3

Continuous Models

3.1 Introduction

In the previous chapter we have considered models having densities with respect to the counting measure. In this section we will consider the case of continuous random variables. Let $\mathcal{G}$ represent the class of all distributions having densities with respect to the Lebesgue measure. We will assume that the true, data generating distribution G and the model family $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ belong to $\mathcal{G}$. Suppose that G and F_θ have densities g and f_θ with respect to the Lebesgue measure.

Let $X_1, \ldots, X_n$ be a random sample from the distribution G which is modeled by $\mathcal{F}$, and we wish to estimate the value of the model parameter θ. As in the case with discrete models, our aim here is to estimate the unknown parameter θ by choosing the model density which gives the closest fit to the data. Unlike the discrete case, however, this poses an immediate challenge; the data are discrete, but the model is continuous, so now there is an obvious incompatibility of measures in constructing a distance between the two. One cannot simply use relative frequencies to represent a nonparametric density estimate of the true data generating distribution in this case.

One strategy for constructing a distance in this case can be to consider a histogram of fixed bin width, say h. If the support of the random variable in question is the real line, one would need a countably infinite sequence of such bins to cover the entire support. One can then compute the empirical probabilities for the bins, and minimize their distance from the corresponding model based bin probabilities. This structure can be routinely extended to multidimensions. Usually, though, an artificial discretization of this type would entail a loss of information.

Instead of discretizing the model, another approach could be to construct a continuous density estimate using some appropriate nonparametric density estimation method such as the one based on kernels. In this case, let

$$g_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} K(x, X_i, h_n) = \int K(x, y, h_n) dG_n(y) \tag{3.1}$$

denote a nonparametric kernel density estimator where $K(x, y, h_n)$ is a smooth kernel function with bandwidth h_n and G_n is the empirical distribution function as obtained from the data. Very often the kernel is chosen as a symmetric density with scale h_n, i.e.,

$$K(x, X_i, h_n) = \frac{1}{h_n} w\left(\frac{x - X_i}{h_n}\right)$$

where $w(\cdot)$ is a symmetric nonnegative function satisfying

$$\int_{-\infty}^{\infty} w(x)dx = 1.$$

We can then estimate θ by minimizing a distance based on disparities defined as

$$\rho_C(g_n^*, f_\theta) = \int C(\delta(x)) f_\theta(x)dx,$$

where the Pearson residual $\delta(x)$ now equals

$$\delta(x) = \frac{g_n^*(x) - f_\theta(x)}{f_\theta(x)}.$$

Under differentiability of the model, the estimating equation now has the form

$$-\nabla \rho_C(g_n^*, f_\theta) = \int_x A(\delta(x)) \nabla f_\theta(x)dx = 0, \tag{3.2}$$

where the residual adjustment function $A(\delta)$ is as described in Equation (2.38). Under appropriate regularity conditions, the solution of this equation represents the minimum distance estimator (based on the disparity ρ_C). In spirit, the rest of the estimation procedure is similar to the discrete case. The residual adjustment function $A(\delta)$ and the disparity generating function $C(\delta)$ continue to provide the same interpretation as in the discrete case in terms of controlling probabilistic outliers.

In practice, however, the addition of the kernel density estimation process leads to substantial difficulties. The theoretical derivation of the asymptotic normality of the minimum distance estimators based on disparities and the description of their other asymptotic properties are far more complex in this case. However, under suitable conditions, the minimum distance estimators based on disparities continue to be first order efficient at the model. In this chapter we present the different approaches in which this problem has been addressed by different authors. Historically, the minimum Hellinger distance estimator remains the first estimator within the class of disparities to be developed for the dual purpose of robustness and asymptotic efficiency. Beran (1977) considered minimum Hellinger distance estimation in continuous models. As mentioned before, his contribution significantly influenced future research in this area. Many authors have considered other aspects of this work in continuous models; see Stather (1981), Tamura and Boos (1986), Simpson (1989b), Eslinger and Woodward (1991), Basu and Lindsay (1994), Cao et al. (1995) and Toma (2008). Wu and Karunamuni (2009) and Karunamuni and Wu (2009) have applied and extended the work of Beran (1977) to semiparametric models. Some researchers have considered the extension of the techniques based on the Hellinger distance to other disparities; see Basu, Sarkar, Vidyashankar (1997), Park and Basu (2003), Bhandari, Basu and Sarkar (2006), Broniatowski and Leorato (2006) and Broniatowski and Keziou (2009). Park and Basu (2004) derived general results for a subclass of disparities under certain conditions.

Donoho and Liu (1988a) argued that minimum distance estimators are "automatically robust" in the sense that the minimum distance functional based on a particular metric changes very little over small neighborhoods of the model based on the same metric subject to Fisher consistency; it also has good breakdown properties with respect to such contamination. In particular, the minimum Hellinger distance estimator has the best stability against Hellinger contamination among Fisher-consistent functionals. This is a very powerful result which provides strong justification of the use of the Hellinger distance in many practical problems. However, the approach of Donoho and Liu explicitly uses the properties of a mathematical metric such as the triangle inequality, and hence does not appear to have straightforward extensions to general classes of statistical distances.

Other authors have considered different applications of these methods in more specialized and extended models, some of which we will briefly describe in Chapter 10. Yang (1991) and Ying (1992) have applied the minimum Hellinger distance and related methods to the case of survival data and censored observations. Woodward, Whitney and Eslinger (1995) and Cutler and Cordero-Braña (1996) have applied these techniques to the case of mixture models. Pak (1996) and Pak and Basu (1998) have considered the linear regression problem through this minimum distance approach. Victoria-Feser and Ronchetti (1997) and Lin and He (2006) have extended these methods to the case of grouped data. Sriram and Vidyashankar (2000), among others, have applied the same to different stochastic process applications; Takada (2009) has applied the simulated minimum Hellinger distance estimator in case of stochastic volatility models. Cheng and Vidyashankar (2006) have approached the problem of adaptive estimation through minimum Hellinger distance techniques.

3.2 Minimum Hellinger Distance Estimation

3.2.1 The Minimum Hellinger Distance Functional

To present Beran's proof of the asymptotic normality of the minimum Hellinger distance estimator, we carry on from where we left off in Section 2.4.1 and first establish additional preliminary results about the minimum Hellinger distance functional. We denote by $T(G)$ the minimum Hellinger distance functional corresponding to the distribution $G \in \mathcal{G}$, which is defined by

$$\mathrm{HD}(g, f_{T(G)}) = \inf_{\theta \in \Theta} \mathrm{HD}(g, f_\theta)$$

provided such a minimum exists. Here the Hellinger distance measure between the continuous densities g and f is defined by

$$\mathrm{HD}(g, f) = 2 \int (g^{1/2} - f^{1/2})^2. \tag{3.3}$$

Suppose that the model family $\{\mathcal{F}\}$ is identifiable, and the conditions of Lemma 2.6 are satisfied. Then, for all $G \in \mathcal{G}$, $T(G)$ exists; also if $T(G)$ is unique, and if G_n be a sequence of distributions for which the corresponding sequence of densities g_n converge to g in the Hellinger metric, the minimum Hellinger distance functional $T(G_n)$ converges to $T(G)$ as $n \to \infty$. The continuity of the minimum Hellinger distance functional ensures the consistency of the minimum Hellinger distance estimator when the corresponding kernel density estimate has the appropriate convergence properties.

When the assumptions of Lemma 2.6 do not strictly hold, but those of Lemma 2.7 can be assumed for an appropriately defined subset of $\mathcal{G}$, the above conclusions remain valid for all G in the above subset of distributions.

We use the notation $s_t = f_t^{1/2}$, and as in Section 2.4, we make further assumptions to make the functional T differentiable. For specified $t \in \Theta$ we assume that there exists a $p \times 1$ vector $\dot{s}_t(x)$ with components in L_2 and a $p \times p$ matrix $\ddot{s}_t(x)$ with components in L_2 such that for every $p \times 1$ real vector e of unit euclidean length and for every scalar α in a neighborhood of zero,

$$\begin{aligned} s_{t+\alpha e}(x) &= s_t(x) + \alpha e^T \dot{s}_t(x) + \alpha e^T \gamma_\alpha(x) & (3.4) \\ \dot{s}_{t+\alpha e}(x) &= \dot{s}_t(x) + \alpha \ddot{s}_t(x) e + \alpha v_\alpha(x) e & (3.5) \end{aligned}$$

where $\gamma_\alpha(x)$ is $p \times 1$ vector, $v_\alpha(x)$ is a $p \times p$ matrix, and each component of γ_α and v_α converges to zero in L_2 as $\alpha \to 0$. We work under these conditions to present the results derived by Beran (1977).

Theorem 3.1. [Beran (1977, Theorem 2)]. *Suppose that (3.4) and (3.5) hold for every $t \in \Theta$, $T(G)$ exists, is unique, and lies in the interior of Θ, $\int \ddot{s}_{T(G)}(x) g^{1/2}(x) dx$ is a nonsingular matrix, and the functional T is continuous at G in the Hellinger topology. Then, for every sequence of densities $\{g_n\}$ converging to g in the Hellinger metric, the corresponding sequence of functionals $T(G_n)$ has the expansion*

$$\begin{aligned} T(G_n) = T(G) &+ \int \vartheta_g(x)[g_n^{1/2}(x) - g^{1/2}(x)]dx \\ &+ \xi_n \int \dot{s}_{T(G)}(x)[g_n^{1/2}(x) - g^{1/2}(x)]d(x) \end{aligned} \tag{3.6}$$

where

$$\vartheta_g(x) = -\left[\int \ddot{s}_{T(G)}(x) g^{1/2}(x) dx\right]^{-1} \dot{s}_{T(G)}(x) \tag{3.7}$$

and ξ_n is a real $p \times p$ matrix which tends to zero as $n \to \infty$. When the true distribution belongs to the model, so that $g = f_\theta$ for some $\theta \in \Theta$, we get

$$\begin{aligned}\vartheta_g(x) = \vartheta_{f_\theta}(x) &= -\left[\int \ddot{s}_\theta(x)s_\theta(x)dx\right]^{-1} \dot{s}_\theta(x)\\ &= \left[\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)dx\right]^{-1} \dot{s}_\theta(x)\\ &= \left[\frac{1}{4}I(\theta)\right]^{-1} \dot{s}_\theta(x), \qquad (3.8)\end{aligned}$$

where $I(\theta)$ represents the Fisher information matrix.

Proof. Let $\theta = T(G)$, and $\theta_n = T(G_n)$. As the differentiability conditions in (3.4) and (3.5) hold, the functionals θ and θ_n must satisfy

$$\int \dot{s}_\theta(x)g^{1/2}(x)dx = 0, \quad \text{and} \quad \int \dot{s}_{\theta_n}(x)g_n^{1/2}(x)dx = 0$$

respectively. Using (3.5), we get

$$\begin{aligned}0 &= \int \dot{s}_{\theta_n}(x)g_n^{1/2}(x)dx\\ &= \int [\dot{s}_\theta(x) + \ddot{s}_\theta(x)(\theta_n - \theta) + v_n(x)(\theta_n - \theta)]g_n^{1/2}(x)dx\end{aligned}$$

where the components of the $p \times p$ matrix $v_n(x)$ converge to zero in L_2 as $n \to \infty$. Thus, for n sufficiently large,

$$\begin{aligned}\theta_n - \theta &= -\left[\int (\ddot{s}_\theta(x) + v_n(x))g_n^{1/2}(x)dx\right]^{-1} \int \dot{s}_\theta(x)g_n^{1/2}(x)dx\\ &= -\left[\int \ddot{s}_\theta(x)g^{1/2}(x)]dx\right]^{-1} \int \dot{s}_\theta(x)[g_n^{1/2}(x) - g^{1/2}(x)]dx\\ &\quad + \xi_n \int \dot{s}_\theta(x)[g_n^{1/2}(x) - g^{1/2}(x)]dx\end{aligned}$$

as was to be proved.

Now suppose that $g = f_\theta$ belongs to the model. Clearly $\int \dot{s}_\theta(x)s_\theta(x)dx = 0$ for all θ in the interior of Θ. It then follows that for every sufficiently small α and every unit vector e,

$$\begin{aligned}0 &= \int \alpha^{-1}[\dot{s}_{\theta+\alpha e}(x)s_{\theta+\alpha e}(x) - \dot{s}_\theta(x)s_\theta(x)]dx\\ &= \int \alpha^{-1}\{[\dot{s}_{\theta+\alpha e}(x) - \dot{s}_\theta(x)]s_\theta(x) + [s_{\theta+\alpha e}(x) - s_\theta(x)]\dot{s}_{\theta+\alpha e}(x)\}dx\\ &= \left[\int \ddot{s}_\theta(x)s_\theta(x)dx + \int \dot{s}_\theta(x)\dot{s}_\theta^T(x)dx\right] e + o(1)\end{aligned}$$

which shows $\left[\int \dot{s}_\theta(x)\dot{s}_\theta^T(x)dx\right]^{-1}\dot{s}_\theta(x) = -\left[\int \ddot{s}_\theta(x)s_\theta(x)dx\right]^{-1}\dot{s}_\theta(x)$. The other equalities of (3.8) are obvious. □

Remark 3.1. In actual computation, a kernel density estimator g_n^* with the right properties will be used to estimate the unknown parameter based on $\rho_C(g_n^*, f_\theta)$. In our functional notation, the estimator may be expressed as $T(G_n^*)$. However, G_n^* is simply the convolution of the empirical with the fixed, known, kernel and we will continue to refer to the estimator obtained by minimizing $\rho_C(g_n^*, f_\theta)$ as $T(G_n)$ in our functional notation.

3.2.2 The Asymptotic Distribution of the Minimum Hellinger Distance Estimator

In Theorem 3.1 we have provided the representation of the functional which we will exploit to find the minimum Hellinger distance estimator of the parameter. Suppose we have a random samples $X_1, \ldots, X_n$ from the distribution G, which is modeled by the parametric family $\mathcal{F}$. Beran suggested the use of the kernel density estimate g_n^* of g given by

$$g_n^*(x) = \frac{1}{n(h_n s_n)} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h_n s_n}\right). \tag{3.9}$$

Here w is a smooth density on the real line, and the bandwidth is the product of h_n and s_n, where the quantity s_n is a robust estimator of scale, while the sequence h_n converges to zero at an appropriate rate. The following theorem lists the conditions under which the convergence of the kernel density estimate g_n^* to g in the Hellinger metric is guaranteed.

Theorem 3.2. [Beran (1977, Theorem 3)]. *Suppose that the kernel density estimate is given by (3.9), and the relevant quantities satisfy the following conditions:*

(i) *w is absolutely continuous and has compact support; w' is bounded.*

(ii) *g is uniformly continuous.*

(iii) $\lim_{n\to\infty} h_n = 0$; $\lim_{n\to\infty} n^{1/2}h_n = \infty$.

(iv) *As $n \to \infty$, $s_n \to s$, a positive constant depending on g.*

Then the kernel density estimate g_n^* converges to G in the Hellinger metric (i.e., $\mathrm{HD}(g_n^*, g) \to 0$ as $n \to \infty$). Thus, if T is a functional which is continuous in the Hellinger metric, then $T(G_n) \to T(G)$ in probability.

Proof. Let the empirical distribution based on the random sample $X_1, \ldots, X_n$ be denoted by G_n. Thus, the kernel density estimate in (3.9) may be expressed as

$$g_n^*(x) = \frac{1}{h_n s_n} \int w\left(\frac{x - y}{h_n s_n}\right) dG_n(y).$$

Also let the density $\tilde{g}_n = (h_n s_n)^{-1} \int w[(h_n s_n)^{-1}(x-y)]dG(y)$ represent the kernel smoothed version of the true model density. We will use $\tilde{g}_n$ as the intermediate tool in establishing the convergence of g_n^* to g. Integration by parts gives

$$|g_n^*(x) - \tilde{g}_n(x)| \le n^{-1/2}(h_n s_n)^{-1} \sup_x |R_n(x)| \int |w'(x)|dx,$$

where $R_n(x) = n^{1/2}[G_n(x) - G(x)]$. On the other hand, suppose $a > 0$ is such that the interval $[-a, a]$ contains the support of w, then

$$|\tilde{g}_n(x) - g(x)| \le \sup_{|t|<a} |g(x - h_n s_n t) - g(x)|.$$

From the above two results it is clear that an appropriately defined $g_n^*(x)$ satisfies $\sup_x |g_n^*(x) - g(x)| \to 0$ with probability 1. Since the functions g_n^* and g are densities, it then follows that

$$\int (g_n^{*1/2}(x) - g^{1/2}(x))^2 dx \to 0$$

with probability 1 for such g_n^* as was to be shown. □

Finally, the asymptotic distribution of the minimum Hellinger distance estimator $T(G_n)$ is considered in the next theorem. This requires quite strong assumptions. Several authors have attempted to derive similar results under weaker conditions, some of which are presented in later sections. Here we present a sketch of Beran's original 1977 proof.

Theorem 3.3. [Beran (1977, Theorem 4)]. *Assume the following conditions.*

(i) w is symmetric about 0 and has compact support.

(ii) w is twice absolutely continuous; w'' is bounded.

(iii) The minimum Hellinger distance functional T satisfies (3.6) and ϑ_g has compact support K^* on which it is continuous.

(iv) $g > 0$ on K^*; g is twice absolutely continuous and g'' is bounded.

(v) $\lim_{n\to\infty} n^{1/2} h_n = \infty$; $\lim n^{1/2} h_n^2 = 0$.

(vi) There exists a positive finite constant s depending on g such that $n^{1/2}(s_n - s)$ is bounded in probability.

Then the limiting distribution of $n^{1/2}[T(G_n) - T(G)]$ under g as $n \to \infty$ is

$$N\left(0, \int \psi_g(x)\psi_g^T(x)g(x)dx\right),$$

where $\psi_g(x) = \vartheta_g(x)/(2g^{1/2}(x))$, and $\vartheta_g(x)$ is as defined in Equation (3.7). In particular, if $g = f_\theta$, the limiting distribution of $n^{1/2}[T(G_n) - T(G)]$ is $N(0, I^{-1}(\theta))$.

Sketch of the Proof. Let $T(G) = \theta$ and $T(G_n) = \theta_n$. Under the given conditions, g_n^* converges, in probability, to g in the Hellinger metric. Now from (3.6)

$$T(G_n) = T(G) + \int \vartheta_g(x)[g_n^{*1/2}(x) - g^{1/2}(x)]dx \\ + \xi_n \int \dot{s}_{T(G)}[g_n^{*1/2}(x) - g^{1/2}(x)]dx$$

where $\xi_n \to 0$ in probability. Thus, the asymptotic normality of

$$n^{1/2}(\theta_n - \theta) = n^{1/2}[T(G_n) - T(G)]$$

is driven by the distribution of the term $n^{1/2} \int \vartheta_g(x)[g_n^{*1/2}(x) - g^{1/2}(x)]dx$. Notice that $\int \vartheta_g(x)g^{1/2}(x)dx = 0$, i.e., $\vartheta_g(x)$ is orthogonal to $g^{1/2}(x)$; also $\vartheta \in L_2$. Thus, our required result will be established if it can be shown that the limiting distribution of $n^{1/2} \int \sigma(x)[g_n^{*1/2}(x) - g^{1/2}(x)]dx$, with $\sigma \in L_2$, σ orthogonal to $g^{1/2}$, and σ supported on K^*, is $N\left(0, \int \varsigma(x)\varsigma^T(x)g(x)dx\right)$, where $\varsigma(x) = \sigma(x)/(2g^{1/2}(x))$.

By an application of the algebraic identity

$$b^{1/2} - a^{1/2} = (b-a)/(2a^{1/2}) - (b-a)^2/[2a^{1/2}(b^{1/2} + a^{1/2})^2] \qquad (3.10)$$

we get

$$n^{1/2}\int \sigma(x)[g_n^{*1/2}(x) - g^{1/2}(x)]^2 dx = n^{1/2}\int \sigma(x)[g_n^*(x) - g(x)]/(2g^{1/2}(x))dx + \zeta_n$$

where ζ_n is the remainder term. To complete the proof, one has to show the following two essential steps.

$$n^{1/2} \int \sigma(x)[g_n^{*1/2}(x) - g^{1/2}(x)]dx \\ - n^{1/2} \int \sigma(x)[g_n^*(x) - g(x)]/(2g^{1/2}(x))dx = o_p(1), \qquad (3.11)$$

so that the remainder term is $o_p(1)$, and

$$n^{1/2} \int \sigma(x)[g_n^*(x) - g(x)]/(2g^{1/2}(x))dx \\ - n^{1/2} \int \frac{\sigma(x)}{2g^{1/2}(x)} d(G_n - G)(x) = o_p(1). \qquad (3.12)$$

The results (3.11) and (3.12), taken together, immediately yield the required result given the conditions on σ.

The proofs of (3.11) and (3.12) are technical and involve complicated mathematics. The reader is referred to Beran (1977, p. 451–452) for complete details of the proof of (3.11). Here we provide the following outline of the proof of (3.12). Let $\varsigma(x) = \sigma(x)/(2g^{1/2}(x))$. Notice that

$$\begin{aligned} n^{1/2}\int \frac{\sigma(x)}{(2g^{1/2}(x))}[g_n^*(x) - g(x)]dx &= n^{1/2}\int \varsigma(x)[g_n^*(x) - g(x)]dx \\ &= n^{1/2}\int \varsigma(x)T_n(x)dx + o_p(1), \end{aligned}$$

where $n^{1/2}T_n(x) = (h_n s)^{-1}\int w((h_n s)^{-1}(x-y)dR_n(y)$, and s is the limiting value of s_n as defined in the statement of the theorem. But

$$n^{1/2}\int \varsigma(x)T_n(x)dx = \int dR_n(y)\int \varsigma(y + h_n sz)w(z)dz. \tag{3.13}$$

and

$$\begin{aligned} &E\left[\int dR_n(y)\int \varsigma(y+h_n sz)w(z)dz - \int \varsigma(y)dR_n(y)\right]^2 \\ &\leq \int w(z)dz \int [\varsigma(y+h_n sz) - \varsigma(y)]^2 g(y)dy. \end{aligned} \tag{3.14}$$

The right-hand side of the last equation tends to zero as $n \to \infty$, which establishes (3.12). In Section 3.3 we will prove the asymptotic results for the minimum Hellinger distance estimators of multivariate location and covariance under more explicit assumptions which directly lead to the convergence of the term in (3.14); see condition (T6) of Theorem 3.4. □

Remark 3.2. The above derivation shows that

$$n^{1/2}(\theta_n - \theta) = n^{1/2}\int \frac{\vartheta_g(x)}{2g^{1/2}(x)}d(G_n - G) + o_p(1).$$

When $g = f_\theta$ belongs to the model family, we get

$$n^{1/2}(\theta_n - \theta) \quad = \quad n^{1/2}\left[\frac{1}{4}I(\theta)\right]^{-1}\int \frac{\dot{s}_\theta(x)}{2f_\theta^{1/2}(x)}d(G_n - F_\theta) + o_p(1)$$

so that the relation

$$\theta_n = \theta + n^{-1/2}I^{-1}(\theta)Z_n(\theta) + o_p(n^{-1/2})$$

continues to hold, where $Z_n(\theta)$ is as in Equation (2.74).

Since Beran's (1977) seminal paper, several other authors have made significant contributions to the problem of minimum Hellinger distance estimation, or more generally, to the problem of minimum distance estimation based on disparities for the continuous model (e.g., Stather, 1981; Tamura and Boos, 1986; Simpson, 1989b; Cao, Cuevas and Fraiman, 1995; Basu, Sarkar and Vidyashankar, 1997b; Park and Basu, 2004; Toma, 2008). Each set of authors makes a contribution in their own way making the literature richer, solving a particular component of the problem, in a particular setting, under their own particular conditions. However, a fully general framework for describing the problem of minimum distance estimation for continuous models does not exist yet, as it does for the discrete case. We will describe the existing techniques under their different settings.

Stather's work, presented in his Ph.D. thesis, represents an important extension of Beran's work. Unfortunately his results are unpublished, so one has to depend on a difficult-to-obtain Ph.D. thesis. Tamura and Boos (1986) studied minimum Hellinger distance estimation in the multivariate context where the location and covariance parameters are of interest. Beran's approach allowed the smoothing parameter to be random but restricted the true distribution to have compact support. Stather allowed infinite support for g, but chose nonrandom bandwidth h_n. Tamura and Boos (1986) also work with a nonrandom h_n and allow an infinite support for g, but their approach has similarities with Beran's proof as well. For actual implementation of the minimum distance method in continuous models based on kernel density estimates it certainly makes sense to use a bandwidth which is a multiple of a robust scale estimate and hence is random. However, the random component is somewhat of a distraction in establishing the asymptotic properties of the estimator. Beran also considered the limiting quantities of the random component to arrive at the final result. Most of the other authors also use nonrandom smoothing parameters. On the whole, the addition of the density estimation component makes the technique of minimum distance estimation significantly more complex compared to the case of discrete models.

The work of Cao et al. (1995) presents another example of the application of the density-based minimum distance method which uses a nonparametric density estimator obtained from the data. However, it is, to a certain degree, different from most of the other cases discussed here, in that these authors only consider the L_1, L_2 and L_∞ metrics as their choice of distances. It is not entirely clear why the authors excluded minimum Hellinger distance estimation from their discussion — a topic which would have fitted admirably with the theme of their paper. Beran (1977), Tamura and Boos (1986) and Simpson (1989b) are not mentioned in the paper. Be that as it may, their approach leads to several interesting results. However, neither of the three metrics considered by the authors belong to the class of disparities, and do not generate fully efficient estimators. Later (Chapter 9) we will discuss another useful method of density-based minimum distance estimation, based on Basu et al. (1998), of which at least the minimum L_2 metric method will be a part. The primary motivation of the method described by Basu et al. (1998) is that in certain cases including the L_2 metric case it is possible to perform the density-based minimum distance estimation routine for continuous models by avoiding direct density estimation. We will see that the asymptotic distribution of the estimator derived in Chapter 9 following Basu et al. (1998) – without data smoothing – matches the asymptotic distribution derived by Cao, Cuevas and

Fraiman (1995, Theorem 2, p. 616) – with data smoothing – for the minimum L_2 case. An estimation process which neither leads to full asymptotic efficiency, nor avoids the complications due to the density estimation, is not among the focused techniques of this book and we do not elaborate further on the approach of Cao, Cuevas and Fraiman (1995) in this book.

3.3 Estimation of Multivariate Location and Covariance

Tamura and Boos (1986) considered minimum Hellinger distance estimation for the multivariate case. The most commonly used kernel density estimate for k dimensional data is

$$g_n^*(x) = (nh_n^k)^{-1} \sum_{i=1}^{n} w((x - X_i)/h_n) \tag{3.15}$$

where w is a density on $\mathbb{R}^k$ and $\{h_n\}$ is a bandwidth sequence with suitable properties. The focus of their work was on determining affine equivariant and affine covariant estimates of multivariate location and covariance respectively (see Tamura and Boos, 1986, Section 2, for relevant definitions). They considered elliptical models having densities of the form

$$f_{\mu,\Sigma}(x) = C_k |\Sigma|^{-1/2} \varrho[(x-\mu)^T \Sigma^{-1} (x-\mu)],$$

for suitable functions C_k and $\varrho(\cdot)$. If an initial affine covariant estimate $\hat{\Sigma}_0$ of covariance is available, a density estimate of the radial type can be constructed as

$$g_n^*(x) = (nh_n^k)^{-1} |\hat{\Sigma}_0|^{-1/2} \sum_{i=1}^{n} w\left(h_n^{-1} ||x - X_i||_{\hat{\Sigma}_0}\right) \tag{3.16}$$

where $||x||_\Sigma^2 = x^T \Sigma^{-1} x$. Such a kernel satisfies the conditions under which the corresponding minimum Hellinger distance estimators of the location and covariance are affine equivariant and affine covariant respectively (Tamura and Boos, 1986, Lemma 2.1). In actual computation Tamura and Boos used the kernel function corresponding to a uniform random vector on the k dimensional unit sphere.

The convergence of the kernel density estimate g_n^* to the true density g in the Hellinger metric is a condition required by all minimum Hellinger distance approaches for consistency and other results. This convergence is implied by the L_1 convergence of g_n^* to g. The L_1 convergence of g_n^* to g is, in turn, implied by the condition $h_n + (nh_n^k)^{-1} \to 0$. Beran (1977), Tamura and Boos (1986) and Simpson (1989b) have all used different bandwidth conditions on he kernel, although all give the necessary convergence in the Hellinger metric.

There is an additional problem faced by the minimum distance method in the multivariate case. The order of the variance of the kernel density estimator is $O((nh_n^k)^{-1})$. However, the maximum absolute bias for the multivariate kernel density estimator, $\sup_x |Eg_n^*(x) - g(x)|$ goes to zero at a fixed rate h_n^2, independently of the dimension k, so that the asymptotic bias term in the minimum Hellinger distance estimator in the multivariate case is NOT $o(n^{-1/2})$. The asymptotic distribution presented by Tamura and Boos, therefore, relate to that of

$$n^{1/2}(T(G_n) - T(G) - B_n)$$

where B_n is the bias term, rather than just that of $n^{1/2}(T(G_n) - T(G))$.

Let $s_\theta = f_\theta^{1/2}$, and let $\dot{s}_\theta$ and $\ddot{s}_\theta$ be the corresponding first and second derivative. Let $\vartheta_g(x)$ be as defined in Theorems 3.1 and 3.3. Let

$$\psi_g(x) = \frac{\vartheta_g(x)}{2g^{1/2}(x)}. \tag{3.17}$$

Let $g_n^*(x)$ be an appropriate kernel density estimator, and let $\tilde{g}_n(x) = E[g_n^*(x)]$. Let $|x|$ and x^2 denote the $k \times 1$ vectors of elementwise absolute and squared values of x, respectively, and $||\cdot||$ represent a norm in $\mathbb{R}^k$. Let $\{\alpha_n\}$ be a sequence of positive numbers tending to infinity with $\lambda_n(x) = \chi_{\{||x|| \le \alpha_n\}}\psi_g(x)$ and $\eta_n(x) = \chi_{\{||x|| > \alpha_n\}}\psi_g(x)$.

Under the above notation, we list below the conditions required by Tamura and Boos for the asymptotic normality proof of the minimum Hellinger distance estimates of multivariate location and scatter.

(T1) Let g_n^* be as defined in Equation (3.15), where w is a symmetric square integrable density on $\mathbb{R}^k$ with compact support S. The bandwidth h_n satisfies $h_n + (nh_n^k)^{-1} \to 0$.

(T2) Either the true distribution belongs to the parametric model and the conditions of Lemma 2.8 are satisfied; or Θ is a compact subset of $\mathbb{R}^p$, the parametric model is identifiable in the sense of Definition 2.2, $f_\theta(x)$ is continuous in θ for almost all x, and $T(G)$ is unique.

(T3) $n \sup_{t \in S} P(||X_1 - h_n t|| > \alpha_n) \to 0$ as $n \to \infty$.

(T4) $(n^{1/2}h_n^k)^{-1} \int |\lambda_n(x)| dx \to 0$ as $n \to \infty$.

(T5) $M_n = \sup_{||x|| < \alpha_n} \sup_{t \in S}\{g(x + h_n t)/g(x)\} = O(1)$ as $n \to \infty$.

(T6) The matrix $\int \psi_g(x)\psi_g^T(x)g(x)dx$ is finite (element-wise), $\sup_{||a||<b} \int \psi_g^2(x + a)g(x)dx$ is finite for some $b > 0$, and $\int [\psi_g(x + a) - \psi_g(x)]^2 g(x)dx \to 0$ as $||a|| \to 0$.

(T7) For each θ in the interior of Θ, $s_\theta = f_\theta^{1/2}$ satisfies the derivative conditions in (3.4) and (3.5), $T(G)$ lies in the interior of Θ, and the matrix $\int \ddot{s}_\theta(x)g^{1/2}(x)dx$ is nonsingular.

Theorem 3.4. [Tamura and Boos (1986, Theorem 4.1)]. *Suppose that conditions (T1) –(T7) hold. Let $X_1, \ldots, X_n$ represent a sample of independent and identically distributed k-vectors having probability density function g. Let $g_n^*(x)$ be a k-dimensional kernel density estimate of the form (3.15). Let G be the true distribution not necessarily in the parametric model. Then the minimum Hellinger distance estimator $T(G_n)$ is consistent, and allows the asymptotic distribution*

$$n^{1/2}[T(G_n) - T(G) - B_n] \to Z^* \sim N\left(0, \int \psi_g(x)\psi_g(x)^T g(x)dx\right),$$

where $B_n = 2C_n^* \int \psi_g(x)\tilde{g}_n^{1/2}(x)g^{1/2}(x)dx$, with $C_n^* \to I$, where I is the p dimensional identity matrix.

Sketch of Proof. Consistency follows from conditions (T1) and (T2) using Lemmas 2.6 and 2.8.

Condition (T7) gives the Taylor series expansion

$$\begin{aligned} T(G_n) - T(G) = &\int \vartheta_g(x)[g_n^{*1/2}(x) - g^{1/2}(x)]dx \\ &+ \xi_n \int \dot{s}_{T(G)}[g_n^{*1/2}(x) - g^{1/2}(x)]dx \end{aligned} \tag{3.18}$$

where $\xi_n \to 0$ in probability as $n \to \infty$. This expansion has already been encountered in Equation (3.6). Note that $\int \vartheta_g(x)g^{1/2}(x)dx = 0$, so that $\int \psi_g(x)g(x)dx$ is also zero where $\vartheta_g(x) = \psi_g(x)2g^{1/2}(x)$. Since $\dot{s}_{T(G)}(x)$ is proportional to $\vartheta_g(x)$, the terms on the right-hand side of Equation (3.18) may be combined, which gives

$$T(G_n) - T(G) = (I + \xi_n^*) \int \vartheta_g(x)[g_n^{*1/2}(x) - g^{1/2}(x)]dx, \tag{3.19}$$

where $\xi_n^* \to 0$ in probability, and I is the p-dimensional identity matrix. This then leads to the result

$$T(G_n) - T(G) - B_n = (I + \xi_n^*) \int \vartheta_g(x)[g_n^{*1/2}(x) - \tilde{g}_n^{1/2}(x)]dx,$$

where

$$B_n = (I + \xi_n^*) \int \vartheta_g(x)[\tilde{g}_n^{1/2}(x) - g^{1/2}(x)]dx.$$

Using the relations $\int \vartheta_g(x)g^{1/2}(x)dx = 0$ and $\vartheta_g(x) = \psi_g(x)2g^{1/2}(x)$, and the fact that $\xi_n^* \to 0$ in probability, the bias can be written as $2C_n^* \displaystyle\int \psi_g(x)\tilde{g}_n^{1/2}(x)g^{1/2}(x)dx$, where $C_n^* \to I$, the p dimensional identity matrix, in probability.

The main task then is to show that $n^{1/2} \int \vartheta_g(x)[g_n^{*1/2} - \tilde{g}_n^{1/2}(x)]dx$ has the appropriate limiting normal distribution. For this purpose, the algebraic identity

$$b^{1/2} - a^{1/2} = (b-a)/(2a^{1/2}) - (b^{1/2} - a^{1/2})^2/(2a^{1/2})$$

is made use of. Note that this identity is the same as in (3.10) employed in Theorem 3.3, but written in a slightly different form. This leads to the relation

$$\begin{aligned}\int \vartheta_g(x)[g_n^{*1/2} - \tilde{g}_n^{1/2}(x)]dx &= \int \psi_g(x)[g_n^*(x) - \tilde{g}_n(x)]dx \\ &\quad - \int \psi_g(x)[g_n^{*1/2} - \tilde{g}_n^{1/2}(x)]^2 dx.\end{aligned}$$

The major steps in the proof involve showing:

(a) $\left| n^{1/2} \int \psi_g(x)[g_n^*(x) - \tilde{g}(x)]dx - n^{1/2}\frac{1}{n}\sum_{i=1}^n \psi_g(X_i) \right| \to 0$ as $n \to \infty$.

(b) $n^{1/2} \int \psi_g(x)[g_n^{*1/2}(x) - g^{1/2}(x)]^2 dx \to 0$ as $n \to \infty$.

To establish (a), let $\tau_{1n}(x) = \int \psi_g(x)[g_n^*(x) - \tilde{g}(x)]dx$. Then

$$n^{1/2}\tau_{1n} = \int\int \psi_g(x) h_n^{-k} w(h_n^{-1}(x-y))dR_n(y)dx, \tag{3.20}$$

where $R_n(y) = n^{1/2}(G_n(y) - G(y))$. Using the Cauchy-Schwarz inequality and condition (T6), the above can be approximated in mean square by $\int \psi_g(y)dR_n(y)$. Notice that the term on the right-hand side of Equation (3.20) is basically the analog of the term $n^{1/2}\int \varsigma(x)T_n(x)dx$ in Theorem 3.3, and the manipulation in item (a) uses relations (3.13) and (3.14) as well.

Let $\tau_{2n} = \int \psi_g(x)[g_n^{*1/2}(x) - \tilde{g}_n^{1/2}(x)]^2 dx$. To establish part (b), notice that

$$\begin{aligned}n^{1/2}\tau_{2n} = n^{1/2} &\int \lambda_n(x)[g_n^{*1/2}(x) - \tilde{g}_n^{1/2}(x)]^2 dx \\ &+ n^{1/2}\int \eta_n(x)[g_n^{*1/2}(x) - \tilde{g}_n^{1/2}(x)]^2 dx.\end{aligned} \tag{3.21}$$

For the term $\int \eta_n(x)[g_n^{*1/2}(x) - \tilde{g}_n^{1/2}(x)]^2 dx$, one completes the square under the integral and looks at all the individual terms. Conditions (T3) and (T6), together with coordinate-wise applications of the Cauchy-Schwarz inequality to each component of the p-vectors show that each of the individual terms is of the order $o_p(n^{-1/2})$, so that the second term on the right-hand side of (3.21) is $o_p(1)$. To prove the convergence of the first term of the right-hand side of (3.21) to zero, it is enough to show that

$$n^{1/2}\int |\lambda_n(x)|[g_n^*(x) - \tilde{g}_n(x)]^2 g^{-1}(x)dx \to 0$$

in probability. However the expected value of the quantity on the left hand side of the above equation is less than

$$\begin{aligned} & (n^{1/2}h_n^k)^{-1}\int |\lambda_n(x)| \int_S g(x+h_nz)g^{-1}(x)w^2(z)dzdx \\ \leq \quad & M_n(n^{1/2}h_n^k)^{-1}\int |\lambda_n(x)|dx \int_S w^2(z)dz, \end{aligned}$$

and the result then follows from conditions (T4) and (T5). □

Tamura and Boos (1986) have briefly discussed the conditions under which Theorem 3.4 is developed. Conditions (T1), (T2) are easily seen to be the standard conditions necessary for the consistency of the estimator. Conditions (T3), (T4) and (T5) relate to the rate at which $\alpha_n \to 0$. For a sequence $\alpha_n = n^l$, $l > 0$, condition (T3) is approximately equivalent to the finiteness of $E||X_1||^{1/l}$. However, (T5) is easier to verify in case of the multivariate normal distribution. (T6) represents some routine model conditions, while (T7) lists the usual smoothness conditions on s_θ.

Condition (T4) also imposes a key restriction on the bandwidth. The conditions of Theorem 3.4 are satisfied for $c_n \sim n^{-(1/2k)+\epsilon}$. Notice that for $k = 1$ this will imply $n^{1/2}c_n \to \infty$. Tamura and Boos argue that the optimal rate of $c_n \sim n^{-1/(k+4)}$ can only be used for $k \leq 3$.

The presence of the bias term B_n is a matter of minor irritation, and for the theorem to be useful, it must be demonstrated that for reasonable sample sizes the bias remains small compared to the variance. Tamura and Boos (1986) have observed this to be true, in simulations from the bivariate normal, for sample sizes smaller than 400.

3.4 A General Structure

Unlike the discrete case, it seems to be difficult to achieve a completely general structure where the theory flows freely for the whole class of minimum distance estimators for the continuous case for all disparities under some general conditions on the model and the residual adjustment function. It is not easy to modify the proofs of Beran (1977), Tamura and Boos (1986) and others so that such a goal may be attained. For the present time we provide a proof by Park and Basu (2004) which does a partial generalization and works under a set of strong conditions on the residual adjustment function. Under the given setup, this produces a general approach and works for all disparities satisfying the assumed conditions. The conditions are strong, and are not satisfied by several common disparities, but the result is still useful and the list of disparities that do satisfy the conditions is also quite substantial. In Section 3.5 we will present an approach based on model smoothing which will encompass a larger class of disparities.

Let the random variable have a distribution G on the real line. Suppose that $X_1, \ldots, X_n$ represent an independently and identically distributed sample generated from G modeled by a parametric family $\mathcal{F}$ as defined in Section 1.1. We assume that the true distribution G and the model distributions F_θ have densities with respect to the Lebesgue measure, and let $\mathcal{G}$ be the class of all distributions with respect to the Lebesgue measure. Let

$$g_n^*(x) = \frac{1}{nh_n} \sum_{i=1}^{n} w\left(\frac{x - X_i}{h_n}\right) \tag{3.22}$$

be the kernel density estimate based on the given data. Let $\rho_C(g_n^*, f_\theta)$ be a distance based on a function C satisfying the disparity conditions. We prove the existence of the corresponding minimum distance functional T and related results in Lemma 3.5. Let $C'(\infty)$ be as defined in Lemma 2.1.

Lemma 3.5. [Park and Basu (2004, Theorem 3.1)]. *We assume that (a) the parameter space Θ is compact, (b) the family of distributions is identifiable, (c) $f_\theta(x)$ is continuous in θ for almost every x, and (d) $C(-1)$ and $C'(\infty)$ (as defined in Lemma 2.1) are finite. Then*

(i) for any $G \in \mathcal{G}$, there exists $\theta \in \Theta$ such that $T(G) = \theta$, and

(ii) for any $F_{\theta_0} \in \mathcal{F}$, $T(F_{\theta_0}) = \theta_0$, uniquely.

Proof. (i) Existence: Denote $D(g, f_\theta) = C(g/f_\theta - 1)f_\theta$. Let $\{\theta_n : \theta_n \in \Theta\}$ be a sequence such that $\theta_n \to \theta$ as $n \to \infty$. Since $D(g, f_{\theta_n}) \to D(g, f_\theta)$ by Assumption (c) and since $\int D(g, f_{\theta_n})$ is finite by Assumption (d) and Lemma 2.2, we have

$$\rho_C(g, f_{\theta_n}) = \int D(g, f_{\theta_n}) \to \int D(g, f_\theta) = \rho_C(g, f_\theta),$$

by a generalized version of the dominated convergence theorem (Royden, 1988, p. 92). Hence $\rho_C(g, f_t)$ is continuous in t and achieves an infimum for $t \in \Theta$ since Θ is compact.

(ii) Uniqueness of the functional at the model: This is immediate from the identifiability assumption on the parametrization and the properties of a statistical distance. □

As in Remark 2.2, the compactness assumption above can be relaxed to include such parameter spaces that may be embedded in a compact space $\bar{\Theta}$, provided the distance $\rho_C(g, f_t)$, viewed as a function of t can be extended to a continuous function on $\bar{\Theta}$.

Theorem 3.6. [Park and Basu (2004, Theorem 3.2)]. *Let $G \in \mathcal{G}$ be the true distribution with density g. Suppose that $|C'(\cdot)|$ is bounded on $[-1, \infty]$. Let $\{g_n\}$ be a sequence of densities, and let $\{G_n\}$ be the corresponding distribution functions. If $T(G)$ is unique, then under the assumptions of Lemma 3.5, the functional $T(\cdot$ is continuous at G in the sense that $T(G_n)$ converges to $T(G)$ whenever $g_n \to g$ in L_1.*

Proof. Suppose that $g_n \to g$ in L_1. Define $\varrho(t) = \rho_C(g, f_t)$ and $\varrho_n(t) = \rho_C(g_n, f_t)$. It follows that

$$|\varrho_n(t) - \varrho(t)| \leq \int |C(\delta_n) - C(\delta)| \, f_t,$$

where $\delta_n = g_n/f_t - 1$ and $\delta = g/f_t - 1$. By the mean value theorem and the boundedness of C' we get,

$$|\varrho_n(t) - \varrho(t)| \leq M \int |\delta_n - \delta| f_t = M \int |g_n - g| \quad \text{for all } t \in \Theta,$$

where $M = \max_\delta |C'(\delta)|$. Since $g_n \to g$ in L_1, we get

$$\sup_t \left| \varrho_n(t) - \varrho(t) \right| \to 0 \ \text{ as } \ n \to \infty. \tag{3.23}$$

Denote $\theta_n = \arg\inf_t \varrho_n(t)$ and $\theta = \arg\inf_t \varrho(t)$. If $\varrho(\theta) \geq \varrho_n(\theta_n)$, then $\varrho(\theta) - \varrho_n(\theta_n) \leq \varrho(\theta_n) - \varrho_n(\theta_n)$, and if $\varrho_n(\theta_n) \geq \varrho(\theta)$, then $\varrho_n(\theta_n) - \varrho(\theta) \leq \varrho_n(\theta) - \varrho(\theta)$. Therefore we have

$$|\varrho_n(\theta_n) - \varrho(\theta)| \leq |\varrho_n(\theta_n) - \varrho(\theta_n)| + |\varrho_n(\theta) - \varrho(\theta)| \leq 2 \sup_t |\varrho_n(t) - \varrho(t)|,$$

which implies $\varrho_n(\theta_n) \to \varrho(\theta)$ as $g_n \to g$ in L_1. Using this and (3.23), we obtain

$$\lim_{n \to \infty} \varrho(\theta_n) = \varrho(\theta). \tag{3.24}$$

Then the proof of the result $\theta_n \to \theta$ as $n \to \infty$ proceeds as in Lemma 2.6 (ii). □

The following is a simple corollary of Theorem 3.6.

Corollary 3.7. [Park and Basu (2004, Corollary 3.3)]. *If the conditions of Theorem 3.6 hold and $g = f_\theta$, then $\theta_n = T(G_n) \to \theta$ and $\rho_C(f_{\theta_n}, f_\theta) \to 0$.*

The second part of the statement follows from the convergence of $\int |f_{\theta_n} - f_\theta|$ to zero from Glick's Theorem (Devroye and Györfi, 1985, p. 10). Also, if Θ cannot be embedded in a compact set, one can extend the results of Theorem 3.6 to the reduced set $\mathcal{G}$, the class of distributions having densities with respect to the dominating measure and satisfying

$$\inf_{t \in \Theta - H} \rho_C(g, f_t) > \rho_C(g, f_{\theta^*})$$

where θ^* belongs to H for some compact subset H of Θ, in the spirit of Lemma 2.7.

For deriving the asymptotic normality of the minimum distance estimator, we will assume that the model is identifiable, $f_\theta(x)$ is twice continuously differentiable with respect to θ, and for any $G \in \mathcal{G}$, $\rho_C(g, f_\theta)$ can be twice differentiated with respect to θ under the integral sign. Sufficient conditions for the above are given in, for example, Park and Basu (2004) as well as many standard texts. The main theorem, presented below, considers the case where the true distribution $G = F_{\theta_0}$ belongs to the model. As in Section 1.1, $\chi(A)$ denotes the indicator for the set A. The proof follows the approach of Tamura and Boos (1986) presented in Theorem 3.4.

Theorem 3.8. [Park and Basu (2004, Theorem 3.4)]. *Let θ_0 be the true value of the parameter, $\{\phi_n\}$ denote any sequence of estimators such that $\phi_n = \theta_0 + o_p(1)$, and $\{\alpha_n\}$ be any sequence of positive real numbers going to ∞. We make the following assumptions:*

(a) $\int |\nabla_{ij} f_{\phi_n} - \nabla_{ij} f_{\theta_0}| = o_p(1)$ *and* $\int |u_{i\phi_n} u_{j\phi_n} f_{\phi_n} - u_{i\theta_0} u_{j\theta_0} f_{\theta_0}| = o_p(1)$, $i, j = 1, \ldots, p$, *for all* $\{\phi_n\}$ *as defined above.*

(b) *The matrix* $I(\theta_0)$ *is finite (element wise), and* $\int u_{i\theta_0}(x+a) u_{j\theta_0}(x+a) f_{\theta_0}(x)dx - \int u_{i\theta_0}(x) u_{j\theta_0}(x) f_{\theta_0}(x)dx \to 0$ *as* $|a| \to 0$, $i, j = 1, \ldots, p$.

(c) *Let the kernel density estimate be as given in Equation (3.22), where* $w(\cdot)$ *is a density which is symmetric about 0, square integrable, and twice continuously differentiable with compact support* S. *The bandwidth* h_n *satisfies* $h_n \to 0$, $n^{1/2} h_n \to \infty$, *and* $n^{1/2} h_n^2 \to 0$.

(d) $\limsup\limits_{n\to\infty} \sup\limits_{y \in \mathcal{A}_n} \int |\nabla_{ij} f_{\theta_0}(x+y) u_{\theta_0}(x)| dx < \infty$ *for* $i, j = 1, \ldots, p$, *where* $\mathcal{A}_n = \{y : y = h_n z,\ z \in S\}$.

(e) $n \sup_{t\in S} P(|X_1 - h_n t| > \alpha_n) \to 0$ *as* $n \to \infty$, *for all* $\{\alpha_n\}$ *as defined above.*

(f) $(n^{1/2} h_n)^{-1} \int \big| u_{\theta_0}(x) \chi(|x| \le \alpha_n) \big| \to 0$, *for all* $\{\alpha_n\}$ *as defined above.*

(g) $\sup_{|x|\le\alpha_n} \sup_{t\in S} \{f_{\theta_0}(x + h_n t)/f_{\theta_0}(x)\} = O(1)$, *as* $n \to \infty$ *for all* $\{\alpha_n\}$ *as defined above.*

(h) $A(\delta)$, $A'(\delta)$, $A'(\delta)(\delta+1)$ *and* $A''(\delta)(\delta+1)$ *are bounded on* $[-1, \infty]$.

Then, $n^{1/2}\big(T(G_n) - T(G)\big)$ *converges in distribution to* $N(0, I^{-1}(\theta_0))$, *where* $\theta_0 = T(G)$, *and* T *is the minimum distance functional based on* ρ_C.

Proof. By condition (c), we have $g_n^*(x) \xrightarrow{\text{a.s}} f_{\theta_0}(x)$ for every x and

$$\int |g_n^*(x) - f_{\theta_0}(x)| dx \to 0,$$

and hence $T(G_n) \xrightarrow{\mathcal{P}} \theta$. Let us denote $\varrho_n(\theta) = \rho_C(g_n^*, f_\theta)$ and $\hat{\theta} = T(G_n)$. Since $\hat{\theta}$ minimizes $\varrho_n(\cdot)$ over Θ, the Taylor series expansion of $\nabla \varrho_n(\hat{\theta})$ at θ_0 yields

$$0 = n^{1/2} \nabla \varrho_n(\hat{\theta}) = n^{1/2} \nabla \varrho_n(\theta_0) + n^{1/2} \nabla_2 \varrho_n(\theta^*)(\hat{\theta} - \theta_0),$$

where θ^* is a point on the line segment joining θ_0 and $\hat{\theta}$. It follows that

$$n^{1/2}(\hat{\theta} - \theta_0) = -[\nabla_2 \varrho_n(\theta^*)]^{-1} n^{1/2} \nabla \varrho_n(\theta_0).$$

Therefore it suffices to show that

$$\nabla_2 \varrho_n(\theta^*) \xrightarrow{\mathcal{P}} I(\theta_0) \qquad (3.25)$$

and

$$-n^{1/2} \nabla \varrho_n(\theta_0) \xrightarrow{\mathcal{D}} Z^* \sim N(0, I(\theta_0)). \qquad (3.26)$$

First we prove (3.25). Let $\delta_n(\theta) = g_n^*/f_\theta - 1$. Differentiating with respect to θ, we have

$$\nabla_2 \varrho_n(\theta) = -\int A(\delta_n) \nabla_2 f_\theta + \int A'(\delta_n)(\delta_n + 1) u_\theta u_\theta^T f_\theta$$

Let $B_1 = \sup_\delta |A(\delta)|$, $B_2 = \sup_\delta |A'(\delta)(\delta + 1)|$; B_1 and B_2 are finite from condition (h); from condition (a) it follows

$$\Big| \int A(\delta_n(\theta^*))(\nabla_2 f_{\theta^*} - \nabla_2 f_{\theta_0}) \Big| \le B_1 \int |\nabla_2 f_{\theta^*} - \nabla_2 f_{\theta_0}| \xrightarrow{\mathcal{P}} 0, \qquad (3.27)$$

and

$$\Big| \int A'(\delta_n(\theta^*))(\delta_n(\theta^*) + 1)(u_{\theta^*} u_{\theta^*}^T f_{\theta^*} - u_{\theta_0} u_{\theta_0}^T f_{\theta_0}) \Big|$$
$$\le B_2 \int |u_{\theta^*} u_{\theta^*}^T f_{\theta^*} - u_{\theta_0} u_{\theta_0}^T f_{\theta_0}| \xrightarrow{\mathcal{P}} 0. \qquad (3.28)$$

Since $A(0) = 0$ and $\delta_n(\theta^*) \to 0$ as $n \to \infty$, using the dominated convergence theorem we have

$$\int A(\delta_n(\theta^*)) \nabla_2 f_{\theta_0} \xrightarrow{\mathcal{P}} 0,$$

and hence by (3.27)

$$\int A(\delta_n(\theta^*)) \nabla_2 f_{\theta^*} \xrightarrow{\mathcal{P}} 0.$$

Similarly, since $A'(0) = 1$ and $\delta_n(\theta^*) \to 0$ as $n \to \infty$, we have

$$\int A'(\delta_n(\theta^*))(\delta_n(\theta^*) + 1) u_{\theta_0} u_{\theta_0}^T f_{\theta_0} \xrightarrow{\mathcal{P}} \int u_{\theta_0} u_{\theta_0}^T f_{\theta_0},$$

by the dominated convergence theorem and hence by (3.28)

$$\int A'(\delta_n(\theta^*))(\delta_n(\theta^*) + 1) u_{\theta^*} u_{\theta^*}^T f_{\theta^*} \xrightarrow{\mathcal{P}} \int u_{\theta_0} u_{\theta_0}^T f_{\theta_0}.$$

Therefore we have (3.25). To prove (3.26) note that

$$-n^{1/2}\nabla\varrho_n(\theta_0) = n^{1/2}\int A(\delta_n(\theta_0))\nabla f_{\theta_0},$$

and it follows from a slight extension of the results of Theorem 3.4 that

$$n^{1/2}\int \delta_n(\theta_0)\nabla f_{\theta_0} = n^{1/2}\int (g_n^* - f_{\theta_0})u_{\theta_0} \xrightarrow{\mathcal{D}} Z^* \sim N(0, I(\theta_0)),$$

with u_{θ_0} playing the role of ψ_g in that theorem. The result also follows directly from Theorem 3.3; however, the latter theorem requires stronger conditions than those assumed in the statement of the present theorem.

It is therefore enough to prove that

$$\left|n^{1/2}\int \big\{A(\delta_n(\theta_0)) - \delta_n(\theta_0)\big\}\nabla f_{\theta_0}\right| \xrightarrow{\mathcal{P}} 0. \tag{3.29}$$

From condition (h) $A'(\delta)$ and $A''(\delta)(\delta+1)$ are bounded. Thus, using Lemma 2.15, we have a finite B such that

$$\big|A(r^2-1) - (r^2-1)\big| \leq B \times (r-1)^2$$

for all $r \geq 0$. Thus,

$$\Big|A(\delta_n(\theta_0)) - \delta_n(\theta_0)\Big| \leq B\left[(g_n^*/f_{\theta_0})^{1/2} - 1\right]^2. \tag{3.30}$$

Using this and (3.29), we have

$$\left|n^{1/2}\int \big\{A(\delta_n(\theta_0)) - \delta_n(\theta_0)\big\}\nabla f_{\theta_0}\right| \leq B\, n^{1/2}\int \{g_n^{*1/2} - f_{\theta_0}^{1/2}\}^2 |u_{\theta_0}|.$$

Now we consider

$$n^{1/2}\int \{g_n^{*1/2} - f_{\theta_0}^{1/2}\}^2 |u_{\theta_0}|.$$

From the inequality $(a-b)^2 \leq (a-c)^2 + (b-c)^2$ for real numbers a, b, c, we get

$$n^{1/2}\int \{g_n^{*1/2} - f_{\theta_0}^{1/2}\}^2 |u_{\theta_0}| \leq T_1 + T_2,$$

where

$$T_1 = n^{1/2}\int \{g_n^{*1/2} - \tilde{g}_n^{1/2}\}^2 |u_{\theta_0}|, \quad T_2 = n^{1/2}\int \{\tilde{g}_n^{1/2} - f_{\theta_0}^{1/2}\}^2 |u_{\theta_0}|.$$

The term T_1 is identical to the term $n^{1/2}\tau_{2n}$ of Theorem 3.4 where the convergence of this term to zero has been established. For T_2, note that the integral $\int\{\tilde{g}_n^{1/2} - f_{\theta_0}^{1/2}\}^2|u_{\theta_0}|$ is essentially of the order of h_n^4, and the given conditions guarantee its convergence to zero. Also see Equation (4.1) of Tamura and Boos (1986), and the succeeding discussion therein. □

3.4.1 Disparities in This Class

The different conditions on the distances required by Park and Basu (2004) to construct the general structure can now be consolidated as follows:

(a) $C(\cdot)$ is strictly convex and thrice differentiable,

(b) $C(\delta) \geq 0$, with equality only at $\delta = 0$ and $C'(0) = 0$, $C''(0) = 1$. Notice that this implies $A(0) = C'(0) = 0$ and $A'(0) = C''(0) = 1$.

(c) $C'(\delta)$, $A(\delta)$, $A'(\delta)$, $A'(\delta)(\delta+1)$ and $A''(\delta)(\delta+1)$ are bounded on $\delta \in [-1, \infty]$.

Although these conditions are quite strong, there is a rich class of distances satisfying the same. We list some of them here. Each family below, except ρ_4, is indexed by a single parameter α. For ρ_4 the tuning parameter is denoted by λ.

$$\begin{aligned}
\rho_1(g,f) &= \frac{1}{2}\int \frac{(g-f)^2}{\alpha\, g + \bar{\alpha}\, f}, \qquad \alpha \in (0,1) \\
\rho_2(g,f) &= \int \frac{(g-f)^2}{\sqrt{\alpha\, g^2 + \bar{\alpha}\, f^2} + \alpha\, g + \bar{\alpha}\, f}, \qquad \alpha \in (0,1) \\
&= \frac{1}{\alpha\bar{\alpha}}\int \Big[\sqrt{\alpha\, g^2 + \bar{\alpha}\, f^2} - \alpha g - \bar{\alpha} f\Big], \\
\rho_3(g,f) &= \frac{1}{2}\int \frac{(g-f)^2}{\sqrt{\alpha\, g^2 + \bar{\alpha}\, f^2}}, \qquad \alpha \in (0,1) \\
\rho_4(g,f) &= \int \Big[\frac{f}{\lambda^2}\{\exp(\lambda - \lambda\, g/f) - 1\} + \frac{g-f}{\lambda}\Big], \qquad \lambda > 0 \\
\rho_5(g,f) &= \int \left[\frac{4}{\alpha\pi} g \tan\left(\frac{\pi\alpha}{2}\frac{g-f}{g+f}\right) - (g-f)\right], \qquad \alpha \in [0,1) \\
\rho_6(g,f) &= \int \left[\frac{2}{\alpha\pi}(g-f)\sin\left(\frac{\pi\alpha}{2}\frac{g-f}{g+f}\right)\right], \qquad \alpha \in [0,1] \\
\rho_7(g,f) &= \int \left[\frac{2}{\alpha\pi}\,(g-f)\tan\left(\frac{\pi\alpha}{2}\frac{g-f}{g+f}\right)\right], \qquad \alpha \in [0,1)
\end{aligned}$$

where $\bar{\alpha} = 1 - \alpha$.

The family ρ_1 is the blended weight chi-square (BWCS) family which we have already encountered. The member of this family for $\alpha = 1/2$ is the symmetric chi-square (SCS). The families ρ_2 and ρ_3 are two other variants of ρ_1

TABLE 3.1
The $C(\cdot)$ and the $A(\cdot)$ functions corresponding to the distances presented in this section. Here $\bar{\alpha} = 1 - \alpha$.

Disparity	$C(\delta)$	$A(\delta)$
ρ_1	$\dfrac{\delta^2}{2(\alpha\delta + 1)}$	$\dfrac{\delta}{\alpha\delta + 1} + \dfrac{\bar{\alpha}}{2}\left[\dfrac{\delta}{\alpha\delta + 1}\right]^2$
ρ_2	$\dfrac{\delta^2}{1 + \alpha\delta + \sqrt{\alpha(\delta+1)^2 + \bar{\alpha}}}$	$\dfrac{1}{\alpha} - \dfrac{1}{\alpha\sqrt{\alpha(\delta+1)^2 + \bar{\alpha}}}$
ρ_3	$\dfrac{\delta^2}{2\sqrt{\alpha(\delta+1)^2 + \bar{\alpha}}}$	$\dfrac{\delta}{\sqrt{\alpha(\delta+1)^2 + \bar{\alpha}}} + \dfrac{\bar{\alpha}}{2}\dfrac{\delta^2}{\left[\sqrt{\alpha(\delta+1)^2 + \bar{\alpha}}\right]^3}$
ρ_4	$\dfrac{e^{-\lambda\delta} - 1 + \lambda\delta}{\lambda^2}$	$\dfrac{(\lambda+1) - [(\lambda+1) + \lambda\delta]e^{-\lambda\delta}}{\lambda^2}$
ρ_5	$\dfrac{4}{\alpha\pi}(\delta+1)\tan\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right) - \delta$	$4\left(\dfrac{\delta+1}{\delta+2}\right)^2 \sec^2\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right) - 1$
ρ_6	$\dfrac{2}{\alpha\pi}\delta\sin\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right)$	$\dfrac{2\delta(\delta+1)}{(\delta+2)^2}\cos\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right) + \dfrac{2}{\alpha\pi}\sin\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right)$
ρ_7	$\dfrac{2}{\alpha\pi}\delta\tan\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right)$	$\dfrac{2\delta(\delta+1)}{(\delta+2)^2}\sec^2\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right) + \dfrac{2}{\alpha\pi}\tan\left(\dfrac{\alpha\pi}{2}\dfrac{\delta}{\delta+2}\right)$

with similar properties. The family ρ_4 is the generalized negative exponential disparity (Bhandari, Basu and Sarkar, 2006) which includes the negative exponential disparity for $\lambda = 1$. The families ρ_5, ρ_6 and ρ_7 are based on trigonometric functions and also satisfy the properties. Note that each of these three families contain the SCS family as limiting cases of $\alpha \to 0$. The families ρ_5 and ρ_6 have been proposed by Park and Basu (2000). In case of both ρ_6 and ρ_7, although these disparities satisfy conditions (a), (b) and (c), more theoretical and empirical investigations are necessary to get a better idea about the performance of the corresponding estimators. Notice that condition (c) above also implies $C'(\infty)$ and $C(-1)$ are finite, the conditions on the disparity that are necessary for establishing the breakdown results.

3.5 The Basu–Lindsay Approach for Continuous Data

In this section we discuss an alternative approach to minimum distance estimation based on disparities for continuous models. Our discussion in this section follows the work of Basu (1991) and Basu and Lindsay (1994). Suppose that $X_1, \ldots, X_n$ represent an independently and identically distributed random sample of size n from a continuous distribution G which has a corresponding density function g with respect to the Lebesgue measure or some other appropriate dominating measure. This is modeled by the family $\mathcal{F}$, and we wish to estimate the parameter θ which represents the best fitting model distribution. Let G_n denote the empirical distribution function. Because of the discrete nature of the data, a disparity between the data and the model cannot be directly constructed. As we have described earlier in this chapter, Beran (1977) took the approach of constructing a nonparametric kernel density estimate g_n^* from the data, and then minimized the distance between this nonparametric density estimate and the model density over the parameter space Θ. Subsequently, several other authors used this approach.

In Beran's approach, the choice of the sequence of kernels (or, more precisely, the sequence of smoothing parameters) becomes critical. The consistency of the kernel density estimator is very important in this case, and complicated conditions have to be imposed on the kernel to make things work properly. The approach taken by Basu and Lindsay (1994) differs from the above in that it proposes that the model be convoluted with the same kernel as well. To distinguish it from the approach of Beran and others, we will refer to the approach based on model smoothing as the Basu–Lindsay approach. Let us denote the kernel integrated version of the model by f_θ^*. In the Basu–Lindsay approach one then constructs a disparity between g_n^* and f_θ^*, and minimizes it over θ to obtain the corresponding minimum distance estimator.

An intuitive rationale for this procedure is as follows. The real purpose here is to minimize some measure of discrepancy between the data and the model. To make the data continuous, an artificial kernel has to be thrown into the system. However, one needs to ensure – through the imposition of suitable conditions on the kernel function and the smoothing parameter – that the additional smoothing effect due to the kernel vanishes asymptotically; in large samples the constructed disparity should really measure the pure discrepancy between the data and the model with minimal or no impact of the kernel and the smoothing parameter. In the Basu–Lindsay approach, one convolutes the model with the same kernel used on the data. In a sense, this compensates for the distortion due to the imposition of the kernel on the data by imposing the same distortion on the model. It is therefore expected that the kernel will play a less important role in the estimation procedure than it plays in Beran's approach, particularly in small samples. As we will see later in this section, one gets consistent estimators of the parameter θ even when the smoothing parameter is held fixed as the sample size increases to infinity.

For a suitable kernel function $K(x, y, h)$, let $g_n^*(x)$ be the kernel density estimate obtained from the data, and f_θ^* be the kernel smoothed model density. These densities may be defined as

$$g_n^*(x) = \int K(x, y, h) dG_n(y) \tag{3.31}$$

and

$$f_\theta^*(x) = \int K(x, y, h) dF_\theta(y) \tag{3.32}$$

respectively. A typical disparity ρ_C based on a disparity generating function C can now be constructed as

$$\rho_C(g_n^*, f_\theta^*) = \int C(\delta(x)) f_\theta^*(x) dx, \tag{3.33}$$

where the Pearson residual is now defined to be

$$\delta(x) = \frac{g_n^*(x) - f_\theta^*(x)}{f_\theta^*(x)},$$

i.e., it is now a residual between the smoothed data and the smoothed model. The minimum distance estimator corresponding to the above disparity is obtained by minimizing the disparity in (3.33) over the parameter space Θ. Under differentiability of the model, this is achieved by solving the estimating equation

$$-\nabla \rho_C(g_n^*, f_\theta^*) = \int A(\delta(x)) \nabla f_\theta^*(x) dx = 0, \tag{3.34}$$

where ∇ represents the gradient with respect to θ.

Since we are dealing with a kernel smoothed version of the model rather than the model itself, it is necessary to replace the maximum likelihood estimator (MLE) with some different reference point. As a natural analog to the MLE, one may choose the estimator which minimizes the likelihood disparity

$$\mathrm{LD}(g_n^*, f_\theta^*) = \int \log(g_n^*(x)/f_\theta^*(x)) g_n^*(x) dx$$

between the smoothed versions of the densities. We call this estimator the MLE*. The subsequent results of this section will establish that all the minimum distance estimators based on disparities presented here are asymptotically equivalent to the MLE* under standard regularity conditions (Theorem 3.19).

The quantities

$$\tilde{u}_\theta(x) = \frac{\nabla f_\theta^*(x)}{f_\theta^*(x)} \tag{3.35}$$

and

$$u_\theta^*(y) = \int \tilde{u}_\theta(x) K(x, y, h) dx \tag{3.36}$$

will be used repeatedly in the rest of the section. We will refer to u_θ^* as the smoothed score function (as opposed to u_θ being the ordinary score function).

In the following, E_θ will represent expectation with respect to the density f_θ.

Lemma 3.9. [Basu and Lindsay (1994, Lemma 3.1)]. *Let g_n^* and f_θ^* be respectively the kernel density estimate obtained from the data and the kernel smoothed model density as defined in (3.31) and (3.32). Then the estimating equation of the MLE* can be written as*

$$-\nabla \mathrm{LD}(g_n^*, f_\theta^*) = \frac{1}{n}\sum u_\theta^*(X_i) = 0. \tag{3.37}$$

Further $E_\theta(u_\theta^*(Y)) = \int u_\theta^*(y) f_\theta(y)dy = 0$ for all θ. Thus, Equation (3.37) represents an unbiased estimating equation.

Proof. By taking a derivative of the likelihood disparity $\mathrm{LD}(g_n^*, f_\theta^*)$ we get

$$\begin{aligned}-\nabla \mathrm{LD}(g_n^*, f_\theta^*) &= \int \frac{g_n^*(x)}{f_\theta^*(x)} \nabla f_\theta^*(x)dx \\ &= \int \tilde{u}_\theta(x)\left[\int K(x,y,h)dG_n(y)\right]dx.\end{aligned}$$

Using Fubini's theorem this becomes

$$\begin{aligned}-\nabla \mathrm{LD}(g_n^*, f_\theta^*) &= \int \left[\int \tilde{u}_\theta(x)K(x,y,h)dx\right]dG_n(y) \\ &= \int u_\theta^*(y)dG_n(y) \\ &= \frac{1}{n}\sum_{i=1}^n u_\theta^*(X_i).\end{aligned}$$

Also,

$$\begin{aligned}E_\theta[u_\theta^*(Y)] &= \int \left[\int \tilde{u}_\theta(x)K(x,y,h)dx\right] f_\theta(y)dy \\ &= \int \tilde{u}_\theta(x)\left[\int K(x,y,h)f_\theta(y)dy\right]dx \\ &= \int \left(\frac{\nabla f_\theta^*(x)}{f_\theta^*(x)}\right) f_\theta^*(x)dx \\ &= \int \nabla f_\theta^*(x)dx = 0.\end{aligned}$$

The estimating function of the MLE* is thus an unbiased estimating function. Therefore the standard asymptotic results for unbiased estimating equations may be used to derive the asymptotic properties of our class of minimum distance estimators. Notice, however, that the relation

$$E_\theta[u_\theta^{*2}(X)] = -E_\theta[\nabla u_\theta^*(X)]$$

is not true. □

3.5.1 Transparent Kernels

The minimum distance estimators generated through the Basu–Lindsay approach are not automatically first-order efficient. However, sometimes it is possible to choose the kernel (relative to the model) in such a way that no information is lost. Suppose $G = F_\theta$ for some $\theta \in \Theta$, i.e., the true distribution belongs to the model. We will show that in some cases the kernel function can be so chosen that for any minimum distance estimator $\hat{\theta}$ within the class of disparities, the asymptotic distribution of $n^{1/2}(\hat{\theta} - \theta)$ is multivariate normal with mean vector zero and asymptotic variance equal to the inverse of the Fisher information matrix; in this sense there is no loss in information. Kernels which allow the generation of first-order efficient estimators as above are called transparent kernels relative to the model. We give a formal mathematical definition of transparent kernels below.

Consider the parametric model $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. Let $u_\theta^*(x)$ be the smoothed score function and $u_\theta(x) = \nabla \log f_\theta(x)$ be the ordinary score function. Then the kernel $K(x, y, h)$ will be called a transparent kernel for the above family of models provided the following relation holds:

$$Au_\theta(x) + B = u_\theta^*(x). \tag{3.38}$$

Here A is a nonsingular $p \times p$ matrix which may depend on the parameter θ and B is a p dimensional vector. However, since both u and u^* have expectation zero with respect to the density f_θ, we get $B = 0$ and the above relation may be simplified and written as

$$Au_\theta(x) = u_\theta^*(x). \tag{3.39}$$

The following lemma is immediate from an inspection of Equations (3.37) and (3.39). Note that the matrix A in Equation (3.39) is nonsingular.

Lemma 3.10. *Suppose that K is a transparent kernel for the family of models $\mathcal{F}$. Then the estimating equation for the MLE* is simply the maximum likelihood score equation, and the MLE* is the ordinary maximum likelihood estimator of θ.*

The above lemma shows that under a transparent kernel, the MLE* is the same as the MLE, and hence all minimum distance estimators, which are asymptotically equivalent to the MLE* are first-order efficient (Theorem 3.19 and Corollary 3.20). We give an example of a transparent kernel below.

Example 3.1. (Normal Model): Suppose that $f_\theta(x)$ is the $N(\mu, \sigma^2)$ density, $\theta = (\mu, \sigma^2)$, and consider K to be the normal kernel with smoothing parameter h (i.e., $K(x, y, h)$ is the $N(y, h^2)$ density at x). Then the smoothed density $f_\theta^*(x)$ is the normal $N(\mu, \sigma^2 + h^2)$ density. We will show that K is a transparent kernel for this family.

Here the ordinary score function and the smoothed score function are given by

$$u_\theta(x) = \begin{pmatrix} \dfrac{x-\mu}{\sigma^2} \\ \dfrac{1}{2\sigma^4}[(x-\mu)^2 - \sigma^2] \end{pmatrix}$$

and

$$u_\theta^*(x) = \begin{pmatrix} \dfrac{x-\mu}{\sigma^2+h^2} \\ \dfrac{1}{2(\sigma^2+h^2)^2}[(x-\mu)^2 - \sigma^2] \end{pmatrix}$$

respectively. Therefore $u_\theta^*(x) = Au_\theta(x)$, where A is the matrix

$$A = \begin{pmatrix} \dfrac{\sigma^2}{\sigma^2+h^2} & 0 \\ 0 & \dfrac{\sigma^4}{(\sigma^2+h^2)^2} \end{pmatrix},$$

and therefore, by definition K is a transparent kernel for this model.

The same may be verified for the m-variate multivariate normal model $\text{MVN}(\mu, \Sigma)$ with a multivariate normal kernel having covariance matrix $h^2 I$ (i.e., having a smoothing parameter of h across each component) where I represents the m dimensional identity matrix. We demonstrate, from first principles, that the MLE* of $\theta = (\mu, \Sigma)$ is the same as the MLE of θ in this case. Notice that the smoothed model density is the $\text{MVN}(\mu, \Sigma + h^2 I)$ density. Let G_n^* represent the distribution function corresponding to g_n^*. Then the estimating equations for the MLE* are, for the parameters in μ,

$$\int (\Sigma + h^2 I)^{-1}(y-\mu)dG_n^*(y) = 0,$$

and, for the parameters in Σ,

$$\int \{(y-\mu)(y-\mu)^T - E_{f_\theta^*}[(Y-\mu)(Y-\mu)^T]\}dG_n^*(y) = 0.$$

Since the distribution of G_n^* is the convolution of G_n and the $\text{MVN}(0, h^2 I)$ density, the solutions to the above estimating equations are just the ordinary maximum likelihood estimators

$$\hat{\mu} = \bar{X}, \quad \hat{\Sigma} = \frac{1}{n}\sum (X_i - \bar{X})(X_i - \bar{X})^T.$$

Once again, there is no loss of information. ‖

3.5.2 The Influence Function of the Minimum Distance Estimators for the Basu–Lindsay Approach

We will set up some more notation and definitions before we prove the next set of results. Corresponding to $\tilde{u}_\theta(x)$ we will define the following quantities to distinguish between the partials:

$$\tilde{u}_{j\theta}(x) = \nabla_j \log f_\theta^*(x).$$

$$\tilde{u}_{jk\theta}(x) = \nabla_{jk} \log f_\theta^*(x).$$

For similar distinction between successive derivatives in case of the smoothed score function we will write

$$u_{j\theta}^*(y) = \int K(x,y,h)\tilde{u}_{j\theta}(x)dx = \nabla_j \int \log f_\theta^*(x)K(x,y,h)dx.$$

$$u_{jk\theta}^*(y) = \int K(x,y,h)\tilde{u}_{jk\theta}(x)dx = \nabla_{jk} \int \log f_\theta^*(x)K(x,y,h)dx.$$

Definition 3.1. Let $J^*(\theta)$ be the $p \times p$ matrix whose jk-th element is given by $E_\theta[-u_{jk\theta}^*(X)]$. It is easy to see that the matrix $J^*(\theta)$ is nonnegative definite.

For the true, unknown, data generating density $g(x)$, let

$$g^*(x) = \int K(x,y,h)g(y)dy$$

be its kernel smoothed version. For a given disparity measure ρ_C, let $\theta^g = T(G)$ be the minimum distance functional defined by the minimization of $\rho_C(g^*, f_\theta^*)$. Let $A(\cdot)$ be the residual adjustment function associated with the disparity ρ_C. We will define $J^{*g}(\theta^g)$ to be the square matrix of dimension p whose jk-th element is given by

$$\int A'(\delta(x))\tilde{u}_{j\theta^g}(x)\tilde{u}_{k\theta^g}(x)g^*(x)dx - \int A(\delta(x))\nabla_{jk} f_{\theta^g}^*(x)dx \tag{3.40}$$

and $u_{\theta^g}^{*g}(y)$ to be the p dimensional vector whose j-th component is

$$\int A'(\delta(x))\tilde{u}_{j\theta^g}(x)K(x,y,h)dx - \int A'(\delta(x))\tilde{u}_{j\theta^g}(x)g^*(x)dx. \tag{3.41}$$

In both of the expressions above, the Pearson residual $\delta(x)$ equals

$$\delta(x) = \frac{g^*(x)}{f_{\theta^g}^*(x)} - 1. \tag{3.42}$$

Let $G_\epsilon(x) = (1-\epsilon)G(x) + \epsilon\wedge_y(x)$, where $\wedge_y$ is as defined in Section 1.1. Let g_ϵ be the corresponding density, and g_ϵ^* be the associated smoothed density.

We denote the functional $T(G_\epsilon)$ obtained via the minimization of $\rho_C(g^*_\epsilon, f^*_\theta)$ as θ_ϵ. It satisfies

$$\int A(\delta_\epsilon(x))\nabla f^*_{\theta_\epsilon}(x)dx = 0, \tag{3.43}$$

where $\delta_\epsilon(x) = g^*_\epsilon(x)/f^*_{\theta_\epsilon}(x) - 1$. Then the influence function $T'(y)$ of the functional T at the distribution G is the first derivative of θ_ϵ evaluated at $\epsilon = 0$, which may be computed by taking a derivative of both sides of (3.43), and solving for $T'(y)$.

The form of the influence function of our minimum distance estimators is presented in the following theorem. The theorem is easily proved by mimicking the proof of Theorem 2.4.

Theorem 3.11. [Basu and Lindsay (1994, Lemma 5.1)]. *Let $G(x)$ be the true data generating distribution, and let $g(x)$ be the corresponding density; let $g^*(x)$ be the kernel smoothed version of $g(x)$. Also let T represent the minimum distance functional corresponding to a particular disparity measure ρ_C, and let $\theta^g = T(G)$. Let $A(\delta)$ be the corresponding residual adjustment function. In this setup, the influence function of the minimum distance functional is given by*

$$T'(y) = [J^{*g}(\theta^g)]^{-1}u^{*g}_{\theta^g}(y), \tag{3.44}$$

where $J^{*g}(\theta^g)$ and $u^{*g}_{\theta^g}$ are as defined above in Equations (3.40) and (3.41).

Corollary 3.12. *Suppose that $G = F_\theta$ for some $\theta \in \Theta$, so that the true distribution belongs to the model. Then the influence function of the minimum distance functional $T(\cdot)$ given in (3.44) reduces to*

$$T'(y) = [J^*(\theta)]^{-1}u^*_\theta(y). \tag{3.45}$$

If in addition K is a transparent kernel for the model f_θ, the influence function in (3.45) has the simple form

$$T'(y) = I^{-1}(\theta)u_\theta(y) \tag{3.46}$$

where $I(\theta)$ is the Fisher information about θ in f_θ and u_θ is the ordinary score function.

Proof. If $G = F_\theta$, it follows that $\theta = \theta^g$. Replacing this in (3.42), we get $\delta(x) = 0$; thus $A(\delta(x)) = 0$ and $A'(\delta(x)) = 1$. Substituting these in Equation (3.40), $J^{*g}(\theta^g)$ becomes

$$\int \tilde{u}_{j\theta}(x)\tilde{u}_{k\theta}(x)f^*_\theta(x)dx = -\int(\nabla_{jk}\log f^*_\theta(x))f^*_\theta(x)dx = -E_\theta[u^*_{jk\theta}(X)].$$

Again, substituting the above in Equation (3.41), we get

$$\int \tilde{u}_{j\theta}(x)K(x,y,h)dx - \int \tilde{u}_{j\theta}(x)f^*_\theta(x)dx = u^*_{j\theta}(\theta).$$

Thus, Equation (3.45) follows.

If in addition K is a transparent kernel for the family, $u^*_\theta(x) = Au_\theta(x)$, and taking derivatives of both sides of this expression one gets

$$\nabla u^*_\theta(x) = A\nabla u_\theta(x) + u_\theta(x)\nabla A. \tag{3.47}$$

Since $E_\theta(u_\theta(X)) = E_\theta(u^*_\theta(X)) = 0$, taking expectation of both sides of (3.47) one gets

$$E[\nabla u^*_\theta(X)] = -AI(\theta). \tag{3.48}$$

But $E[-\nabla u^*_\theta(X)] = J^*(\theta)$ by definition. Substituting this in Equation (3.48), we get $J^*(\theta) = AI(\theta)$. Thus

$$[J^*(\theta)]^{-1}u^*_\theta(y) = [AI(\theta)]^{-1}Au_\theta(y) = I^{-1}(\theta)A^{-1}Au_\theta(y) = I^{-1}(\theta)u_\theta(y)$$

and relation (3.46) holds. □

Remark 3.3. It is intuitively clear that the smoothed score $u^*_\theta(x)$ converges to the ordinary score function $u_\theta(x)$ when the bandwidth is allowed to go to zero. As a result, the coefficient matrix A defined in Equation (3.39) converges to the p-dimensional identity matrix in this case. This phenomenon may be directly observed in Example 3.1.

3.5.3 The Asymptotic Distribution of the Minimum Distance Estimators

We are now ready to prove the asymptotic distribution of the minimum distance estimators within the class of disparities. Most of the lemmas and theorems of this section are straightforward extensions of the results of Section 2.5. Thus, in the proofs of this section we will generally appeal to the logic of Section 2.5, and only emphasize the additional arguments necessary for the description of the kernel smoothed versions.

We assume that a random sample $X_1, \ldots, X_n$ is drawn from the true distribution G (having density function g), and let $K(x, y, h)$ be the kernel function used in the construction of the smoothed densities. A typical density in the model family is denoted by f_θ, while its kernel smoothed version is represented by f^*_θ; similarly g^* represents the kernel smoothed version of the true density. We are interested in the properties of the minimum distance estimator corresponding to the disparity ρ_C. Let θ^g represent the best fitting parameter which satisfies

$$\rho_C(g^*, f^*_{\theta^g}) = \min_{\theta\in\Theta} \rho_C(g^*, f^*_\theta).$$

The Pearson residuals in this context are defined as $\delta_n(x) = g^*_n(x)/f^*_\theta(x) - 1$ and $\delta_g(x) = g^*(x)/f^*_\theta(x) - 1$.

The first lemma follows in a straightforward manner from the definition in Equation (3.31).

Lemma 3.13. *Provided it exists,* $\mathrm{Var}_g(g_n^*(x)) = \frac{1}{n}\lambda(x)$, *where* $\lambda(x)$ *is given by*

$$\lambda(x) = \int K^2(x,y,h)g(y)dy - [g^*(x)]^2.$$

We will assume that the kernel function K is bounded. From now on let

$$K(x,y,h) \leq M(h) < \infty,$$

where $M(h)$ depends on h, but not on x or y. Then

$$\begin{aligned}\lambda(x) &\leq \int K^2(x,y,h)g(y)dy \\ &\leq M(h)\int K(x,y,h)g(y)dy \\ &= M(h)g^*(x). \qquad (3.49)\end{aligned}$$

Lemma 3.14. $n^{1/4}(g_n^{*1/2}(x) - g^{1/2}(x)) \to 0$ *with probability* 1 *if* $\lambda(x) < \infty$.

Proof. One needs an assumption of the finiteness of the variance of $g_n^*(x)$ for the application of the strong law of large numbers. Once that is settled by the assumption of the finiteness of $\lambda(x)$, the rest of the proof is similar to that of Lemma 2.9. □

We will assume that the residual adjustment function $A(\delta)$ is regular in the sense of Definition 2.3 of Chapter 2. As in the case of the discrete models, we will define the Hellinger residuals for the structure under continuous models as

$$\Delta_n(x) = \frac{g_n^{*1/2}(x)}{f_\theta^{*1/2}(x)} - 1$$

$$\Delta_g(x) = \frac{g^{*1/2}(x)}{f_\theta^{*1/2}(x)} - 1.$$

Let $Y_n(x) = n^{1/2}(\Delta_n(x) - \Delta_g(x))^2$.

The following lemma is the analog of Lemma 2.13 and provides the bounds which are useful later in proving the main theorem of the section. The proof essentially retraces the steps of Lemma 2.13. Let $\delta_n(x) = g_n^*(x)/f_\theta^*(x) - 1$ and $\delta_g(x) = g^*(x)/f_\theta^*(x) - 1$.

Lemma 3.15. [Basu and Lindsay (1994, Lemma 6.3)]. *For any* $k \in [0,2]$ *we have*

(i) $E[Y_n^k(x)] \leq E[|\delta_n - \delta_g|]^k n^{k/2} \leq (\lambda^{1/2}(x)/f_\theta^*(x))^k$.

(ii) $E[|\delta_n - \delta_g|] \leq (\lambda^{1/2}(x)/f_\theta^*(x))$.

Lemma 3.16. [Basu and Lindsay (1994, Lemma 6.4)]. $\lim_n E[Y_n^k(x)] = 0$ *for* $k \in [0,2)$.

Proof. From Lemma 3.15 (i), $\sup_n E[Y_n^2(x)] < \dfrac{\lambda(x)}{(f_\theta^*(x))^2} < \infty$. The rest of the proof is similar to Lemma 2.14. □

Let $a_n(x) = A(\delta_n(x)) - A(\delta_g(x))$ and $b_n(x) = (\delta_n(x) - \delta_g(x))A'(\delta_g(x))$. We will find the limiting distribution of $S_{1n} = n^{1/2} \int a_n(x) \nabla f_\theta^*(x) dx$ by showing it to be equal to the limiting distribution of $S_{2n} = n^{1/2} \int b_n(x) \nabla f_\theta^*(x) dx$ which is easier to find.

Assumption 3.1. The smoothed version of the true density g satisfies

$$\int g^{*1/2}(x) |\tilde{u}_\theta(x)| dx < \infty. \tag{3.50}$$

.

Lemma 3.17. [Basu and Lindsay (1994, Lemma 6.5)]. *If $A(\cdot)$ is a regular RAF in the sense of Definition 2.3, and if Assumption 3.1 is satisfied, then*

$$\lim_n E\left|S_{1n} - S_{2n}\right| = 0.$$

Proof. Let $\tau_n(x) = n^{1/2}|a_n(x) - b_n(x)|$. Then by Lemma 2.14,

$$E|S_{1n} - S_{2n}| \le \int E(\tau_n(x)) |\nabla f_\theta^*(x)| dx \le B \int E(Y_n(x)) |\nabla f_\theta^*(x)| dx. \tag{3.51}$$

But by Lemma 3.15 (i), and Equation (3.49),

$$E(Y_n(x)) \le \frac{\lambda^{1/2}(x)}{f_\theta^*(x)} \le \frac{M^{1/2}(h) g^{*1/2}(x)}{f_\theta^*(x)},$$

so that $BM^{1/2}(h) \int g^{*1/2}(x)|\tilde{u}_\theta(x)| dx$ bounds the integral on the right-hand side of (3.51) which is bounded by Assumption 3.1. The rest of the proof is similar to that of Lemma 2.16. □

By an application of Markov's inequality it follows that $S_{1n} - S_{2n} \to 0$ in probability.

Corollary 3.18. *Suppose that*

$$V = \text{Var}\Big[\int K(x, X, h) A'(\delta_g(x)) \tilde{u}_\theta(x) dx\Big]$$

is finite and Assumption 3.1 holds. Then for a regular RAF,

$$S_{1n} \xrightarrow{\mathcal{D}} Z^* \sim N(0, V).$$

Proof. The quantity in question $S_{1n} = n^{1/2} \int a_n(x)\nabla f^*_\theta(x)dx$. The asymptotic distribution of this is the same as that of $S_{2n} = n^{1/2} \int b_n(x)\nabla f^*_\theta(x)dx$ by the previous lemma, which can be written as

$$\begin{aligned} & n^{1/2} \int (\delta_n(x) - \delta_g(x))A'(\delta_g(x))\nabla f^*_\theta(x)dx \\ = \; & n^{1/2} \int (g^*_n(x) - g^*(x))A'(\delta_g(x))\tilde{u}_\theta(x)dx \\ = \; & n^{1/2} \frac{1}{n} \sum_{i=1}^{n} \int (K(x, X_i, h) - E(K(x, X_i, h))A'(\delta_g(x))\tilde{u}_\theta(x)dx \end{aligned}$$

and the result follows by a simple application of the central limit theorem. □

Definition 3.2. We will say that the kernel integrated family of distributions $f^*_\theta(x)$ is smooth if the conditions of Lehmann (1983, p. 409, p. 429) are satisfied with $f^*_\theta(x)$ is place of $f_\theta(x)$.

Suppose $X_1, \ldots X_n$ are n independent and identically distributed observations from a continuous distribution G modeled by $\mathscr{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$. Let g and f_θ represent the corresponding densities, and let g^* and f^*_θ be the corresponding kernel smoothed versions. Consider a disparity $\rho_C(g^*_n, f^*_\theta)$ where g^*_n is the kernel density estimate based on the data. Let $C(\cdot)$ and $A(\cdot)$ be the associated disparity generating and residual adjustment functions respectively, and let θ^g be the best fitting value of the parameter. Let

$$\delta^g_n(x) = \frac{g^*_n(x)}{f^*_\theta(x)} - 1$$

and

$$\delta^g_g(x) = \frac{g^*(x)}{f^*_\theta(x)} - 1.$$

We make the following assumptions for the proof of our main theorem.

(B1) The family $\mathscr{F}$ is identifiable in the sense of Definition 2.2.

(B2) The probability density functions f_θ of the model distributions have common support so that the set $\mathscr{X} = \{x : f_\theta(x) > 0\}$ is independent of θ. Also the true density g is compatible with the model family $\{f_\theta\}$ in the sense of Definition 2.4.

(B3) The family of kernel integrated densities $\{f^*_\theta\}$ is smooth in the sense of Definition 3.2.

(B4) The matrix $J^{*g}(\theta^g)$ as defined in Equation (3.40) is positive definite.

(B5) The quantities

$$\int g^{*1/2}(x)|\tilde{u}_{j\theta}(x)|dx, \int g^{*1/2}(x)|\tilde{u}_{j\theta}(x)\tilde{u}_{k\theta}(x)|dx, \int g^{*1/2}(x)|\tilde{u}_{jk\theta}(x)|dx$$

are bounded for all j and k and all θ in an open neighborhood ω of θ^g.

(B6) For almost all x there exist functions $M_{jkl}(x)$, $M_{jk,l}(x)$, $M_{j,k,l}(x)$ that dominate, in absolute value, $\tilde{u}_{jkl\theta}(x)$, $\tilde{u}_{jk\theta}(x)\tilde{u}_{l\theta}(x)$, $\tilde{u}_{j\theta}(x)\tilde{u}_{k\theta}(x)\tilde{u}_{l\theta}(x)$ for all j, k, l, and that are uniformly bounded in expectation with respect to g^* and f_θ^* for all $\theta \in \omega$.

(B7) The RAF $A(\delta)$ is regular in the sense of Definition 2.3, and K_1 and K_2 represent the bounds for $A'(\delta)$ and $A''(\delta)(1+\delta)$ respectively.

Theorem 3.19. [Basu and Lindsay (1994, Theorem 6.1.)]. *Suppose that conditions (B1)–(B7) hold. Then there exists a consistent sequence θ_n of roots to the minimum disparity estimating equations in (3.34). Also the asymptotic distribution of $n^{1/2}(\theta_n - \theta^g)$ is multivariate normal with mean vector 0 and covariance matrix*

$$[J^{*g}(\theta^g)]^{-1} V_g [J^{*g}(\theta^g)]^{-1}$$

where V_g is the quantity defined in Corollary 3.18, evaluated at $\theta = \theta^g$.

Proof. The proof is essentially the same as the proof of Theorem 2.19, and here we point out just the minor differences.

In case of the linear terms, the difference

$$\left| \int A'(\delta_n^g(x)) \nabla_j f_\theta^*(x) dx - \int A'(\delta_g^g(x)) \nabla_j f_\theta^*(x) dx \right|$$

is now bounded by

$$K_1 \int |\delta_n^g(x) - \delta_g^g(x)| |\nabla_j f_{\theta^g}^*(x)|.$$

But by Lemma 3.15 (ii) and Assumption 3.1,

$$\begin{aligned} E\Big[K_1 \int \big|\delta_n^g(x) - \delta_g^g(x)\big| \, \big|\nabla_j f_{\theta^g}^*(x)\big| dx\Big] &\leq K_1 \int \lambda^{1/2}(x) |\tilde{u}_{j\theta^g}(x)| dx \\ &\leq K_1 M^{1/2}(h) \int g^{*1/2}(x) |u_{j\theta^g}(x)| dx \\ &< \infty, \end{aligned}$$

so the linear term converges as desired.

In case of the quadratic term, the convergences

$$\left| \int (A'(\delta_n^g(x))(1+\delta_n^g) - A'(\delta_g^g(x))(1+\delta_g^g)) \tilde{u}_{j\theta^g}(x) \tilde{u}_{k\theta^g}(x) f_{\theta^g}^*(x) dx \right| \to 0$$

and

$$\left| \int A(\delta_n^g) \nabla_{jk} f_{\theta^g}^*(x) dx - \int A(\delta_g^g) \nabla_{jk} f_{\theta^g}^*(x) dx \right| \to 0$$

hold in probability, so that $\nabla_k \int A(\delta_n^g(x)) \nabla_j f_\theta^*(x) dx$ converges in probability to

$$-\int A'(\delta_g^g(x))(1+\delta_g^g(x)) \tilde{u}_{j\theta^g}(x) \tilde{u}_{k\theta^g}(x) f_{\theta^g}^*(x) dx + \int A(\delta_g^g(x)) \nabla_{jk} f_{\theta^g}^*(x) dx,$$

which is the negative of the (j,k)-th term of the matrix $J^{*g}(\theta^g)$. The rest of the proof is similar to Theorem 2.19. □

Corollary 3.20. *Assume the conditions of Theorem 3.19. In addition, let us suppose that the true distribution belongs to the model ($G = F_\theta$ for some $\theta \in \Theta$) and Kis a transparent kernel for the model family. Then the minimum distance estimator θ_n, $n^{1/2}(\theta_n - \theta^g)$ has an asymptotic normal distribution with mean 0 and covariance matrix $I^{-1}(\theta)$, where $I(\theta)$ is the Fisher information matrix about θ in f_θ.*

Proof. If $G = F_\theta$, $V_g = \mathrm{Var}_\theta(u^*_\theta(X))$ and $J^{*g}(\theta^g) = J^*(\theta)$. If in addition K is a transparent kernel then $u^*_\theta(x) = Au_\theta(x)$ and $J^*(\theta) = AI(\theta)$. Then

$$\begin{aligned}[J^{*g}(\theta^g)]^{-1}V_g[J^{*g}(\theta^g)]^{-1} &= [J^*(\theta)]^{-1}\mathrm{Var}_\theta(u^*_\theta(X))[J^*(\theta)]^{-1}\\ &= [J^*(\theta)]^{-1}\mathrm{Var}_\theta(u^*_\theta(X))[J^*(\theta)^T]^{-1}\\ &= I^{-1}(\theta)A^{-1}AI(\theta)A^T[A^T]^{-1}I^{-1}(\theta)\\ &= I^{-1}(\theta).\end{aligned}$$

□

Remark 3.4. Theorem 3.19 shows that all minimum distance estimators are asymptotically equivalent to the MLE* for the Basu–Lindsay approach. Corollary 3.20, in addition, shows that when the true distribution belongs to the model and the kernel is a transparent kernel, the relation

$$\theta_n = \theta + n^{-1/2}I^{-1}(\theta)Z_n(\theta) + o_p(n^{-1/2}) \tag{3.52}$$

continues to hold where Z_n is as defined in Equation (2.74). Thus, all minimum distance estimators that are asymptotically equivalent to the MLE* are also equivalent to the ordinarily maximum likelihood estimator under these conditions.

3.6 Examples

Example 3.2. This example involves Short's data for the determination of the parallax of the sun, the angle subtended by the earth's radius, as if viewed and measured from the surface of the sun. From this angle and available knowledge of the physical dimensions of the earth, the mean distance from earth to sun can be easily determined. The raw observations are presented in Data Set 2 of Stigler (1977).

To calculate our minimum distance estimators under the normal model, we have used the kernel density function with the Epanechnikov kernel $w(x) = 0.75(1-x^2)$ for $|x| < 1$. Following Devroye and Györfi (1985, pp. 107–108), the

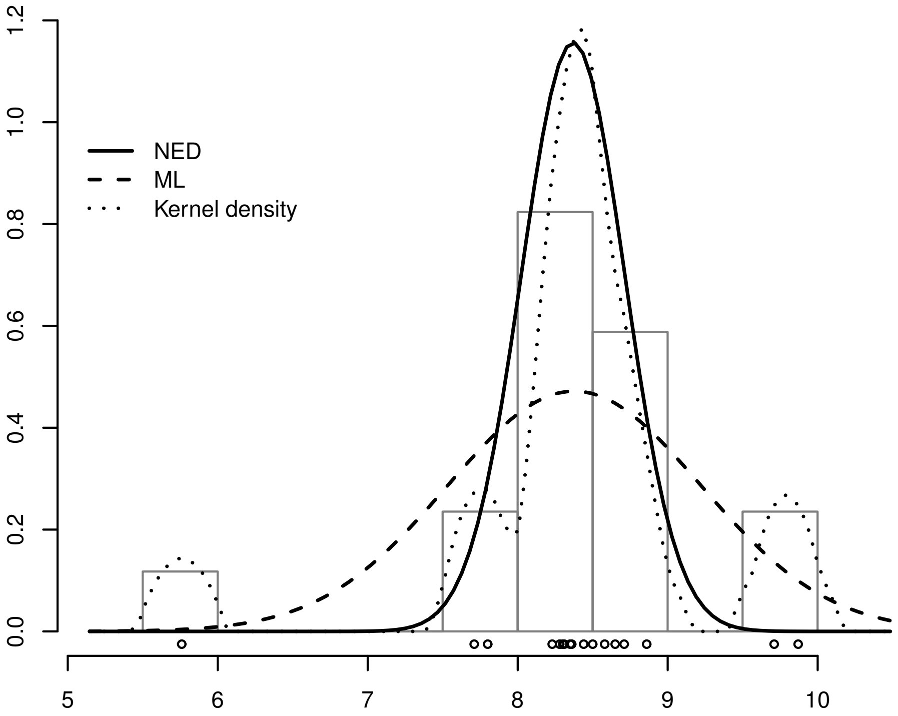

FIGURE 3.1
Normal density fits to Short's data.

mean L_1 criterion with the Epanechnikov kernel and the Gaussian probability density function leads to a bandwidth of the form

$$h_n = (15e)^{1/5}(\pi/32)^{1/10}\sigma n^{-1/5} = 1.66\sigma n^{-1/5}$$

TABLE 3.2
Estimates of the location and scale parameters for Short's data.

	LD	LD+D	HD	$\text{PD}_{-0.9}$	PCS
$\hat{\mu}$	8.378	8.541	8.385	8.388	8.198
$\hat{\sigma}$	0.846	0.552	0.348	0.273	1.046

	NED	$\text{BWHD}_{1/3}$	SCS	$\text{BWCS}_{0.2}$	$\text{GKL}_{1/3}$
$\hat{\mu}$	8.369	8.513	8.378	8.572	8.380
$\hat{\sigma}$	0.345	0.611	0.332	0.596	0.347

where σ is the standard deviation. We have used

$$\hat{\sigma} = \text{median}_i\{X_i - \text{median}_j\{X_j\}\}/0.6745$$

in place of σ in the above expression of the bandwidth. The maximum likelihood estimates of location (μ) and scale (σ) are 8.378 and 0.846, respectively. After removing the large outlier at 5.76, the maximum likelihood estimates of the location and scale become 8.541 and 0.552. In Table 3.2, we provide the minimum distance estimates of the location and scale parameters under several minimum distance methods. The LD column in Table 3.2 represents the maximum likelihood estimates for the full data and the LD+D column represents the outlier deleted maximum likelihood estimates after removing the large outlier at 5.76. Among the other minimum distance estimates, except the minimum PCS estimate, all the rest seem of offer some degree of resistance to the outliers. The minimum HD, NED, SCS and $\text{GKL}_{1/3}$ estimates are all quite close to each other and neatly downweight the large outliers. The $\text{PD}_{-0.9}$ statistic has the strongest downweighting effect. Even the minimum $\text{BWHD}_{1/3}$ and the minimum $\text{BWCS}_{0.2}$ estimates are significantly improved compared to the maximum likelihood estimator.

In Figure 3.1 we present the histogram of the observed data, on which we superimpose the kernel density estimate, as well as the normal fits corresponding to maximum likelihood and the minimum negative exponential disparity. The actual observations are indicated in the base of the figure. Notice that the maximum likelihood estimate of the scale parameter is widely different from the robust estimates. The outlier deleted maximum likelihood estimate of scale, as presented in Table 3.2, is substantially closer to the robust estimates; however, it is still not in the same ballpark. This is because this estimate is still highly influenced by the two moderate outliers close to 10, whereas the robust estimates model the central part of the data and effectively eliminate these observations as well. ‖

Example 3.3. This example involves Newcomb's light speed data (Stigler, 1977, Table 5). The histogram, a kernel density, and the normal fits using the maximum likelihood estimate and the minimum SCS estimate are presented in Figure 3.2. The data set shows a nice unimodal structure, and the normal model would have provided an excellent fit to the data except for the two large outliers. The minimum SCS estimate automatically discounts these large observations, unlike the maximum likelihood estimate. Table 3.3 provides the values of the different minimum distance estimates of the location and scale parameters. Once again the minimum SCS, the minimum HD, the minimum $\text{GKL}_{1/3}$ and the minimum NED estimates are remarkably close to each other. This time the minimum $\text{BWHD}_{1/3}$ and $\text{BWCS}_{0.2}$ estimates are also in this cluster. In this example, we have used the same kernel and the same formula for the bandwidth as in Example 3.2. Clearly the robust minimum distance estimators automatically discount the effects of the large outliers, unlike the maximum likelihood estimator. ‖

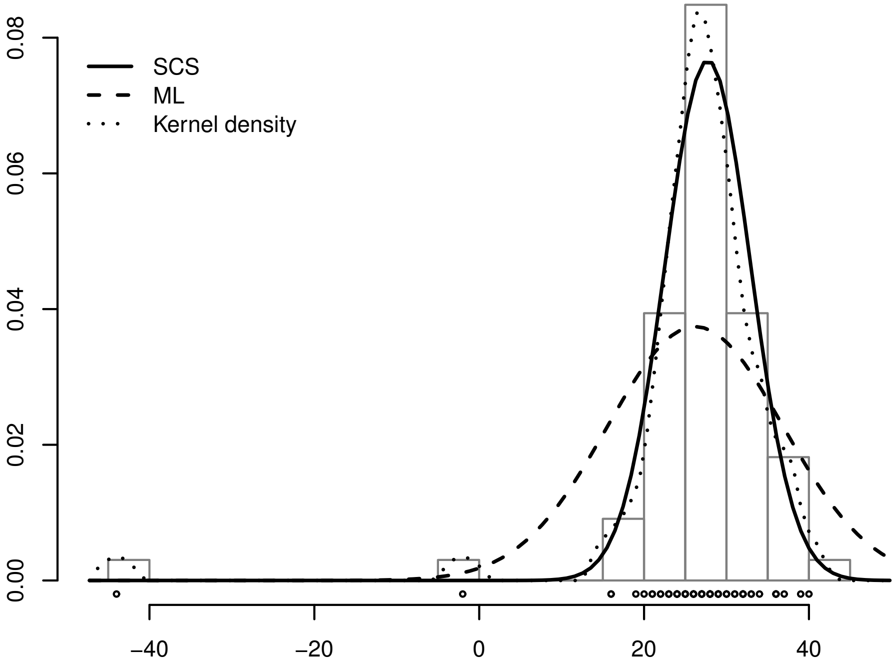

FIGURE 3.2
Normal density fits to the Newcomb data.

TABLE 3.3
Estimates of the location and scale parameters for the Newcomb data.

	LD	LD+D	HD	$PD_{-0.9}$	PCS
$\hat{\mu}$	26.212	27.750	27.738	27.710	14.405
$\hat{\sigma}$	10.664	5.044	5.127	4.746	22.242

	NED	$BWHD_{1/3}$	SCS	$BWCS_{0.2}$	$GKL_{1/3}$
$\hat{\mu}$	27.744	27.744	27.734	27.764	27.741
$\hat{\sigma}$	5.264	5.255	5.204	5.303	5.198

Example 3.4. This example concerns an experiment to test a method of reducing faults on telephone lines (Welch, 1987). Notice again that this data set has a normal structure with one, nonconforming, large outlier. This data set was previously analyzed by Simpson (1989b) using the Hellinger distance. The LD+D column in Table 3.5 represents the maximum likelihood estimate

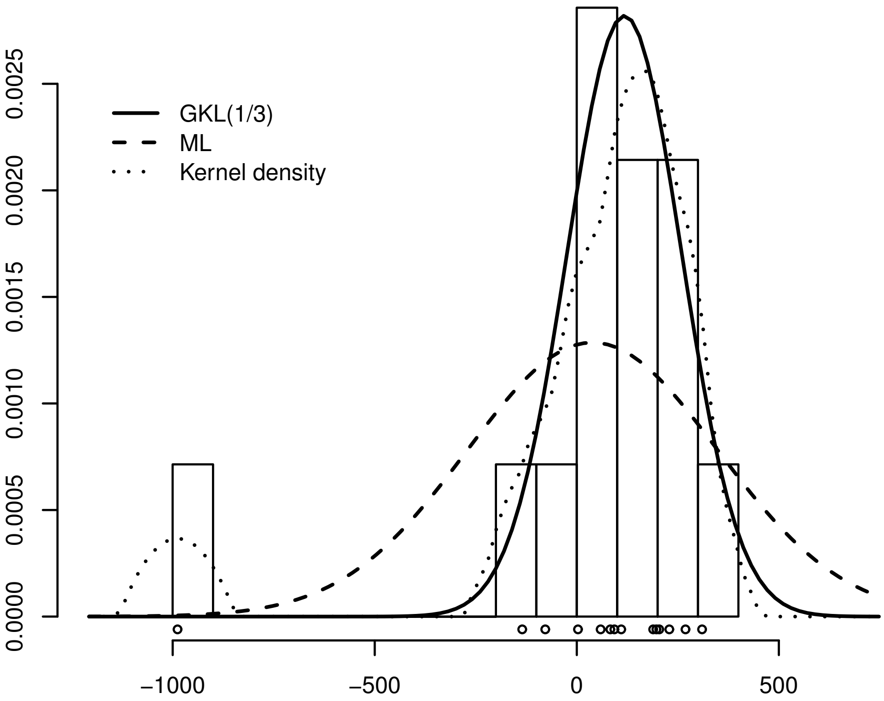

FIGURE 3.3
Normal density fits to the Telephone-line fault data.

after the large outlier (−988) is deleted from the data set. Figure 3.3 presents the kernel density estimate as well as the fits corresponding to the maximum likelihood estimate and the minimum $\mathrm{GKL}_{1/3}$ estimate. All the robust estimates again belong to the same cluster and effectively eliminate the outlier. Here we have used the same kernel and the same formula for the bandwidth as in the two previous examples. ‖

All the three examples bear out the same story that has been observed in the examples in Chapter 2. The observations 1–5 in Section 2.6 are all relevant in the above examples. In spite of the fact that the influence function of all our minimum distance estimators are unbounded, a large outlier has no impact on any of our robust estimates which practically eliminates such observations from the data set; this is true for the minimum $\mathrm{BWHD}_{1/3}$ and the minimum $\mathrm{BWCS}_{0.2}$ estimates as well, although mild outliers produce a somewhat tentative response from these estimates. On the other hand, the maximum likelihood estimator, and, to a greater degree, the minimum PCS estimator, appear to be completely overwhelmed by any small deviation from the assumed model among the observed data.

TABLE 3.4
Estimates of the location and scale parameters for the Telephone-line fault data.

	LD	LD+D	HD	$\mathrm{PD}_{-0.9}$	PCS
$\hat{\mu}$	38.929	117.923	116.769	114.646	-21.701
$\hat{\sigma}$	310.232	127.614	137.612	122.678	418.612

	NED	$\mathrm{BWHD}_{1/3}$	SCS	$\mathrm{BWCS}_{0.2}$	$\mathrm{GKL}_{1/3}$
$\hat{\mu}$	119.093	117.600	118.623	119.179	117.794
$\hat{\sigma}$	145.430	142.560	142.553	146.837	141.499

Example 3.5. This example gives some indication of the possible benefits of the effect of model smoothing compared to the case where there is none. For comparison we consider the pseudo random sample of size 40 from the $N(0,1)$ distribution presented by Beran (1977). We used a normal kernel with several different values of the smoothing parameter h_n and determined the minimum Hellinger distance estimates of the location parameter μ and the scale parameter σ using the Basu–Lindsay approach. The obtained values of the estimates are presented in Table 3.5.

For the same data Beran determined the minimum Hellinger distance estimator of the parameters without model smoothing using the Epanechnikov kernel $w(x) = 0.75(1 - x^2)$ for $|x| \leq 1$ with

$$\hat{\mu}^{(0)} = \tilde{\mu} = \mathrm{median}\{X_i\}, \quad \hat{\sigma}^{(0)} = \tilde{\sigma} = (0.674)^{-1}\mathrm{median}\left\{|X_i - \hat{\mu}^{(0)}|\right\}$$

as the initial estimates. His scale estimate s_n was set at $\hat{\sigma}^{(0)}$, and the form of the kernel density estimate is as given in Equation (3.9).

We present the following observations on the basis of the values in Table 3.5, as well as the associated computational experience.

1. For the Basu–Lindsay approach, the estimates $\hat{\mu}$ and $\hat{\sigma}$ tend, respectively, toward the maximum likelihood estimates of the respective parameters, 0.1584 and 1.012, with increasing values of the smoothing parameter. This phenomenon will also be observed later on in other examples and for other distances. It illustrates a general observation for the Basu–Lindsay approach. As the value of the smoothing parameter increases, the minimum distance estimate ultimately settles on the maximum likelihood estimate. Most of the time, however, the parameters change slowly with increasing bandwidth.

2. An equally important observation is the following. While the estimate of the scale parameter in Table 3.5 is fairly stable over h, its rate of

TABLE 3.5
Estimates of the location ($\hat{\mu}$) and scale ($\hat{\sigma}$) parameters for Beran's data under the Basu–Lindsay approach.

h	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\hat{\mu}$	0.1495	0.1528	0.1551	0.1564	0.1572	0.1577	0.1580
$\hat{\sigma}$	0.9855	0.9903	0.9931	0.9949	0.9960	0.9967	0.9972

TABLE 3.6
Estimates of the location and scale parameters, as reported by Beran (1977).

h_n	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\hat{\mu}$	0.132	0.137	0.141	0.143	0.146	0.148	0.149
$\hat{\sigma}$	0.962	0.977	0.992	1.007	1.023	1.039	1.056

change with the smoothing parameter is significantly faster for the Beran approach; this can be observed from the results reported by Beran (1977, Table 1); for comparison, these numbers are presented here in Table 3.6. As the value of h_n changes from 0.4 to 1.0, the estimate of the scale parameter changes from 0.962 to 1.056. This is what one should expect. Larger values of the smoothing parameter spread out the density estimate. The estimated value of the scale parameter must show a corresponding increase to match the more spread out density. Thus, in small samples the choice of the smoothing parameter becomes critical. But in the Basu–Lindsay approach there is a corresponding spreading out of the model which compensates for the above. As a result, the variation in the estimate of scale in this approach is restricted between 0.9855 and 0.9972 as the smoothing parameter varies between 0.4 and 1.0.

That the scale estimates presented by him increased with increasing bandwidth was, of course, noticed by Beran who suggested additional considerations, such as closeness to the classical estimators, for choosing the value of the appropriate bandwidth.

3. Another fact related to the above issues which is not reflected directly in the table above but has been observed by the authors in repeated simulations is that smaller values of the smoothing parameter lead to more robust estimates (as opposed to more MLE-like estimates for larger values of the same) for the Basu–Lindsay approach. However, the density estimates do become spikier for smaller values h, and the optimization process becomes relatively more unstable. Typically one would require a larger number of iterations for the corresponding root solving process to converge. ‖