+1]})$. Taking the expectation[3] we obtain the L-estimator for $B(p)$

$$\tilde{B}_n(p) = \sum_{i=0}^{n} h(y_{[i]}, y_{[i+1]}) \binom{n}{i} p^i (1-p).^{n-i}, \quad p \in [0,1],$$

with $h(y_{[i]}, y_{[i+1]}) \in [y_{[i]}, y_{[i+1]})$ a measurable function. When $h(y_{[i]}, y_{[i+1]}) = (1-p)y_{[i]} + py_{[i+1]}$ the Bernstein polynomial $B_n(p)$ defined in Definition 3.1 is obtained, and for $h(y_{[i]}, y_{[i+1]}) = \frac{1}{2}y_{[i]} + \frac{1}{2}y_{[i+1]}$ the Kantorovič polynomial

$$\tilde{K}_n(p) = \sum_{i=0}^{n} \frac{y_{[i]} + y_{[i+1]}}{2} \binom{n}{i} p^i (1-p).^{n-i},$$

[2] In literature, sometimes the notation $\mathrm{Q}(\cdot)$ is used for epistemic probabilities instead of $\mathrm{P}(\cdot)$, which is used for 'physical' probabilities. We do not distinguish in notation, and let the interpretation of P be defined by the context.

[3] Note that

$$\binom{n+1}{i} \int_0^1 p^i (1-p)^{n+1-i} \, \mathrm{d}p = \frac{1}{n+1}.$$

follows. The Bernstein polynomial approximation $B_n(p)$ is very convenient because the corresponding derivative

$$b_n(p) = \sum_{i=0}^{n} (y_{[i+1]} - y_{[i]}) \binom{n}{i} (n+1) p^i (1-p)^{n-i}$$

is strictly positive and continuous. If one differentiates the Kantorovič polynomial a less pleasant expression is obtained.

Polynomial estimators B_n and K_n (without Ψ-transformation) were already given and analyzed in MUÑOZ PEREZ AND FERNÁNDEZ PALACÍN (1987), and similar formulas for the density function in VITALE (1975). MUÑOZ PEREZ AND FERNÁNDEZ PALACÍN (1987) prove that $\lim_n \mathbf{E}\, B_n(p) = B(p)$ uniformly on $[0,1]$ in case $B(p)$ is continuous with bounded derivative. Furthermore, concerning the integrated risk, they provide

$$\mathbf{E} \int |B_n - B| \sim o(n^{-1/2}), \quad \text{and} \quad \mathbf{E} \int |\mathbf{E}\, B_n - B| \sim o(n^{-1/2}),$$

and, when H is also twice differentiable on $(0,1)$ with continuous derivative on $[0,1]$,

$$\mathbf{E} \int |B_n - \mathbf{E}\, B_n| \sim o(n^{-1/2}).$$

In Groningen, the study of polynomial quantile estimators started with DEHLING ET AL. (1991) , involving a study of the Islamic Mean. The Islamic Mean is obtained by taking pairwise averages of the n order statistics, after that taking pairwise averages of the $n-1$ averages, etcetera, until a single number is obtained. This L-statistic

$$T_n = \sum_{i=1}^{n} x_{[i]} \binom{n-1}{i-1} 2^{-n+1},$$

is very similar to $B_n(\frac{1}{2})$. DEHLING ET AL. (1991) show that the Islamic Mean, and hence $B_n(\frac{1}{2})$ are interesting alternatives to the sample median for estimating the population median.

DE BRUIN ET AL. (1999) extended this suggestion to the case of the entire quantile function, using semi-Bayesian and Fisherian arguments. This article also introduced the use of an 'initial guess' to increase the performance (at least in case some a priori knowledge or intuition is present). Not the original quantile distribution function $H = F^{-1}$, but one transformed to the unit interval, $B = \Psi \circ F = (F \circ \Psi)^{-1}$, is estimated. The estimator is transformed back using Ψ to obtain an estimator for H, and eventually F and f. The estimator $B_n(p)$ thus obtained in Definition 3.1 is, as can be easily seen, a linear function of the order statistics (of the transformed sample), where Bernstein polynomials are used as smooth weight functions, or kernels.

The example in Figure 3.2 illustrates the performance of the density estimation method based on a sample of size $n = 100$ from the Beta$(3,2)$-density, and $\psi = 1$.

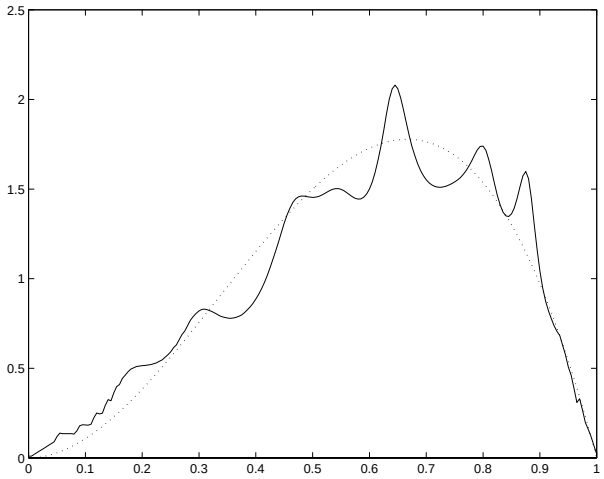


Figure 3.2: Visualization of the estimation by $f_n(x)$.

3.3 Asymptotic properties of the quantile function estimate

In this section it is established that the quantile density estimates $b_n(p)$ can be equipped with the standard error

$$\frac{b_n(p)}{\sqrt[4]{4\pi np(1-p)}}.$$

In Section 3.4, a similar result will be suggested for the density estimation function f_n which, however, does not hold in general. The proof involving the quantile estimation function (Theorem 3.1) requires two lemmas. An introduction is as follows[4].

The values $y_{[1]}, \ldots, y_{[n]}$ are ordered outcomes of an independent random sample from the distribution on $[0,1]$ with distribution function $F \circ \Psi^{-1} = B^{-1}$. If we define $u_{[i]} = F(\Psi^{-1}(y_{[i]})) = B^{-1}(y_{[i]})$, then we have that $(u_{[1]}, \ldots, u_{[n]})$ is the ordered outcome of an independent random sample from the uniform distribution on $[0,1]$. Hence, replacing $y_{[i]}$ by $B(u_{[i]})$, we obtain, by applying the Mean Value Theorem, that $v_i \in [u_{[i]}, u_{[i+1]}]$ $(i = 0, \ldots, n)$ exist such that

$$\begin{aligned} b_n(p) &= \sum_{i=0}^{n} \left(B(u_{[i+1]}) - B(u_{[i]})\right)(n+1)\binom{n}{i}p^i(1-p)^{n-i} \\ &= \sum_{i=0}^{n} b(v_i)\left(u_{[i+1]} - u_{[i]}\right)(n+1)\binom{n}{i}p^i(1-p)^{n-i}. \end{aligned}$$

[4]Reading the technicalities in this and the following section is not a necessity for a good understanding of the remainder of the chapter.

The expectations $b_{n,i} = \binom{n}{i}p^i(1-p)^{n-i}$ of the weights of the $b(v_i)$ are largest if $i \approx np$ and, hence, $v_i \approx p$. This suggests to replace $b(v_i)$ by $b(p)$ Using the approximation $\tilde{b}_n(p)$ of $b_n(p)$ thus obtained, we shall have to study the error $\tilde{b}_n(p) - b_n(p)$. This will be done in Lemma 3.2, for which we need the following result.

Lemma 3.1 *With respect to $\tilde{b}_n(p)$ just defined, we have*

$$\mathcal{L}\, n^{1/4}(\tilde{b}_n(p) - b(p)) \to \mathcal{N}\left(0, \frac{b(p)^2}{\sqrt{4\pi p(1-p)}}\right).$$

Proof
Note that the coverages $u_{[i+1]} - u_{[i]}$ are outcomes of random variables $U_{[i+1]} - U_{[i]}$ which have the same joint distribution as the r.v.'s $E_i/(E_0+\ldots+E_n)$ where $E_0, \ldots, E_n$ constitute an independent random sample from the standard negative-exponential distribution. Hence

$$\mathcal{L}\, \tilde{b}_n(p) = \mathcal{L}\left(\frac{b(p)\sum_{i=0}^n E_i\binom{n}{i}p^i(1-p)^{n-i}}{\sum_{i=0}^n E_i/(n+1)}\right).$$

The denominator in the right-hand side has a Gamma$(n+1, n+1)$-distribution. As it converges to 1 (in probability), it suffices to show that

$$\mathcal{L}\, n^{1/4}\left(\sum_{i=0}^n (E_i - 1)b_{n,i}\right) \to \mathcal{N}\left(0, \frac{1}{\sqrt{4\pi p(1-p)}}\right),$$

where $b_{n,i} = \binom{n}{i}p^i(1-p)^{n-i} = \mathrm{P}(S = s)$ if S has the binomial distribution $\mathrm{B}(n,p)$. The expectation of $\sum E_i b_{n,i}$ is obviously zero, the variance is equal to

$$\mathrm{Var}\left(\sum_{i=0}^n (E_i - 1)b_{n,i}\right) = \sum_{i=0}^n b_{n,i}^2 = \mathrm{P}(S_1 - S_2 = 0),$$

where $S_1, S_2 \sim \mathrm{B}(n,p)$ are independent and, as a consequence, $S_1 - S_2$ has a distribution on $\{-n, -n+1, \ldots, n\}$ which is asymptotically $\mathcal{N}(0, 2np(1-p))$. Using a local form of the Central Limit Theorem, we have that[5]

$$\mathrm{P}(S_1 - S_2 = 0) \approx \frac{1}{\sqrt{2\pi \cdot 2np(1-p)}}.$$

[5] An alternative derivation, is by considering S_1 and S_2 in a bivariate sense:

$$\begin{aligned}
\mathrm{P}(S_1 = S_2) &\approx \int_{-\infty}^{\infty}\int_{u-\frac{1}{2}}^{u+\frac{1}{2}} f_{S_1,S_2}(u,v)\,\mathrm{d}v\mathrm{d}u \\
&\approx \sqrt{2}\int_{-\infty}^{\infty} f_{S_1,S_2}(u,u)\,\mathrm{d}u \\
&= \frac{1}{\sqrt{2}\pi np(1-p)}\int_{-\infty}^{\infty}\exp\left\{-\frac{(u-np)^2}{np(1-p)}\right\}\mathrm{d}u \\
&= \frac{1}{\sqrt{4\pi np(1-p)}}.
\end{aligned}$$

Therefore, $\mathrm{Var}(\sum_{i=0}^{n}(E_i-1)b_{n,i}) \approx (4\pi np(1-p))^{-1/2}$, and the first step of the proof is complete.

Expectation and variance being dealt with, the asymptotic normality will follow from the Lyapunov Central Limit Theorem. This theorem implies

$$\mathcal{L}\left(\frac{\sum_{i=0}^{n} E_i b_{n,i} - 1}{\sqrt{\mathrm{Var}(\sum_{i=0}^{n} E_i b_{n,i})}}\right) \to \mathcal{N}(0,1)$$

if the condition

$$\frac{\mathbf{E}\ \sum_{i=0}^{n} |(E_i-1)b_{n,i}|^3}{(\mathrm{Var}(\sum_{i=0}^{n} E_i b_{n,i}))^{3/2}} \to 0 \quad \text{as} \quad n \to \infty$$

is satisfied. The asymptotic behaviour of the denominator has been obtained in the above as $(\sqrt{4\pi np(1-p)})^{-3/2}$. Hence it suffices to establish that the numerator is $o(n^{-3/4})$ or, equivalently, that $\sum b_{n,i}^3$ behaves that way. Note that, using notations similar to before,

$$\sum_{i=0}^{n} b_{n,i}^3 = \mathrm{P}(S_1 = S_2 = S_3) = \mathrm{P}(S_1 - S_2 = S_1 - S_3 = 0)$$

where $(S_1 - S_2, S_1 - S_3)'$ has a distribution on a subset of $\{(i,j) \mid i,j \in \{-n, -n+1, \ldots, n\}\}$, this distribution being asymptotically normal with vector of expectations $(0,0)'$, variances $2np(1-p)$ and covariance $np(1-p)$. Finally, using a local form of the multivariate Central Limit Theorem we obtain that

$$\begin{aligned} \mathrm{P}(S_1 - S_2 = S_1 - S_3 = 0) &\approx \left(4\pi^2 (np(1-p))^2 \det \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\right)^{-1/2} \\ &= (2\pi np(1-p))^{-1} 3^{-1/2}. \end{aligned}$$

Hence the numerator behaves like n^{-1} which is obviously $o(n^{-3/4})$. □

Lemma 3.2 *Under weak regularity assumptions*

$$n^{1/4}(b_n(p) - \tilde{b}_n(p)) \to 0 \qquad \textit{in probability.}$$

Proof
Reformulate

$$n^{1/4}\left(b_n(p) - \tilde{b}_n(p)\right) = n^{1/4} \sum_{i=0}^{n} \left(b(V_i) - b(p)\right)\left(U_{[i+1]} - U_{[i]}\right)(n+1)b_{n,i},$$

where V_i is a random variable satisfying $U_{[i]} \le V_i \le U_{[i+1]}$. About f we know that it is smooth and strictly positive on a given interval (a,b). Almost equivalently, $b(p)$

will be assumed to be smooth and strictly positive. For the sake of convenience, the additional assumption is made that b is differentiable with $|b'(p)| \leq M$ for all p for some finite M. The Mean Value Theorem states that $n^{1/4}(b_n(p) - \tilde{b_n}(p))$ can be rewritten as

$$n^{1/4} \sum_{i=0}^{n} b'(W_i)(V_i - p)(U_{[i+1]} - U_{[i]})(n+1)b_{n,i}$$

with W_i between p and V_i. Hence

$$\begin{aligned} |n^{1/4}(b_n(p) - \tilde{b_n}(p))| &\leq M n^{1/4} \sum_{i=0}^{n} |V_i - p|(U_{[i+1]} - U_{[i]})(n+1)b_{n,i} \\ &\leq M n^{1/4} \sum_{i=0}^{n} \left(|U_{[i+1]} - p| + |U_{[i]} - p|\right)\left(U_{[i+1]} - U_{[i]}\right)(n+1)b_{n,i} \end{aligned}$$

because $V_i \in [U_{[i]}, U_{[i+1]}]$. To establish that

$$n^{1/4} \sum_{i=0}^{n} \left|U_{[i+1]} - p\right|\left(U_{[i+1]} - U_{[i]}\right)(n+1)b_{n,i} \to 0$$

in probability, it suffices to prove that the expectation of the left-hand side tends to 0. The other contribution follows similarly. The Cauchy-Schwartz inequality provides

$$\mathbf{E}\,|U_{[i+1]} - p|(U_{[i+1]} - U_{[i]}) \leq \sqrt{\mathbf{E}\,(U_{[i+1]} - p)^2}\,\frac{\sqrt{2}}{n+1},$$

because the 'coverage' $(U_{[i+1]} - U_{[i]}) \sim \text{Beta}(1, n)$ has expectation $\mathbf{E}\,(U_{[i+1]} - U_{[i]})^2 = \frac{2}{(n+1)(n+2)} < \frac{2}{(n+1)^2}$. Hence it suffices to establish that

$$n^{1/4} \sum_{i=0}^{n} \sqrt{\mathbf{E}\,(U_{[i+1]} - p)^2}\, b_{n,i} \to 0,$$

where $U_{[i+1]} \sim \text{Beta}(i+1, n-i)$ is such that

$$\mathbf{E}\,(U_{[i+1]} - p)^2 = \frac{(i+1)(n-i)}{(n+2)(n+1)^2}\left(\frac{i+1}{n+1} - p\right)^2.$$

To complete the proof we use the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for nonnegative a and b, and note that it suffices to establish that both

$$n^{1/4} \sum_{i=0}^{n} \sqrt{\frac{(i+1)(n-i)}{(n+2)(n+1)^2}}\, b_{n,i} = n^{1/4}\mathbf{E}\,\sqrt{\frac{(S+1)(n-S)}{(n+2)(n+1)^2}} \to 0,$$

and

$$n^{1/4}\sum_{i=0}^{n}\left|\frac{i+1}{n+1}-p\right|b_{n,i}=n^{1/4}\mathbf{E}\left|\frac{S+1}{n+1}-p\right|\to 0.$$

As both statements are trivial, the lemma holds. □

Theorem 3.1 *The quantile-density estimate defined in Section 3.2 has the following limiting density*

$$\mathcal{L}\, n^{1/4}(b_n(p)-b(p))\to\mathcal{N}\left(0,\frac{b(p)^2}{\sqrt{4\pi p(1-p)}}\right).$$

Proof
This follows immediately from Lemma 3.1 and Lemma 3.2. □

3.4 Asymptotic properties of the probability density function estimate

Applications in discriminant analysis and pattern recognition require estimates of $f(x)$, not of $b(p)$. The relationship between these population concepts is as follows. Starting from the distribution function F of X_1 and the corresponding quantile function $H=F^{-1}$, we have that $Y_1=\Psi(X_1)$ has distribution function $F\circ\Psi^{-1}$ and quantile function $B=\Psi\circ F^{-1}$. Hence $H=\Psi^{-1}\circ B$. For the densities $f=F'$, $h=H'$ and $b=B'$ we have that $f(x)=1/h(F(x))$ and $h(p)=1/f(H(p))$ while

$$h(p)=\frac{b(p)}{\psi(\Psi^{-1}(B(p)))}.$$

Hence

$$f(H(p))=\frac{\psi(\Psi^{-1}(B(p)))}{b(p)}$$

and

$$f(x)=\frac{\psi(\Psi^{-1}(B(F(x))))}{b(F(x))}=\frac{\psi(x)}{b(F(x))}.$$

In Section 3.2 sampling analogues were defined.

Theorem 3.2 suggests that $n^{1/4}(f_n(x)-f(x))$ might have a normal limiting distribution analogue to that of $n^{1/4}(b_n(p)-b(p))$. Unfortunately we were not able to establish such result in general. However, if $\Psi=F$, then the required result can be proven as follows.

Theorem 3.2 *If $\Psi = F$ then the difference between the density estimate defined in Section 3.2 and the true density $f = \psi$ satisfies*

$$\mathcal{L}\, n^{1/4}(f_n(x) - f(x)) \to \mathcal{N}\left(0, \frac{f(x)^2}{\sqrt{4\pi F(x)(1-F(x))}}\right).$$

Proof
Note that

$$\begin{aligned} f_n(x) - f(x) &= \frac{\psi(x)\,[b(F(x)) - b_n(F_n(x))]}{b_n(F_n(x))\; b(F(x))} \\ &\quad \frac{\psi(x)\,[b_n(F(x)) - b_n(F_n(x))] - \psi(x)\,[b_n(F(x)) - b(F(x))]}{b_n(F_n(x))\; b(F(x))}. \end{aligned}$$

We claim that the second part dominates the first one. To establish the asymptotic distribution of the second part, note that Theorem 3.1 implies

$$\mathcal{L}\, n^{1/4}\,(b_n(F(x)) - b(F(x))) \to \mathcal{N}\left(0, \frac{b(F(x))^2}{\sqrt{4\pi F(x)(1-F(x))}}\right).$$

Furthermore, not spelling out the details, we have

$$\frac{\psi(x)^2}{b_n(F_n(x))^2 b(F(x))^2}\,\frac{b(F(x))^2}{\sqrt{4\pi F(x)(1-F(x))}} \approx \frac{f(x)^2}{\sqrt{4\pi F(x)(1-F(x))}}.$$

Finally we complete the proof by, indeed, establishing that $b_n(F(x)) - b_n(F_n(x))$ is of smaller order of magnitude than $b_n(F(x)) - b(F(x))$, which is $O(n^{-1/4})$, as is shown in the following lemma. □

Lemma 3.3 *If $\Psi = F$ then*

$$b_n(F_n(x)) - b_n(F(x)) = o(n^{-1/4}).$$

Proof
The proof is similar to that of Lemma 3.1. A difficulty is that $b_n'(p)$ does *not* tend to $b'(p)$, the difference being $O(n^{1/4})$. Compare $b_n(p)$ with $b_n(r)$ $(0 \le p, r \le 1)$.

$$b_n(p) - b_n(r) = \sum_{i=0}^{n} (u_{[i+1]} - u_{[i]})(n+1)(\mathrm{P}(B_{n,p} = i) - \mathrm{P}(B_{n,r} = i))$$

where $B_{n,p} \sim \mathrm{Bin}(n,p)$ and the $u_{[i]}$ are ordered outcomes of the standard uniform distribution. This equation can be considered as the outcome of

$$\frac{\sum_{i=0}^{n} E_i(\mathrm{P}(B_{n,p} = i) - \mathrm{P}(B_{n,r} = i))}{\sum_{i=0}^{n} E_i/(n+1)}$$

with $E_0, \ldots, E_n$ a random sample from the standard negative exponential distribution and the denominator is practically 1. The expectation of the numerator is

$$\sum_{i=0}^{n} \mathrm{P}(B_{n,p} = i) - \sum_{i=0}^{n} \mathrm{P}(B_{n,r} = i) = 0,$$

and the variance is

$$\begin{aligned}
& \sum_{i=0}^{n} \left(\mathrm{P}(B_{n,p} = i) - \mathrm{P}(B_{n,r} = i)\right)^2 \\
= \;& \mathrm{P}(B_{n,p}^{(1)} = B_{n,p}^{(2)}) + \mathrm{P}(B_{n,p}^{(1)} = B_{n,r}^{(2)}) - 2\mathrm{P}(B_{n,p}^{(1)} = B_{n,r}^{(1)}) \\
\approx \;& \frac{1}{\sqrt{4\pi np(1-p)}} + \frac{1}{\sqrt{4\pi nr(1-r)}} - 2\frac{\exp\left[-\frac{1}{2}\left(\frac{p-r}{p(1-p)+r(1-r)}\right)^2\right]}{\sqrt{2\pi n(p(1-p)+r(1-r))}} \\
= \;& \frac{1}{\sqrt{4\pi n}}\left(\frac{1}{\sqrt{p(1-p)}} + \frac{1}{\sqrt{r(1-r)}} - \frac{2\sqrt{2}}{\sqrt{p(1-p)+r(1-r)}}\right) + \\
& + \frac{2}{\sqrt{4\pi n}} \frac{\exp\left[-\frac{1}{2}\left(\frac{p-r}{p(1-p)+r(1-r)}\right)^2\right] - 1}{\sqrt{p(1-p)+r(1-r)}}.
\end{aligned}$$

For the first terms, see the proof of Lemma 3.1. The last term follows from

$$B_{n,p}^{(1)} - B_{n,r}^{(1)} \sim \mathcal{N}\left(n(p-r),\ n(p(1-p)+r(1-r))\right).$$

As $p = F_n(x)$ tends to $r = F(x)$, the entire expression for the variance is, indeed, $o(n^{-1/2})$ □

If the restriction $\Psi = F$ is abandoned then the methods of proof no longer apply, since they depend on the interpretation of $b_n(p) - b_n(r)$ as outcome of the random variable specified.

Under certain regularity conditions for ψ (and f), it might be possible to prove a similar theorem for the general case, but we were unable to accomplish such result.

However, our theory is developed for situations where the practical researcher has some good intuition about f. Numerical analyses (Section 3.7) suggest that in those situations our method provides a useful contribution to existing density estimation theory. In these cases, and especially when Ψ and F are not too irregular, the behaviour of our density estimate will not be much different from that for the ideal situation, and the, now approximate, standard-error

$$\frac{f(x)}{\sqrt[4]{4\pi nF(x)(1-F(x))}}$$

will usually suffice.

3.5 Comparison with other methods

Comparison with kernel estimation methods
If one compares f_n with the 'usual' kernel estimates k_n (see, e.g., SILVERMAN, 1986), then there are some theoretical advantages of f_n but these are outweighed by practical advantages of k_n, even if the initial guess ψ of f is reliable (DE BRUIN ET AL., 1999). Some theoretical arguments are as follows.

The asymptotic distribution of f_n is studied via that of b_n (see Sections 3.3 and 3.4). This results in the approximation

$$\lim_{n\to\infty} \mathcal{L}\, n^{1/4}(f_n(x) - f(x)) \approx \mathcal{N}(0, \sigma^2(x))$$

where

$$\sigma(x) = \frac{f(x)}{\sqrt[4]{4\pi F(x)(1-F(x))}}.$$

The approximation-sign can be replaced by an equality-sign if $\Psi = F$. The rate of convergence $n^{-1/4}$ is not very satisfactory. Furthermore, the estimate f_n is not sufficiently smooth: the error ${f_n}' - f'$ in the derivative does not vanish, it is of order $O(n^{1/4})$ (Section 3.4). The expected L_1-error of f_n is of the order $n^{-1/4}$ because, using the notation $Z \sim \mathcal{N}(0,1)$, we have

$$\begin{aligned}
\mathbf{E} \int |f_n(x) - f(x)|\, \mathrm{d}x &= \int \mathbf{E}\, |f_n(x) - f(x)|\, \mathrm{d}x \\
&\approx n^{-1/4} \mathbf{E}\, |Z| \int \sigma(x)\, \mathrm{d}x \\
&= n^{-1/4} \sqrt{\frac{2}{\pi}} \frac{1}{\sqrt[4]{4\pi}} \int \frac{\mathrm{d}F(x)}{\sqrt[4]{F(x)(1-F(x))}} \\
&= n^{-1/4} 2(\Gamma(3/4))^2 \pi^{-5/4} \\
&= .72 n^{-1/4}
\end{aligned}$$

Note that this asymptotic behaviour does neither depend on f nor on the initial guess ψ. DE BRUIN ET AL. (1999) provides simulation results in the case that f follows the Beta(2,1)-density. These results are compared with the above approximate values in Table 3.1. It follows from this table that standard errors based on $\sigma(x)$ are, perhaps, a bit conservative.

Kernel estimators seem to perform better. The rate of convergence of kernel estimates with optimal bandwidths (this requires knowledge of the unknown f) is such that $n^{2/5}(k_n(x) - f(x))$ has a limiting distribution, while ${k_n}' - f'$ vanishes in probability as $n \to \infty$, at least under regularity conditions (see, e.g., SILVERMAN, 1978, and SCHUSTER, 1969). As knowledge of f cannot be presupposed, DE BRUIN ET AL. (1999, Figure 4) performed an extensive numerical comparison between f_n and

n	10	13	25	50	75	100	150	200
DE BRUIN ET AL.	.31	.31	.27	.23	.22	.20	.19	.18
$.72n^{-1/4}$	.41	.38	.32	.27	.24	.23	.21	.19

Table 3.1: Comparison between the numerical results and the theory

kernel estimates k_n with bi-weight kernels and bandwidths based on likelihood cross-validation. This provided (for a special density f) that the L_1-error of k_n is significantly smaller than that of f_n, on the average. If the L_1-error of f_n can be decreased by about 15% then k_n and the modification of f_n would be about equally good, on the average.

Comparison with Bernstein polynomial estimations of the density function
VITALE (1975) introduced a Bernstein polynomial approach to estimate the density function directly (instead of via the quantile function). Basically, his method provides a Bernstein expansion of the histogram. After transformation to the unit interval, the estimate of Vitale is defined by

$$\tilde{f}_n^{(m)}(x) = \frac{(m+1)}{n} \sum_{j=0}^{m} h_n^{(m)}(j) \binom{n}{j} x^j (1-x)^{m-j}$$

where $h_n^{(m)}(j) = \#\{x_i \in [\frac{j}{m+1}, \frac{j+1}{m+1}]\}$. This representation is a linear combination of beta densities with random coefficients based on the observations. The parameter m is used as the degree of the polynomial expansion, and thus works as a smoothing parameter. In the next section we shall introduce a different method for smoothing, U-statistic symmetrization. Vitale shows that, under the weak assumptions that f is bounded and $f''(x)$ exists, the optimal choice (in the sense of minimal average mean squared error) of m is $m \sim n^{2/5}$. This follows from

$$\begin{aligned} \text{Bias}(x) &\sim \frac{1}{m} \frac{\tilde{f}'(x)(1-2x) + \tilde{f}''(x)x(1-x)}{2} \\ \text{Var}\,(x) &\sim \frac{m^{1/2}}{n} \frac{\tilde{f}(x)}{2\sqrt{\pi x(1-x)}} \end{aligned}$$

and trying to minimize the rate of $\text{bias}^2(x) + \text{Var}\,(x)$, yielding a convergence rate of the mean squared error of $n^{-4/5}$ for $0 < x < 1$ and $n^{-3/5}$ in the endpoints.

PETRONE (1999) suggested Bayesian methods using a Bernstein polynomial prior with alternative weight functions $h_n^{(m)}$ and GHOSAL (2001) derived the asymptotic results that if the real density f is a Bernstein polynomial, then the convergence rate of the posterior is equal to $n^{-1/2} \log n$, otherwise this rate is $n^{-1/3}(\log n)^{5/6}$ (under the same weak assumptions as in VITALE, 1975).

3.6 Improvements through *U*-statistic symmetrization

How to improve f_n, that is the key question of this section. Theoretical arguments behind B_n, b_n and, hence, f_n are spelled out in Section 3.2 and DE BRUIN ET AL. (1999). A practical argument in favor of these estimates is the strict positivity of b_n. This should certainly not be given up.

A breakthrough appeared when we tried to compute B_n, b_n and f_n for large values of n. To overcome numerical difficulties it was decided to split the sample into two subsamples of size $\frac{1}{2}n$ and to take the arithmetic average of these estimates. This average turned out to be smoother and more accurate than the estimate based on the entire sample. This experience is in line with the asymptotic distribution of f_n and it formed the basis of the U-statistic symmetrization explained below, after some preparations.

A natural way of smoothing is to partition the sample into $k = m^{-1}n$ sub-samples of size m each (for the sake of simplicity, assume $n = km$), and to take the arithmetic average of the resulting estimates: for each subsample $y_1^{(h)}, \ldots, y_m^{(h)}, (h = 1, \ldots, k)$ the density estimate $f_{m,h}$ is derived from

$$B_m(p|y_1^{(h)}, \ldots, y_m^{(h)}), \quad h = 1, \ldots, k$$

in the same way as f_n was derived from B_n. Define $\bar{f}_{n,m}$ as the average of the k estimates $f_{m,h}$. The approximation

$$\mathcal{L}\, m^{1/4}(f_{m,h}(x) - f(x)) \approx \mathcal{N}(0, \sigma^2(x))$$

with $\sigma(x)$ defined in the beginning of Section 3.5, implies that

$$\mathcal{L}\, m^{1/4}(\bar{f}_{n,m}(x) - f(x)) \approx \mathcal{N}(0, k^{-1}\sigma^2(x)) = \mathcal{N}(0, n^{-1}m\sigma^2(x))$$

or, equivalently, that the standard error of $\bar{f}_{n,m}(x)$ is $m^{-1/4}k^{-1/2}\sigma(x)$, which equals $m^{1/4}n^{-1/2}\sigma(x)$, i.e. $k^{-1/4}$ times that of $f_n(x)$. This suggests that the standard error is reduced by 15% or more if $(\frac{m}{n})^{1/4} < .85$ or, equivalently, if m is less than $.53n$ (see the remark in the previous section). Note that a special case appears when the sample is split into $m = \sqrt{n}$ equal parts, the resulting estimate $\bar{f}_{\sqrt{n},\sqrt{n}}$ is consistent and converges with rate $n^{-3/8}$ (if $f = \psi$).

To remove the permutation dependence (and the assumption that $k, m \in \mathbb{N}$), a U-statistic symmetrization will be applied. It will have the effect that the degree n of the polynomial expression $b_n(p)$ is lowered to the degree m of the estimate $b_n^{(m)}(p)$ to be derived. The positivity of the quantile density estimate will not be affected.

Let $B_m(p|y_1, \ldots, y_m)$ denote the polynomial approximation introduced in the previous section, applied to the m observations $y_1, \ldots, y_m$ (with $y_i = \Psi(x_i)$). Note that the degree of this polynomial is $m + 1$. The U-statistic

$$B_n^{(m)}(p) = \binom{n}{m}^{-1} \sum_{1 \le \alpha_1 < \ldots < \alpha_m \le n} B_m(p|y_{\alpha_1}, \ldots, y_{\alpha_m})$$

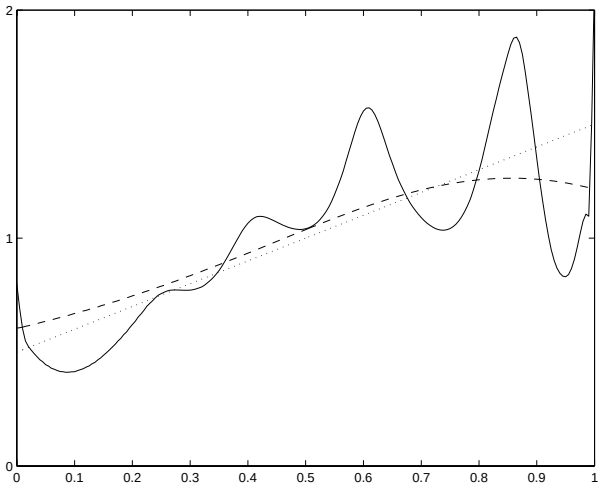


Figure 3.3: Visualization of the improvement discussed in Section 3.6.

is the estimate of $B(p)$ of interest. It can be rewritten as the L-statistic

$$p^{m+1} + \sum_{j=1}^{m} \binom{m+1}{j} p^j (1-p)^{m+1-j} \sum_{i=j}^{n-m+j} \frac{\binom{i-1}{j-1}\binom{n-i}{m-j}}{\binom{n}{m}} y_{[i]}.$$

So, $B_n^{(m)}$ is the average of the quantile function estimates based on all size-m subsets of the sample. From this we get $F_n^{(m)} = {B_n^{(m)}}^{-1} \circ \Psi$, $f_n^{(m)} = {F_n^{(m)}}'$, etcetera. For $m \to \infty$, the $B_m(p|\ldots)$ are consistent estimators of $(\Psi^{-1} \circ H)(p)$. For m fixed, $B_n^{(m)}(p)$ is an unbiased estimate of $\mathbf{E}\,(B_m(p|X_1, \ldots, X_m))$ which, of course, depends on the underlying distribution function of X_1. In principle the (exact) theory of HOEFFDING (1948) about U-statistics is applicable but the complexity of the formulas and the dependence on the unknown f, F, H, etcetera, of the asymptotic variances precludes application in practice. It is reasonable to conjecture that the rate of convergence is better than $n^{-3/8}$ if $k = n^{1/2}$ is taken.

The bias of $f_n^{(m)}$ will be larger than that of the original estimator $f_n(x)$. For a good comparison between this and other methods, one has to look at, e.g., the Mean (Integrated) Squared Error. We rely on the computational experiences to be discussed in Section 3.7, because theoretical analyses are both too complicated and too much dependent on 'irrelevant asymptotics'. In practice, m is chosen to be less than, say 10 or 20, and asymptotic theory becomes questionable in such cases.

The example in Figure 3.3 illustrates the errors involved in using f_n and $f_n^{(m)}$ and suggests that improvement indeed is gained by lowering the degree of the estimating

	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
0	.000	.050	.100	.150	.200	.250	.300	.350	.400	.450	.500
1	.025	.043	.067	.100	.132	.165	.196	.233	.272	.299	.336
2	.040	.047	.057	.078	.100	.124	.146	.177	.210	.225	.256
3	.051	.051	.057	.068	.086	.103	.119	.146	.176	.181	.208
4	.059	.056	.060	.067	.080	.094	.104	.128	.156	.155	.179
5	.066	.006	.063	.068	.078	.091	.096	.118	.144	.137	.160
6	.072	.064	.067	.070	.079	.090	.093	.112	.136	.125	.147
7	.078	.068	.070	.073	.080	.090	.091	.108	.130	.117	.139
8	.082	.071	.073	.075	.081	.092	.091	.107	.127	.113	.133
9	.086	.074	.076	.078	.084	.094	.092	.106	.125	.110	.130
10	.090	.077	.078	.081	.086	.096	.093	.107	.124	.108	.128
11	.094	.080	.081	.084	.088	.098	.095	.107	.124	.107	.127
12	.097	.083	.084	.087	.090	.100	.097	.108	.124	.107	.126
13	.100	.086	.086	.089	.092	.103	.099	.109	.124	.108	.125
14	.103	.089	.089	.092	.094	.105	.101	.110	.125	.108	.125
15	.105	.091	.091	.094	.096	.107	.103	.110	.126	.109	.125

Table 3.2: Average total variation distances for 50 replications, for different choices of a (horizontal axis) and m (vertical), with a sample size of $n = 100$. Underlined are the optimal choices for m.

polynomial. A sample of size $n = 100$ is taken from the distribution on $(0, 1)$ with density $f(x) = \frac{1}{2} + x$ (the dotted line). Using the uniform density as initial guess, the densities $f_n^{(m)}$ $(m = 0, \ldots, 100)$ were computed and those corresponding to $m = 6$ (dashed line) and $m = 100$ (solid line) are displayed.

3.7 Simulation studies

In the current context it is obvious that m should depend on the sample size and on the reliability of the initial guess. To investigate the structure of this dependence, extensive Monte Carlo simulations have been performed. Results will be presented and discussed in this section.

Given some density f, consider (estimates of) the expected value of the total variation distance $||f_n^{(m)} - f||_1$ as a function of m. The optimal m, i.e. the value for which this expected L_1-distance is minimum, depends on the sample size and on the reliability of ψ. This optimal m is determined using the special densities

$$f_a(x) = (1 - a) + 2ax \quad |a| \leq 1, 0 \leq x \leq 1,$$

and the standard-uniform density $\psi(x) \equiv 1$ as initial guess. Figure 3.3 provides a visualization for $a = \frac{1}{2}$ (for explanation, see the last paragraph of the previous section).

	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00
25	0	0	1	2	2	3	4	5	5	6	7
50	0	1	2	3	4	5	6	7	8	9	10
75	0	1	2	3	4	6	7	9	10	11	11
100	0	1	2	4	5	7	8	10	11	13	14
150	0	2	3	5	7	8	10	12	14	16	17
200	0	2	4	6	8	10	12	14	16	18	20
250	0	2	4	6	9	11	13	15	17	19	21

Table 3.3: Optimal m for the cases f_a with various a (horizontal axis) and n (vertical).

Computations were performed for a variety of values of n, as well of a. Variations in a lead to variations in the reliability of the initial guess through the relation $||f_a - \psi||_1 = \frac{1}{2}|a|$.

Table 3.2 displays results involving the choice $n = 100$. For all a in the table, a random sample of size 100 is drawn from f_a. Density estimates $f_{100}^{(1)}, \ldots, f_{100}^{(15)}$ are constructed, and total variation distances $||f_a - f_{100}^{(\cdot)}||_1$ are calculated. This is repeated 49 times, and reported are, for each pair (m, a), the average L_1-errors for the $N = 50$ replications. For example, for $f_{.4}(x) = .6 + .8x$ and a sample of size 100, a choice of $m = 5$ seems best, since the average total variation distance is lowest for $m = 5$.

Similar analyses have been performed for a variety of other choices of n. For each pair (n, a) in Table 3.3, a random sample (of size n) is drawn from f_a. For all possible values of m ($m = 0, 1, \ldots, n$), $||f_a - f_n^{(m)}||_1$ is calculated. This process is replicated 249 times and Table 3.3 displays those m for which the average total variation distance between true and estimated density was smallest. These simulations suggest that the optimal m is approximately proportional to $n^{1/2}$ and $||f_a - \psi||_1$. The expression $m^* = 1.3n^{1/2}a = 2.6n^{1/2}||f_a - \psi||_1$ seems reasonable. The densities f_a studied were rather regular with only one sign change of $f_a - \psi$.

Note that the quantile densities corresponding to the f_a are not polynomials. In case $f_{1/2} = \frac{1}{2} + x$ we have a corresponding quantile density $2/\sqrt{1 + 8p}$, which is not a polynomial though, of course, (low degree) polynomial approximations can be fairly accurate. To study the choice of m in alternative situations, a second simulation study was carried out, involving five examples, all with the unit interval as support, $\psi(x) = 1$, and $||f - \psi||_1 = .25$. The examined densities are

$$\begin{aligned} f_{\mathrm{I}}(x) &= .5 + x \\ f_{\mathrm{II}}(x) &= 1.5 - |2x - 1| \\ f_{\mathrm{III}}(x) &= 1.616 - 2.464x(1 - x^2) \\ f_{\mathrm{IV}}(x) &= 1.091 - .427\sin(4.712x) \\ f_{\mathrm{V}}(x) &= .741 + 1.301\ \mathrm{e}^{-5x}, \end{aligned}$$

m	0	4	5	6	7	8	10	15	20	30
f_{I}	.250	.096	.091	.089	.089	.090	.093	.097	.106	.122
f_{II}	.250	.182	.167	.155	.145	.138	.129	.122	.124	.134
f_{III}	.250	.184	.168	.156	.146	.138	.128	.123	.125	.134
f_{IV}	.250	.147	.138	.132	.127	.124	.119	.116	.119	.130
f_{V}	.250	.104	.096	.092	.090	.089	.090	.098	.107	.124
theor.	.250	.102	.108	.113	.117	.121	.128	.141	.152	.169

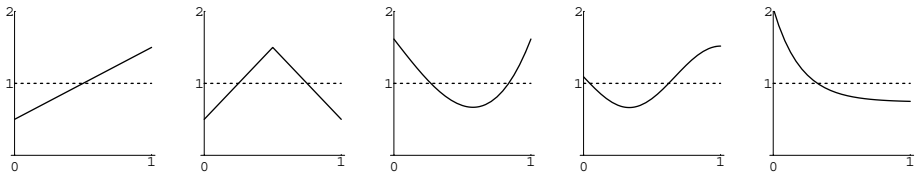
Table 3.4: Performance of $f_{\mathrm{I}}, \ldots, f_{\mathrm{V}}$ and a theoretical approximation. Displayed values are $\frac{1}{R}\Sigma||f_{\cdot} - f_{\cdot,n}^{(m)}||_1$, the underlined values are the minima for each f.

see Figure 3.4. We believe that these densities are 'representative' for the distributions occurring in practice. Analyzing these five densities will give insight in the effect of the type of density on the optimal m. For cases I and V, deviations $\psi(x) - f(x)$ are more regular (involving only one crossing of 0) than for the cases II to IV where two crossings occur. This interpretation of 'representativity' is in line with NEYMAN (1937) where 'smooth alternative probability density functions were [...] presented [...] in terms of probability density functions that have few intersections with the null probability density function' (RAYNER AND BEST, 1989, p. 45). It is reasonable to expect that the optimal m is smallest in cases I and V.

To quantify the optimal m in these 5 cases, a sample of size $n = 100$ is drawn from each density, and the corresponding $f_{100}^{(m)}$ and L_1-errors $||f_{100}^{(m)} - f||_1$ are computed for all possible values of m ($m = 0, 1, \ldots, 100$). This process is performed $R = 500$ times. The average L_1-error, for a deliberate choice of values of m, is reported in Table 3.4. The last row of this table contains the values $.72m^{1/4}n^{-1/2}$, mentioned in Table 3.1. These values provide a crude approximation to the expected L_1-error, especially those for the optimal m (the italicized ones), since they approximate the average L_1-error of the obviously less accurate estimates $\bar{f}_{n,m}$ studied in Section 3.6. The case $m = 0$ involves a degeneracy. If $m = 0$ is actually used, then the value .250 appears as the constant true value.

It follows from the first row of Table 3.4 that, in case $f = f_{\mathrm{I}}$, the expected L_1-error is minimum if m is 6 or 7 (the fourth row of Table 3.3 provides the value 7, take

Figure 3.4: From left to right: densities f_{I} up to f_{V}, discussed in Section 3.6



$a = \frac{1}{2}$). The second, third and fourth row provide that m should be about twice as large if the behaviour of f is less regular. Row five shows, as expected, that m must be chosen substantially smaller if these f's are replaced by the more regular one f_{V}. Both simulation studies taken together indicate that m should depend on the sample size n, the value v of $||f - \psi||_1$ where optimality is required, and the (expected) regularity of f. The latter will be quantified by the symbol w, where w is equal to 1 if the number of sign changes of $\psi - f$ is equal to 1, like in the cases of f_a, f_{I} and f_{V}, and is equal to 2 if two or more sign changes are occurring (cases f_{II} to f_{IV}).

Conclusion
Our simulation studies suggest that the optimal m is approximately equal to

$$m^* = 2.6\, n^{1/2}\, v\, w$$

Note that $m^* = 6.5$ in the case $n = 100$, $v = .25$, $w = 1$ studied in the fourth row of Table 3.3 and in the first row of Table 3.4. The choice $m^* = 13$ is indicated for f_{II}, f_{III} and f_{IV}. In practice f is unknown and the research worker will have to choose the value v of $||f - \psi||_1$ and the value $w \in \{0, 1\}$ where 'optimality is required'.

Robustness
As a final remark, we mention that $||f_n^{(m_1)} - f_n^{(m_2)}||_1$ is very small if m_1 and m_2 are 'not much different', as is exemplified in Figure 3.5 and in Table 3.5. This is, of course, of considerable importance, because it would have been unfortunate if minor variations in the smoothing parameter m induce major variations in the resulting density estimate.

Figure 3.5 indicates this: using one random sample of size $n = 100$ from the Beta$(3, 2)$-distribution and the initial guess $\psi \sim$ Beta$(3, 3)$, density estimates $f_{100}^{(9)}, \ldots, f_{100}^{(13)}$ are constructed. These estimates with smoothing parameters $m = 9, \ldots, 13$ are 'practically identical' in the light of the true values which are $O(1)$.

Another validation of our suggestion is in Table 3.5. A random sample of size $n = 100$ is drawn from $f(x) = \frac{1}{2} + x$ on $(0, 1)$, with $\psi(x) = 1$. On basis of this sample, estimates $f_{100}^{(m)}$, with m=4, 6, 7, 8, 12, 20 are constructed. This process is repeated 249 times, the averages of the R=250 L_1-distances $||f_{100}^{(m_1)} - f_{100}^{(m_2)}||_1$ are shown.

Table 3.5: Indication of the robustness. See the text for full details.

	4	6	7	8	12	20
4	.000	.025	.033	.041	.063	.090
6		.000	.009	.017	.039	.068
7			.000	.008	.031	.060
8				.000	.023	.054
12					.000	.031

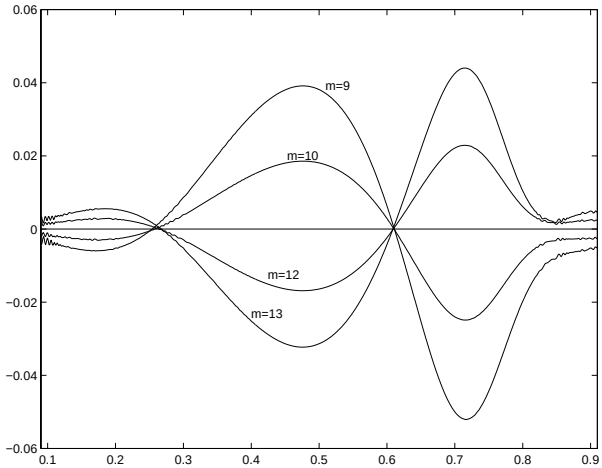


Figure 3.5: Indication of the robustness of the choice of m. Displayed are the differences $f_{100}^{(9)} - f_{100}^{(11)}, \ldots, f_{100}^{(13)} - f_{100}^{(11)}$.

Another desired robustness property is that minor deviations in *initial guess* should not lead to major variations in the induced density.

3.8 Specifying the initial guess

Application of the previous sections requires the specification of an initial guess ψ and the choice of m. Our theory is based on the assumption that the initial guess ψ of the true but unknown density f is made a priori. This assumption is not necessarily unrealistic. Suppose, for example, that the distribution under 'standard conditions' is fairly well known and that the distribution under certain non-standard conditions is required. If the deviations are expected to be fairly small, then the distribution under standard conditions can be used as initial guess.

In other situations some 'data peeping' cannot be dispensed with. A standard situation is that where the distribution has support $(a, b) = (-\infty, \infty)$ and classical statisticians like to make normality assumptions but do not have reliable prior information about expectation μ and variance σ^2. These true values will then be estimated from the sample, preferably by using the sample mean $\bar{x}$ and the sample variance s^2 because the underlying estimators are best unbiased if the normality holds.

The theory of nonparametric density estimation is motivated by the idea that, in practice, parametric models like that discussed above, are never exactly true. They may provide useful approximations but one would like to adapt these, especially if the data suggest that the assumptions are false. That is what we tried to do in the preceding sections by 'fine-tuning' the initial guess ψ.

Though one should not forget that the theory was based on the assumption that the initial guess was made up a priori, it is a matter of 'good statistical practice' to discuss some methods to have a look at the data and to adapt the initial guess if necessary. We already discussed the possibility to use the $\mathcal{N}(\bar{x}, s^2)$ distribution as initial guess. It is well known that this choice maximizes the Shannon entropy $-\int f(x) \log f(x)\, \mathrm{d}x$ under the restriction that f is a probability density on $(-\infty, \infty)$ with $\int x f(x)\, \mathrm{d}x = \bar{x}$ and $\int (x - \bar{x})^2 f(x)\, \mathrm{d}x = s^2$. This *maximum entropy* approach allows some generalization which we like to discuss.

Maximum Entropy approach

For explanatory works on maximum entropy, see JAYNES (1996), GOLAN ET AL. (1996), ZELLNER AND HIGHFIELD (1988) (and Section 2.4). Shannon's entropy is 'an axiomatic method to define a unique function to measure the uncertainty of a collection of events' (GOLAN ET AL., 1996). and is defined as the measure

$$H(f) \equiv -\int_a^b f(u) \log(f(u))\, \mathrm{d}u$$

in the continuous case. When no initial information is present other than the finiteness and specifications of a and b, the entropy measure reaches its maximum when $f(x) = 1/(b-a)\mathbf{1}_{(a,b)}(x)$. If $a = -\infty$ or $b = \infty$ (or both) then, obviously, $\max_f H(f) = \infty$. When a priori knowledge is presented in the form of postulated values for expectations such as $\int x f(x)\, \mathrm{d}x = \bar{x}$ and $\int x^2 f(x)\, \mathrm{d}x = s^2 + \bar{x}^2$, the maximum entropy approach will yield the distribution f that maximizes the entropy $H(f)$ subject to the constraints given by the knowledge presented. We already mentioned that the density of $\mathcal{N}(\bar{x}, s^2)$ appears. According to Jaynes, this distribution 'can be realized in the greatest number of ways consistent with what we know' (GOLAN ET AL., 1996).

Example

Suppose the following information is available. The support of the unknown density of interest f is $(0, 10)$, and the first moment is $\mathbf{E}\ f(x) = 6.3$. We like to take the maximum entropy estimate of f as initial guess ψ. It is constructed by maximizing

$$-\int_0^{10} \psi(x) \log(\psi(x))\, \mathrm{d}x$$

subject to constraints

$$\begin{aligned}
\psi(x) &\geq 0 && \text{nonnegativity of } \psi \\
\int_0^{10} \psi(x)\, \mathrm{d}x &= 1 && \text{total probability 1} \\
\int_0^{10} x\psi(x)\, \mathrm{d}x &= 6.3 && \mathbf{E}\ \psi(x) = 6.3.
\end{aligned}$$

The 'MaxEnt' density $\psi(x) = \mathrm{e}^{\cdot 162x}/25$ obtained (by constructing the Lagrangian function, and setting the partial derivatives zero, see Section 2.4) may be used as initial guess of the density f.

Conclusion
The maximum entropy principle and, especially, the explicit way its solution can be obtained for a large variety of cases makes it a valuable tool (see, e.g., GOLAN ET AL., 1996, JAYNES, 1996, and ZELLNER AND HIGHFIELD, 1988). In practice, however, if one has a distribution at some interval, say $[0, 1]$, with its first 2 moments prescribed then a standard family, say $\beta(f, g)$, may also be used (it would correspond to the maximum entropy method if $\beta_1(u) = \log u$ and $\beta_2(u) = \log(1-u)$, and $\mathbf{E}(\log X)$ and $\mathbf{E}(\log(1-X))$ are prescribed). On Jaynes and maximum entropy, ZABELL (1988) states:

> 'The right-wing totalitarians [...] believe that once an axiom system is adopted, one must accept without question every consequence that flows from it. One searches within one's heart, discovers the basic properties of belief and inference, christens them axioms, and then all else follows as logical consequence. [...] One representative of this position is E.T. Jaynes, who dates his adherence to Bayesianism to the time when he encountered Cox's axiomatization of epistemic probability, and who views the Shannon axioms for entropy as an unanswerable foundation for his method of maximum entropy. This position errs in giving the axioms too distinguished a position.'

What if the true density is normal?
Returning to the general issue of specifying the 'inputs' (ψ and m) of our nonparametric density estimate (Section 3.6) and restricting the attention to $a = -\infty$ and $b = \infty$, the idea of using a normal density as initial guess sounds plausible.

The sample values $x_{[1]}, \ldots, x_{[n]}$ can be used to construct an initial guess Ψ in the form of a normal distribution with parameters estimated from the sample. The density of $\mathcal{N}(\bar{x}, s^2)$ seems natural in this respect, though alternative estimates[6] are not necessarily worse. The issue of interest is whether our estimates $f_n^{(m)}$ provide a satisfactory method for fine-tuning such normal a priori guess. The choice of m is of particular interest in this case. To make this choice, a specification of $v = ||f - \psi||_1$ is needed. It is interesting to study the L_1 distance if f *truly* is a $\mathcal{N}(\mu, \sigma^2)$ density and ψ is its estimate based on $\hat{x}$ and s^2. In practice the true density f will not be of this normal kind and, hence, the values of m to be suggested for the normal case, should be regarded as lower bounds of the choice of m to be made in practice.

A partial analytic result is provided by using

$$\begin{aligned} ||\psi - f||_1 &= ||\mathcal{N}(\bar{x}, s^2) - \mathcal{N}(\mu, \sigma^2)||_1 \\ &\leq ||\mathcal{N}(\bar{x}, s^2) - \mathcal{N}(\mu, s^2)||_1 + ||\mathcal{N}(\mu, s^2) - \mathcal{N}(\mu, \sigma^2)||_1, \end{aligned}$$

[6] e.g. such that $x_{[1]}$ corresponds to the $(n+1)^{-1}$th quantile and $x_{[n]}$ to the $1-(n+1)^{-1}$th quantile, or such that a normal density is fitted to the quantiles through Normal Probability Plot Regression (see Chapter 5). Of particular interest in our context are the estimators $\hat{f}$ that, given the data, try to minimize $||f - \hat{f}||_1$. We, however, content ourselves with $\mathcal{N}(\bar{x}, s^2)$ as indicated.

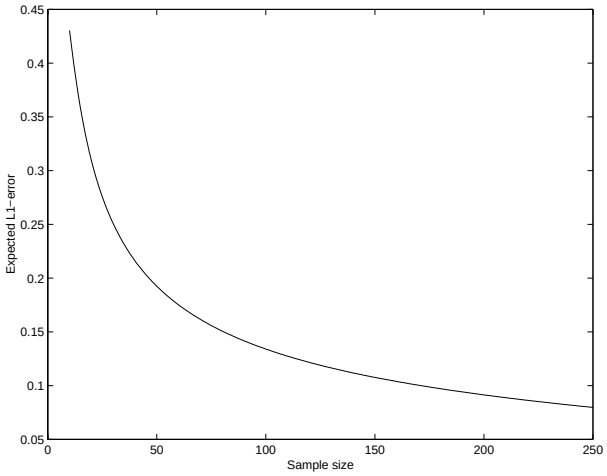


Figure 3.6: Expected L_1-error versus sample size

on account of the triangular inequality, and it is, theoretically, possible to obtain explicit results. We do have the following partial results: in case of unknown μ and known σ^2, the exact result

$$\mathcal{L}\ ||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\hat{\mu},\sigma^2)||_1 = 2\Phi\left(\tfrac{|\hat{\mu}-\mu|}{2\sigma}\right)-1,$$

where, e.g., $\hat{\mu}=\bar{x}$, applies. A similar but more complicated formula holds for $||\mathcal{N}(\mu,s^2)-\mathcal{N}(\mu,\sigma^2)||_1$

$$\mathcal{L}\ ||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\mu,\hat{\sigma^2})||_1 = 2\left|\Phi\left(\frac{a(\sigma^2,\hat{\sigma^2})}{\sigma}\right)-\Phi\left(\frac{a(\sigma^2,\hat{\sigma^2})}{\hat{\sigma}}\right)\right|$$

where $a(\sigma^2,\hat{\sigma^2})=\sqrt{\log\frac{\hat{\sigma^2}}{\sigma^2}\frac{\hat{\sigma^2}\sigma^2}{|\hat{\sigma^2}-\sigma^2|}}$ and $\hat{\sigma^2}$ is, e.g., the sample variance.

More useful might be the asymptotic approximation (BROWN, 1995)

$$\lim_{n\to\infty}\mathrm{P}\left(||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\hat{\mu},\hat{\sigma^2})||_1<\frac{2z_{1-\alpha}}{\sqrt{n\pi\mathrm{e}}}\right)=1-\alpha.$$

When, like in our case, both μ and σ^2 are unknown, BROWN (1995) provides the asymptotic inequality

$$\lim_{n\to\infty}\mathrm{P}\left(||\mathcal{N}(\mu,\sigma^2)-\mathcal{N}(\hat{\mu},\hat{\sigma^2})||_1<\sqrt{\frac{-2\log\alpha}{n}}\right)\geq 1-\alpha.$$

n	10	15	25	50	75	100	250	500	1000
$v_{\min}$	1.39	1.13	.86	.60	.49	.43	.27	.19	.14

Table 3.6: Expected L_1-error versus sample size

and establishes that this holds for *all* two parameter densities with both parameters unknown.

A precise, exact, and explicit result is not attainable. That is why we will establish a value for the total variation numerically (which is quite easily possible).

Taking, without loss of generality, $\mu = 0$ and $\sigma^2 = 1$, standard sampling theory provides $\bar{X} \sim \mathcal{N}(0, \frac{\sigma^2}{n})$ and $(n-1)S^2 \sim \chi^2_{n-1}$, ($\bar{X}$ and S^2 being statistically independent). We computed $||\mathcal{N}(0,1) - \mathcal{N}_(a,b)||_1$ for all (a,b) in a sufficiently precise lattice on $\mathbb{R}$. Following that, we computed $\mathbf{E}_{a=\hat{X}} \mathbf{E}_{b=S^2} ||\mathcal{N}(0,1) - \mathcal{N}(a,b)||_1$. See Figure 3.6 and Table 3.6 for a display and note that the expected L_1-error decreases when the sample size increases. These results are very important. Note that in this case where the initial guess ψ corresponds to a normal density estimated from the data, we have the peculiarity that $v \approx 4.3n^{-1/2}$. The crucial input v in the rule of thumb $m = 2.6\sqrt{n}\ v\ w$ (Section 3.10) should depend on n as well. This provides a lower bound $11w$ for m, independent of the sample size. Recall that $w = 1$ if only one zero of $\psi - f$ is expected, and 2 if the behaviour is less regular Two normal densities, in general, intersect at two points, hence $w = 2$ seems the appropriate choice.

These lower bounds for m and v correspond to the situation where f is truly normal. The researcher has to 'add a value on top of v, or m, describing his confidence in the notion that f is normal'.

3.9 Extending the theory to the bivariate case

It is not straightforward to generalize our univariate density estimate $f_n^{(m)}(x)$ to some bivariate estimate $f_n^{(m)}(x,y)$. The reason is that the concept of quantile distribution does not translate easily to the bivariate case. Associated to this in a semi-parametric way are *copula functions*. For the joint distribution H of X and Y with marginals F and G, there is a so called copula function $C : [0,1]^2 \to [0,1]$ such that

$$H(x,y) = C(F(x), G(y)), \quad \forall x, y.$$

References in this field are SCHWEIZER AND SKLAR (1983), MARSHALL (1996), and SKLAR (1996). ALBERS (1998) developed a semi-parametric model to estimate bivariate densities under the assumption of some form of positive association, e.g. positive quadrant dependence (see LEHMANN, 1966), based on the techniques in DE BRUIN ET AL. (1999). It is possible to construct a copula such that the $H(x,y)$ inferred is indeed positive quadrant dependent, e.g. $H(x,y) = \min(F(x), G(y))$ (MARSHALL,

1996). This method was not very successful, partly due to the wiggling behaviour of the (marginal) density estimates, and partly because the over-simplification in taking only one parameter to model dependence.

Another approach is that of TENBUSCH (1994) who extended the uni-variate Bernstein polynomial estimation method of VITALE (1975, see Section 3.5) to the bivariate case. When (X, Y) is defined on the unit square, the density estimate is

$$\tilde{f}_n^{(m)}(x,y) = \frac{(m+1)^2}{n} \sum_{i=0}^{m} \sum_{j=0}^{m} h_n^{(m)}(i,j) \binom{n}{i} x^i (1-x)^{m-i} \binom{n}{j} y^j (1-y)^{m-j}$$

with $h_n^{(m)(i,j)} = \#\{(x_i, y_i) \in [\frac{i}{m+1}, \frac{i+1}{m+1}] \times [\frac{j}{m+1}, \frac{j+1}{m+1}]\}$. Tenbusch claims that for this method the quality of the estimators is comparable with the quality of the usual bivariate kernel estimators. A drawback of his method is that it is seriously affected by order-preserving transformations of the data.

To overcome the drawbacks of Albers's original method and of Tenbusch's method, the following approach might be of interest. Consider that the problem is reduced to that of estimating the densities of a large number of well-chosen linear combination. To be specific, if the bivariate data suggests a normal initial guess, then we would prefer to transform the data linearly such that the initial guess corresponds to the $\mathcal{N}_2(0_2, I_2)$ distribution. Next consider J linear combinations

$$(\bar{x}_i, \bar{y}_i) \mapsto (\cos\varphi_j)\bar{x}_i + (\sin\varphi_j)\bar{y}_i \quad \text{where } \varphi_j = j\tfrac{\pi}{n}, \quad j = 0, 1, \ldots, J-1, \quad \forall i.$$

For each linear combination $\phi = \mathcal{N}(0,1)$ is used as the initial guess. Next, applying the procedures of this chapter, we obtain J estimated densities for these linear combinations. These marginal densities can then be combined into one estimate $h_n^{(m)}(x,y)$ of $h(x,y)$, e.g., by using numerical methods familiar to MRI-scan specialists.

3.10 Discussion

The nonparametric density estimates f_n and $f_n^{(m)}$ require that an initial guess ψ of the true density f is made. In principle, the initial guess should be made *a priori* because the theory is based on this assumption. In practice some data peeping may be necessary, see Section 3.8. The subjectivity involved in the specification of the initial guess remains visible if m is small. If one chooses $m = n$ such that $f_n^{(n)} = f_n$ corresponds to the estimator studied in DE BRUIN ET AL. (1999) , then the data speaks almost exclusively and the information provided by ψ is almost completely ignored. That is why f_n cannot compete with estimates that take more information into account. The situation becomes more interesting, but also more complicated, if one accepts the idea that ψ is 'reliable' in the sense that the difference $||\psi - f||_1$ between ψ and the true but unknown f may be expected to be less than some constant c which is substantially smaller than the upper bound 2. If, e.g., one is willing to believe that $||\psi - f||_1$ is less than .50 then one will try to choose m such that m is

'optimal' if $v = ||\psi - f||_1$ is .25. The rule of thumb $m^* = 2.6n^{1/2}\ v\ w$ mentioned at the end of Section 3.6 will then provide $m^* = 6.5w$ if $n = 100$ and, hence, $m^* = 6.5$ if one sign change of $\psi - f$ is expected. This implies that the 'uniqueness' of f_n is preserved, at least to some extent. This, however, depends on the belief that v is something like .25. In an interesting case from practice (see Section 3.8), values of v 'not less than $4.3n^{-1/2}$ are suggested which in the case $n = 100$ provides that $v \geq .43$ and $m^* = 11$, respectively 22, if $w = 1$, respectively 2. The difference between the m^* values suggested is less important than one might expect. Table 3.5 suggests that the L_1-errors are more affected by the sample one has to evaluate than by the degree m of the estimate $b_n^{(m)}$ of the quantile density.

The 'unique' nonparametric density estimate $f_n^{(m)}$ thus defined has an expected L_1-error which is considerably smaller than that of f_n if the true density f is not too much different from the initial guess ψ. It is natural that $f_n^{(m)}$ will, under these conditions, be a better estimate than any kernel estimate if the bandwidth is based on likelihood cross-validation (because of the additional knowledge incorporated in $f_n^{(m)}$). This was supported by simulations. The situation may change if the kernel and bandwidth are also chosen on the basis of the initial guess ψ. It will depend on the true density f whether the estimate $f_n^{(m)}$, with $m = 2.6n^{1/2}v\ w$ (or if ψ has been obtained by a preliminary parametric analysis of the data, $m = 10w$), has the smallest expected L_1-error or some kernel estimate with a bandwidth determined such that it is optimal for $f = \psi$.

To summarize, we claim (on basis of our simulation experiments) that $f_n^{(m)}$ with $m = .65n^{1/2}w$ (the case $v = .25$), possibly with $w = 1$, is quite reasonable in practice and that the explicit and usable formula

$$m^{1/4}n^{-1/2}\hat{\sigma}(x) = \frac{(.65\ w)^{1/4}}{n^{3/8}} \frac{f_n^{(m)}(x)}{\sqrt[4]{4\pi F_n^{(m)}(x)(1 - F_n^{(m)}(x))}}$$

for the corresponding standard error is not unreasonable.

The theory of this chapter is of particular interest because it tries to combine that what is supposed to be 'good' in the schools of the classical statisticians, the nonparametric statisticians and the Bayesians. Such *multi-modal approach* is always somewhat questionable because arguments can be weighted differently. A peculiar feature is that instead of a restrictive Bayesian a priori distribution, a less demanding initial guess is exploited. This suggests that a relatively new type of statistics is involved.