2

Statistical Distances

2.1 Introduction

An important component of statistical modeling is the quantification of the amount of discrepancy between data and the model through an appropriate divergence. Based on a sample of n independent and identically distributed observations, such divergences may be constructed, for example, between the empirical distribution function and its population version, or a nonparametric density estimate obtained from the data (constructed, if necessary, using an appropriate density estimation method such as the kernel density estimation) and the probability density function at the model. The first one represents a divergence between two distribution functions, while the second one is the divergence between two probability density functions. In this book, our primary attention will be on the density-based approach. A prominent example of the early use of the density-based idea is the chi-square distance of Pearson (1900).

It is important to make it clear at this stage that many of the density-based divergences (and some of the other divergences as well) that are utilized for different purposes in the statistical literature are not mathematical distances in the sense of being metrics. Most of them are not symmetric in their arguments. This is not very important for statistical purposes. In fact, in many cases it is the asymmetry in the structure of these divergences which has a major role in imparting some of the desirable properties to the estimators generated by them. What is statistically important is that these measures should be nonnegative, and should be equal to zero if and only if the data match the model exactly. Any divergence which satisfies the above two properties will be referred to here as a "statistical distance." This entire book is devoted to the use of statistical distances in statistical inference, with emphasis on the density-based divergences. In a loose sense, we will often drop the term "statistical," and refer to these divergence measures as "distances." In effect, the word "distance" will be used interchangeably with the word "divergence" or, as will be defined later, with the word "disparity."

Unless specifically mentioned otherwise, the scenario under consideration will be the one where n independent and identically distributed observations have been obtained from a distribution G modeled by the parametric family $\mathcal{F}$ as defined in Section 1.1. Our major interest here is in inference problems

involving the parameter θ. We will briefly discuss more specialized problems in later chapters.

Since we are interested in density-based distances, we will assume that the density function f_θ of each model element F_θ exists with respect to an appropriate dominating measure, as does the density function of the true distribution. We will, in fact, often represent the model family in terms of the family of densities $\{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$.

2.2 Distances Based on Distribution Functions

Suppose that $X_1, \ldots, X_n$ represent a random sample of independent and identically distributed observations drawn from an unknown continuous distribution, and let

$$G_n(x) = \frac{1}{n} \sum_{i=1}^{n} \chi(X_i \leq x)$$

be the empirical distribution function where $\chi(A)$ represents the indicator function for the event A. Let $\{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a parametric family of model distributions used to describe the true distribution. A general measure of distance between G_n and F_θ will be denoted by $\rho(G_n, F_\theta)$. Then the weighted Kolmogorov–Smirnov distance is defined by

$$\rho_{\mathrm{KS}}(G_n, F_\theta) = \sup_{-\infty<z<\infty} |G_n(z) - F_\theta(z)| \sqrt{\psi(F_\theta(z))},$$

where $\psi(u) = 1$ gives the usual Kolmogorov–Smirnov distance measure. A minimum distance estimator corresponding to the Kolmogorov–Smirnov distance (or the weighted Kolmogorov–Smirnov distance if appropriate) can be obtained by minimizing the above distance over the parameter space Θ.

The Kolmogorov–Smirnov distance has been used for many different purposes. For example, Kolmogorov–Smirnov statistics have long been used in testing for one dimensional probability distributions by comparing the data with a known and fixed reference distribution. Consider the ordinary Kolmogorov–Smirnov distance

$$\sup_z |G_n(z) - G(z)| \tag{2.1}$$

for testing the null hypothesis that the known distribution $G(\cdot)$ represents the true data generating distribution. When G is continuous, and the null hypothesis is true, the statistic

$$D_n = n^{1/2} \sup_z |G_n(z) - G(z)|$$

has an asymptotic distribution linked to a standard Brownian bridge. The

asymptotic critical values of the statistic D_n can be obtained (Kolmogorov, 1933) by noting the convergence

$$P(D_n \leq t) \to H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t^2}.$$

The test has wide applicability because the asymptotic distribution of the statistic does not depend on the null distribution G; also see Smirnov (1939). Historically, the Kolmogorov–Smirnov metric in (2.1) generated the first goodness-of-fit test which is (pointwise) consistent against any alternative (e.g., Lehmann and Romano, 2008). Massey (1951) appears to be the first to refer to it as the Kolmogorov–Smirnov test. The test is also useful for testing the equality of two distribution functions.

A useful adaptation of the Kolmogorov–Smirnov test is the Lilliefors test (Lilliefors, 1967), which provides a test for normality, and corrects for the bias in the ordinary Kolmogorov–Smirnov test for normality. However, it is generally outperformed by the Shapiro-Wilk statistic (Shapiro and Wilk, 1965) or the Anderson–Darling statistic (Anderson and Darling, 1952) as tests for normality, in terms of the attained power.

The minimum Kolmogorov–Smirnov distance estimator enjoys several nice properties. For example, Drossos and Philippou (1980) show that among many other minimum distance estimators, the minimum Kolmogorov–Smirnov distance estimator enjoys the invariance of the maximum likelihood estimator. Parr and Schucany (1980) have shown that the minimum distance estimator based on the Kolmogorov–Smirnov distance is competitive with several of its rivals. In other applications, Durbin (1975) considered the distribution of the Kolmogorov–Smirnov statistic in terms of a Fourier transform and produced explicit results for the exponential case. In Margolin and Maurer (1976), the results for the exponential case are derived from general results for order statistics, and computationally efficient approximations to these distribution functions are obtained.

The Kolmogorov–Smirnov distance is an extremely popular distance in statistics and many other scientific disciplines. It has been used, for example, in pattern recognition, image comparison and image segmentation, signature verification, credit scoring and library design, just to name a few application domains. Some examples of recent applications of the distance include Benson and Nikitin (1995), Rassokhin and Agrafiotis (2000) and Bockstaele et al. (2006). Weber et al. (2006) present and implement an algorithm for computing the parameter estimates in a univariate probability model for a continuous random variable that minimizes the Kolmogorov–Smirnov test statistic.

However, Donoho and Liu (1988b) have demonstrated certain "pathologies" of some minimum distance estimators, including the minimum Kolmogorov–Smirnov distance estimator; for example, the asymptotic variance of the minimum Kolmogorov–Smirnov distance estimator is unbounded over small Kolmogorov–Smirnov neighborhoods of the model, a consequence

of the Kolmogorov–Smirnov distance being "non-Hilbertian." Also see Millar (1981) for some other properties of this method of estimation.

The ordinary Kolmogorov–Smirnov distance and the associated inference relate to continuous underlying distributions, and the overwhelming majority of the available literature in this area operates under the continuity assumption. Appropriate modifications are needed for the applicability of these methods in discrete and discontinuous distributions. This has been addressed by, among others, Schmid (1958), Noether (1963), Coberly and Lewis (1972), Conover (1972), Petitt and Stevens (1977), Wood and Altavela (1978) and Gleser (1985). Bartels et al. (1978) provide some computational details in this connection.

Another prominent member of the class of distances based on distribution functions is the Cramér–von Mises distance. The weighted Cramér–von Mises distance between the empirical distribution function G_n and the model distribution function F_θ is given by

$$\rho_{\mathrm{CM}}(G_n, F_\theta) = \int_{-\infty}^{\infty} (G_n(z) - F_\theta(z))^2 \psi(F_\theta(z)) dF_\theta(z), \tag{2.2}$$

where $\psi(u) = 1$ gives the usual Cramér–von Mises distance, and $\psi(u) = [u(1-u)]^{-1}$ generates the Anderson–Darling measure (Anderson and Darling, 1952). The ordinary Cramér–von Mises distance is often denoted by the symbol ω^2 in the literature.

The Cramér–von Mises criterion is widely used for testing goodness-of-fit, as well as for parametric estimation. To test the goodness-of-fit of a given, fixed distribution G, Cramér (1928) suggested an integral measure obtained by integrating a weighted version of the squared residual $(G_n(x) - G(x))^2$. It appears that von Mises (1931) suggested an equivalent test independently. The form of the Cramér–von Mises distance that is currently being used by practitioners includes the modifications of Smirnov (1936, 1937), and the test has also been called the Cramér-Smirnov test in the literature. Darling (1955, 1957) gives an excellent account of these tests. Also see Anderson (1962).

For the one sample case, given ordered observations $X_1, \ldots, X_n$ obtained through an independently and identically distributed sample, the Cramér–von Mises statistic may be described as

$$n\rho_{\mathrm{CM}}(G_n, G) = n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - G(X_i) \right]. \tag{2.3}$$

We reject that hypothesis that the data come from the distribution G for large observed values of the statistic.

In the two-sample case, let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be the observed values, arranged in increasing order, in independent and identically distributed samples from the two populations. Suppose we are interested in testing for the equality of the two populations. In this case, the analog of the test statistic

in Equation (2.3) is given by

$$\frac{U}{nm(n+m)} - \frac{4mn-1}{6(m+n)},$$

where $U = n\sum_{i=1}^{n}(r_i - i)^2 + m\sum_{j=1}^{m}(s_j - j)^2$, and r_i and s_j are the ranks of the observations X_i and Y_j respectively in the pooled sample (Anderson, 1962). Available tables of the statistic U facilitate the comparison and conclusion based on this statistic.

Millar (1981) considered weighted Cramér–von Mises distance estimation from a decision theoretic standpoint. Boos (1981) used a weighted Cramér–von Mises distance to estimate the parameter in a location family. Parr and De Wet (1981) extended the results of Boos to more general settings outside the location model. Wiens (1987) discussed the weighted Cramér–von Mises distance estimation of a location parameter and developed robust estimators which have optimal minimax variance properties in gross error neighborhoods of the target model. Öztürk and Hettmansperger (1997) considered generalized weighted Cramér–von Mises distance estimators which have high efficiency at the true model and stable behavior at the neighborhood of the target model. Heathcote and Silvapulle (1981) and Hettmansperger et al. (1994) used the unweighted Cramér–von Mises distance for the simultaneous estimation of location and scale parameters.

In the k-sample one way analysis of variance problem, Brown (1982) developed a multiple response permutation procedure based on within-group sums of absolute rank differences. The relevant distribution turns out to be a convolution of $k-1$ copies of the usual Cramér–von Mises distribution.

The Anderson–Darling distance (Anderson and Darling, 1952), a special case of (2.2) for $\psi(u) = [u(1-u)]^{-1}$ provides one of the most powerful tests for normality against most alternatives (e.g., Stephens, 1974). Also see Boos (1982) for a general discussion. For details of the application of the Anderson–Darling test in normal, exponential, Gumbel and Weibull distributions, see Shorack and Wellner (1986). Stephens (1979) discusses the case of the logistic distribution. Tables of critical values for the statistic for some specific distributions are given in Pearson and Hartley (1972).

See Kiefer (1959) for another application of the Anderson–Darling statistic to the k-sample problem. For some other aspects of statistical analysis based on the Anderson–Darling statistic, the reader is referred to Scholz and Stephens (1987), Thas and Ottoy (2003) and Mansuy (2005).

2.3 Density-Based Distances

Within the class of density-based distances, our focus in this book will be on the family of chi-square type distances, generally called ϕ-divergences, f-

divergences, g-divergences, or disparities (see Csiszár, 1963, 1967a,b; Ali and Silvey, 1966; Lindsay, 1994; Pardo, 2006). The primary reason for focusing our attention on this family of distances is that all the minimum distance procedures based on these distances generate estimators which are asymptotically fully efficient and many of them have remarkably strong robustness properties. In addition, the likelihood based methods (maximum likelihood estimation, likelihood ratio test) can often be considered as special cases of this approach, so that the entire framework can be viewed as a generalized approach containing the likelihood based methods as well as other robust minimum distance procedures as special cases.

For ease of presentation, we will first introduce these distances for discrete models. The distances for the continuous models will be introduced subsequently. Some other distances will be discussed in later chapters.

2.3.1 The Distances in Discrete Models

Parametric estimation based on minimum chi-square type methods has been studied by many authors. Pardo (2006) provides a nice description of the minimum distance methods in connection with these distances; however, the coverage of the latter work is primarily limited to the case of discrete models with finite support based on the multinomial distribution. In addition, the robustness angle is not emphasized in Pardo (2006). In contrast, we will consider the more general countable support case, extend this later to the case of continuous models, and present the issue of robustness as the distinguishing theme of this book. To be specific, we will follow the approach of Lindsay (1994) to describe our minimum distance methodology; this is primarily because it allows a neat illustration of the adjustment between robustness and efficiency through the residual adjustment function and the Pearson residual to be described later in this chapter.

Let $X_1, \ldots, X_n$ represent a sequence of independent and identically distributed observations from a distribution G having a probability density function g with respect to the counting measure. Without loss of generality, we will assume that the support of the distribution G is $\mathcal{X} = \{0, 1, 2, \ldots, \}$. Let $d_n(x)$ represent the relative frequency of the value x in the random sample described in the previous paragraph. Let the parametric model family $\mathcal{F}$, which models the true data generating distrubtion G, be as defined in Section 1.1. We will denote by $\mathcal{G}$ the class of all distributions having densities with respect to the counting measure (or the appropriate dominating measure in other cases), and we will assume this class to be convex. We will also assume that both G and $\mathcal{F}$ belong to $\mathcal{G}$.

One of our main aims in this context is to estimate the parameter θ efficiently and robustly. In minimum distance estimation, as mentioned before, we estimate the parameter θ by determining the element of the model family which provides the closest match to the data in terms of the distance under consideration. In the discrete setup described here, this can be achieved by

quantifying the separation between the vectors $\underset{\sim}{d} = (d_n(0), d_n(1), \ldots)^T$ and $\underset{\sim}{f_\theta} = (f_\theta(0), f_\theta(1), \ldots)^T$, which are probability vectors satisfying

$$\sum_{x=0}^{\infty} d_n(x) = \sum_{x=0}^{\infty} f_\theta(x) = 1. \tag{2.4}$$

One way to quantify the separation between the vectors $\underset{\sim}{d}$ and $\underset{\sim}{f_\theta}$ is through the class of disparities (Lindsay, 1994); see also Csiszár (1963, 1967a,b) and Ali and Silvey (1966). In the following, we formally define a disparity.

Definition 2.1. Let C be a thrice differentiable, strictly convex function on $[-1, \infty)$, satisfying

$$C(0) = 0. \tag{2.5}$$

Let the *Pearson residual* at the value x be defined by

$$\delta(x) = \frac{d_n(x)}{f_\theta(x)} - 1. \tag{2.6}$$

(We will denote it by $\delta_n(x)$ whenever the dependence on n has to be made explicit). Then the disparity between $\underset{\sim}{d}$ and $\underset{\sim}{f_\theta}$ generated by C is given by

$$\rho_C(d_n, f_\theta) = \sum_{x=0}^{\infty} C(\delta(x)) f_\theta(x). \tag{2.7}$$

For simplicity of notation we will refer to the disparity by the expression on the left-hand side of Equation (2.7), rather than as $\rho_C(\underset{\sim}{d}, \underset{\sim}{f_\theta})$.

The conditions imposed on the function $C(\cdot)$ by Definition 2.1 will be called the disparity conditions. In the spirit of the statistical distance notation, we will often refer to the quantity in (2.7) based on a function C as in Definition 2.1 as a distance satisfying the disparity conditions. We will refer to the function C as the disparity generating function.

Notice that the right-hand side of Equation (2.7) is the expectation of $C(\delta(X))$ with respect to the density $f_\theta(x)$. Since C is a strictly convex function, it follows from Jensen's inequality and Equations (2.4) and (2.5) that

$$\sum_{x=0}^{\infty} C(\delta(x)) f_\theta(x) \geq C(E_{f_\theta}(\delta(X))) = C(0) = 0,$$

establishing the result that the disparity defined in Equation (2.7) is always nonnegative. Notice that since $E_{f_\theta}(C(\delta(x))) = E_{f_\theta}(C(\delta(x)) - k\delta(x))$ for any scalar k, there may be different versions of the disparity generating function C generating the same disparity.

By the strict convexity of the disparity generating function C, the disparity in (2.7) is zero only when $d_n \equiv f_\theta$, identically. Thus, the disparity satisfies the basic requirements of a statistical distance. For simplicity of presentation, in the following we will write the expression on the right-hand side of Equation (2.7) as $\sum C(\delta) f_\theta$.

Specific forms of the function C generate many well known disparities. For example, $C(\delta) = (\delta + 1)\log(\delta + 1) - \delta$ generates the well known likelihood disparity (LD) given by

$$\mathrm{LD}(d_n, f_\theta) = \sum [d_n \log(d_n/f_\theta) + (f_\theta - d_n)] = \sum d_n \log(d_n/f_\theta), \tag{2.8}$$

which is a form of the Kullback–Leibler divergence (Kullback and Leibler, 1951). However, it is the symmetric opposite of likelihood disparity that we will refer to as the Kullback–Leibler divergence (KLD), which has the form

$$\mathrm{KLD}(d_n, f_\theta) = \sum [f_\theta \log(f_\theta/d_n) + (d_n - f_\theta)] = \sum f_\theta \log(f_\theta/d_n). \tag{2.9}$$

This distance corresponds to $C(\delta) = \delta - \log(\delta + 1)$. The (twice, squared) Hellinger distance (HD) has the form

$$\mathrm{HD}(d_n, f_\theta) = 2 \sum [{d_n}^{1/2} - {f_\theta}^{1/2}]^2 \tag{2.10}$$

which corresponds to $C(\delta) = 2((\delta + 1)^{1/2} - 1)^2$. The Pearson's chi-square (divided by 2) is defined as

$$\mathrm{PCS}(d_n, f_\theta) = \sum \frac{(d_n - f_\theta)^2}{2 f_\theta}, \tag{2.11}$$

where $C(\delta) = \delta^2/2$, and the Neyman's chi-square (divided by 2) is defined as

$$\mathrm{NCS}(d_n, f_\theta) = \sum \frac{(d_n - f_\theta)^2}{2 d_n}, \tag{2.12}$$

where $C(\delta) = \frac{\delta^2}{2(\delta+1)}$. For the rest of the chapter the terms HD, PCS and NCS will indicate the quantities in Equations (2.10), (2.11) and (2.12) respectively.

There are several important subfamilies of the class of disparities, each generating several common divergences. These include the Cressie–Read family (Cressie and Read, 1984) of power divergences (PD), indexed by a real parameter $\lambda \in (-\infty, \infty)$, and having the form

$$\mathrm{PD}_\lambda(d_n, f_\theta) = \frac{1}{\lambda(\lambda + 1)} \sum d_n \left[\left(\frac{d_n}{f_\theta} \right)^\lambda - 1 \right]. \tag{2.13}$$

Cressie and Read (1984) denoted the power divergence with tuning parameter λ as $I^{\boldsymbol{\lambda}}$. For the sake of a uniform notation, we will refer to it as PD_λ. Notice that for values of $\lambda = 1, 0, -1/2, -1$ and -2 the Cressie–Read form

in Equation (2.13) generates the Pearson's chi-square (PCS), the likelihood disparity (LD), the Hellinger distance (HD), the Kullback–Leibler divergence (KLD) and the Neyman's chi-square (NCS) respectively. The LD and KLD are not directly obtainable from Equation (2.13) by replacing $\lambda = 0$ and -1 in its expression; however, they are the continuous limits of the expression on the right hand side of Equation (2.13) as $\lambda \to 0$ and $\lambda \to -1$ respectively.

It is also easy to see that the Hellinger distance ($\lambda = -1/2$) is the only distance metric within the Cressie–Read family, and the divergences that are equally spaced on either side of $\lambda = -1/2$ in the λ scale are symmetric opposites of each other obtained by interchanging d_n and f_θ. Thus, one gets the KLD ($\lambda = -1 = -1/2 - 1/2$) by interchanging d_n and f_θ in the expression for LD ($\lambda = 0 = -1/2 + 1/2$), and the NCS ($\lambda = -2 = -1/2 - 3/2$) by interchanging d_n and f_θ in the expression for the PCS ($\lambda = 1 = -1/2 + 3/2$). In fact, one can write $\mathrm{PD}_\lambda(d_n, f_\theta)$ alternatively as

$$\mathrm{PD}^*_\alpha(d_n, f_\theta) = \frac{4}{1-\alpha^2} \sum_x d_n \left[1 - \left(\frac{f_\theta}{d_n}\right)^{(1+\alpha)/2}\right]. \tag{2.14}$$

Notice that under this formulation $\mathrm{PD}_\lambda = \mathrm{PD}^*_\alpha$, with $\alpha = -(1 + 2\lambda)$. The PD^*_α distance is symmetric for $\alpha = 0$ (which is the HD), and the choices α and $-\alpha$ generate distances which are symmetric opposites of each other when $\alpha \neq 0$. Jimenez and Shao (2001) have called distances corresponding to α and $-\alpha$ adjoints of each other. The Hellinger distance is self adjoint. There are some advantages of the form in (2.14). However, in the rest of this book we will continue to use the form for the power divergence family given in (2.13) or the modified form in (2.15) to conform to the huge volume of research that has followed up on the inspired work of Cressie and Read (1984).

While we know that any distance generated by a function $C(\cdot)$ satisfying the disparity conditions is nonnegative, it is sometimes of substantial benefit to be able to express the disparity in such a way that each individual term in the summation of the right-hand side of Equation (2.7) is nonnegative (this is not an automatic property for an arbitrary version of the disparity generating function C). Often we can achieve this with a little extra effort. For example, with the Cressie–Read family, one can write the disparity PD_λ as

$$\mathrm{PD}_\lambda(d_n, f_\theta) = \sum \left\{ \frac{1}{\lambda(\lambda+1)}\, d_n \left[\left(\frac{d_n}{f_\theta}\right)^\lambda - 1\right] + \frac{f_\theta - d_n}{\lambda+1} \right\}, \tag{2.15}$$

and it is a simple matter to check that all the terms in the right-hand side of Equation (2.15) are nonnegative. In terms of the function $C(\cdot)$, this amounts to redefining the function, without changing the value of the disparity, so that $C(\delta)$ is a nonnegative convex function with $C(0) = 0$ as its minimum value. Since $C(\cdot)$ is convex, this essentially means introducing the additional restriction

$$C'(0) = 0, \tag{2.16}$$

where C' represents the first derivative of C with respect to its argument (similarly C'' will represent the second derivative of C). This is the reason for considering, for example, the alternative expression $\sum[d_n \log(d_n/f_\theta) + (f_\theta - d_n)]$ when defining the likelihood disparity in Equation (2.8) (rather than just the usual $\sum d_n \log(d_n/f_\theta)$). Defining $C(\cdot)$ to be nonnegative aids the interpretation, and is also helpful in deriving some of the asymptotic properties that we will consider later. However, this is not a basic disparity condition, and depending on our mathematical convenience, we will also make use of disparity generating functions which are not necessarily of this type.

The C function for the Cressie–Read family of distances under the formulation in Equation (2.15) is given by

$$C_\lambda(\delta) = \frac{(\delta+1)^{\lambda+1} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}. \tag{2.17}$$

Other subfamilies within the class of disparities include the blended weight Hellinger distance (Lindsay, 1994; Basu and Lindsay, 1994; Shin, Basu and Sarkar, 1995), as a function of a tuning parameter $\alpha \in [0,1]$, as

$$\text{BWHD}_\alpha(d_n, f_\theta) = \frac{1}{2}\sum \frac{(d_n - f_\theta)^2}{(\alpha {d_n}^{1/2} + \bar{\alpha} f_\theta^{1/2})^2}, \tag{2.18}$$

where $\bar{\alpha} = 1 - \alpha$. The C function for this is given by

$$C_\alpha(\delta) = \frac{1}{2}\frac{\delta^2}{[\alpha(\delta+1)^{1/2} + \bar{\alpha}]^2}. \tag{2.19}$$

For $\alpha = 0, 1/2$, and 1, this family generates the Pearson's chi-square, the Hellinger distance and the Neyman's chi-square respectively. Although the above authors have used the BWHD_α only for $\alpha \in [0,1]$, in practice there is no conceptual difficulty in allowing the range of the tuning parameter to be $(-\infty, \infty)$. The disparities on the right-hand side of (2.18) satisfy the conditions for statistical distances for all $\alpha \in (-\infty, \infty)$, which is a consequence of the squared term in the denominator.

Another such family is the blended weight chi-square divergence (Lindsay, 1994; Shin, Basu and Sarkar, 1996); this is given, as a function of a tuning parameter α, as

$$\text{BWCS}_\alpha(d_n, f_\theta) = \frac{1}{2}\sum \frac{(d_n - f_\theta)^2}{\alpha d_n + \bar{\alpha} f_\theta}, \tag{2.20}$$

where $\bar{\alpha} = 1 - \alpha$. The C function for this is given by

$$C_\alpha(\delta) = \frac{1}{2}\frac{\delta^2}{[\alpha(\delta+1) + \bar{\alpha}]}. \tag{2.21}$$

This family generates the Pearson's chi-square and the Neyman's chi-square

for $\alpha = 0$ and 1 respectively. Note that the particular choice $\alpha = 1/2$ generates the symmetric chi-square (SCS) measure, given by

$$\text{SCS}(d_n, f_\theta) = \sum \frac{(d_n - f_\theta)^2}{d_n + f_\theta}, \tag{2.22}$$

which leads to another genuine distance within the class of disparities.

The family of generalized negative exponential disparities (Bhandari, Basu and Sarkar, 2006), as a function of the tuning parameter $\lambda \geq 0$, corresponds to the C function

$$C_\lambda(\delta) = \begin{cases} (e^{-\lambda\delta} + \lambda\delta - 1)/\lambda^2, & \text{if } \lambda > 0 \\ \delta^2/2 & \text{if } \lambda = 0. \end{cases} \tag{2.23}$$

The $\lambda = 0$ case, which is obtained by taking the limit of the expression for $\lambda > 0$ as $\lambda \to 0$, corresponds to the Pearson's chi-square. The ordinary negative exponential disparity, which corresponds to $\lambda = 1$, has the form $\text{NED}(d_n, f_\theta) = \sum (e^{-\delta} + \delta - 1) f_\theta$.

The disparities for the generalized Kullback–Leibler (GKL) divergence family have been defined by Park and Basu (2003) as

$$\text{GKL}_\tau(d_n, f_\theta) = \sum \left[\frac{d_n}{\bar{\tau}} \log(d_n/f_\theta) - (\frac{d_n}{\bar{\tau}} + \frac{f_\theta}{\tau}) \log(\tau \frac{d_n}{f_\theta} + \bar{\tau}) \right], \tag{2.24}$$

where $\bar{\tau} = 1 - \tau$. In this case, the C function is given by

$$C_\tau(\delta) = \frac{\delta + 1}{1 - \tau} \log(\delta + 1) - \frac{\tau\delta + 1}{\tau(1 - \tau)} \log(\tau\delta + 1). \tag{2.25}$$

The range of the tuning parameter is $\tau \in [0, 1]$. The disparities for the cases $\tau = 0$ and $\tau = 1$ are the limiting cases as $\tau \to 0$ and $\tau \to 1$, and in these cases the disparities equal the LD and KLD measures respectively.

A motivation for the construction of the GKL family, which produces a smooth bridge between LD and KLD is as follows. It is easy to see that for any two probability density functions g and f,

$$\text{GKL}_\tau(g, f) = \min_p \{\tau \text{LD}(g, p) + (1 - \tau)\text{KLD}(p, f)\}, \tag{2.26}$$

where the minimization is over the density p. The right-hand side of the above equation is actually the solution of the likelihood ratio testing problem which minimizes the likelihood disparity subject to $p \in B_f = \{t : \text{KLD}(t, f) \leq c\}$, and τ is an appropriate constant depending on c. A reversal of the roles of LD and KLD generates the celebrated power divergence family of Cressie and Read (1984).

In addition to the restriction imposed on the disparities in Equation (2.16), sometimes it is convenient, particularly in the goodness-of-fit testing scenario, to impose the restriction

$$C''(0) = 1 \tag{2.27}$$

on the defining function $C(\cdot)$. Given any convex function C having the property $C(0) = 0$, the disparity (2.7) can be centered and rescaled to the form

$$\rho_{C^*}(d_n, f_\theta) = \sum C^*(\delta) f_\theta = \sum \left(\frac{C(\delta) - C'(0)\delta}{C''(0)} \right) f_\theta, \qquad (2.28)$$

provided $C''(0) \neq 0$. Notice that $C^*(0) = 0$, and if C is convex, so is C^*. In addition, C^* satisfies the conditions (2.16) and (2.27). This does not change the estimating properties of the disparity in the sense that if $\hat{\theta}$ is the minimizer of ρ_C, then it is also the minimizer of ρ_{C^*}. All the C functions presented in Equations (2.17), (2.19), (2.21), (2.23), and (2.25) satisfy conditions (2.16) and (2.27). In the rest of this book, unless otherwise mentioned, we will assume that the C functions are standardized to satisfy $C'(0) = 0$ and $C''(0) = 1$.

Some other less known disparities have been discussed in Park and Basu (2004).

A disparity function takes the value zero when the two arguments are identical. The following two lemmas establish the upper bound of the disparity function for two arbitrary densities g and f and some ancillary results; also see Vajda (1972) and our discussion in Chapter 11. Since these results below are general and encompass both discrete and continuous models, we express the disparities with integrals in the two lemmas below.

Lemma 2.1. *Consider two probability density functions g and f, and let $\rho_C(g, f) = \int C(g/f - 1)f$ represent a disparity between the densities g and f, where the function C satisfies the disparity conditions. Let $D(g, f) = C(g/f - 1)f$, and define $C'(\infty) = \lim_{\delta \to \infty}[C(\delta)/\delta]$. Then we have the following results.*

(i) $D(g, f) \leq D(0, f)\,\chi(g \leq f) + D(g, 0)\,\chi(f < g) \leq D(0, f) + D(g, 0)$,

(ii) $D(g, f) \leq C(-1)f + C'(\infty)g$,

where $\chi(A)$ is the indicator for the set A.

Proof. First, for $g \in [0, f]$ with fixed f, look at $D(g, f)$ as a function of g.

$$\frac{\partial}{\partial g} D(g, f) = C'\left(\frac{g}{f} - 1\right) < 0, \quad \forall g \in (0, f)$$

since $C(\cdot)$ is decreasing for $\delta < 0$. Hence $D(g, f) \leq D(0, f)$ for $g \in [0, f]$. Note that $D(g, f) \leq C(-1)f$ for $\forall g \in [0, f]$.

Next, for $f \in (0, g)$ with fixed g, look at $D(g, f)$ as a function of f.

$$\frac{\partial}{\partial f} D(g, f) = -C'\left(\frac{g}{f} - 1\right)\frac{g}{f} + C\left(\frac{g}{f} - 1\right) = -A\left(\frac{g}{f} - 1\right) \leq 0, \quad \forall f \in (0, g)$$

since $A(\delta)$ is an increasing function with $A(0) = 0$. Hence $D(g, f) \leq D(g, 0)$ for $f \in [0, g)$. This proves part (i). Note that

$$D(g, 0) = \lim_{t \to 0} C(g/t - 1)t = C'(\infty)g,$$

so that part (ii) holds. □

Lemma 2.2. *Suppose that $C(-1)$ and $C'(\infty)$ are finite. Then the disparity $\rho_C(g, f)$ is bounded above by $C(-1) + C'(\infty)$.*

Proof. This follows easily from Lemma 2.1 (ii). □

2.3.2 More on the Hellinger Distance

In many ways, the Hellinger distance is the focal point of the theme of this book. Notice that the actual Hellinger distance is equal to

$$\left\{ \sum \left({d_n}^{1/2} - {f_\theta}^{1/2} \right)^2 \right\}^{1/2} = \left\{ \frac{1}{2} \mathrm{HD}(d_n, f_\theta) \right\}^{1/2} \tag{2.29}$$

where $\mathrm{HD}(d_n, f_\theta)$ is the measure defined in Equation (2.10). The distance in (2.29) is a genuine metric satisfying the triangle inequality, and the disparity measure $\mathrm{HD}(d_n, f_\theta)$ based on it is by far the most popular disparity in robust minimum distance literature. Notice that the measure in (2.10) may be written as

$$\mathrm{HD}(d_n, f_\theta) = 2 \left(2 - 2 \sum_x {d_n}^{1/2} {f_\theta}^{1/2} \right) = 4 \left(1 - \sum_x {d_n}^{1/2} {f_\theta}^{1/2} \right),$$

making it a one to one function of the term $\sum_x {d_n}^{1/2} {f_\theta}^{1/2}$. Observe that this term is also linked to the distance measure

$$\mathrm{B}(d_n, f_\theta) = -\log \left(\sum_x {d_n}^{1/2} {f_\theta}^{1/2} \right) \tag{2.30}$$

given by Bhattacharyya (1943), and the equivalence of $\mathrm{HD}(d_n, f_\theta)$ in (2.10) and the Bhattacharyya distance in (2.30) may be expressed as

$$\mathrm{HD}(d_n, f_\theta) = 4(1 - e^{-\mathrm{B}(d_n, f_\theta)}).$$

The term $\sum_x {d_n}^{1/2} {f_\theta}^{1/2}$ is sometimes referred to as the Bhattacharyya coefficient, and can be thought of as an approximate measure of the amount of overlap between two probability densities. Although it does not satisfy the triangle inequality, the Bhattacharyya distance is nonnegative, and equals zero if and only if $d_n \equiv f_\theta$, identically. See Kailath (1967), Djouadi et al. (1990), and Aherne et al. (1997), among others, for some interesting applications of the Bhattacharyya distance in solving different problems in various disciplines.

The Hellinger distance is named after Ernst David Hellinger, a German mathematician who was active in the first half of the twentieth century. He introduced a type of integral called the Hellinger integral (Hellinger, 1909) which led to the construction of the Hellinger distance. The Hellinger distance is also referred to as the Matusita distance (Matusita, 1954; Kirmani, 1971) or

the Jeffreys–Matusita distance in the literature. Both the Bhattacharyya distance and the Matusita (Hellinger) distance are extensively used as measures of separation between probability densities in many practical problems such as remote sensing (for example Landgrebe, 2003; Canty, 2007). The disparity goodness-of-fit statistic (see Chapter 8) in multinomial models based on the Hellinger distance is also referred to as the Freeman-Tukey statistic in the literature (for example Freeman and Tukey, 1950; Read, 1993).

Bhattacharyya's distance may be looked upon as a special case of the Rényi divergence (Rényi, 1961; Leise and Vajda, 1987) given by

$$\mathrm{RD}_r(d_n, f_\theta) = \frac{1}{r(r-1)} \log \left(\sum_x d_n^r(x) f_\theta^{1-r}(x) \right), \quad r \neq 0, 1. \tag{2.31}$$

The distances at $r = 0, 1$ are obtained as the limiting cases for those values, which gives

$$\mathrm{RD}_1(d_n, f_\theta) = \lim_{r \to 1} \mathrm{RD}_r(d_n, f_\theta) = \mathrm{LD}(d_n, f_\theta),$$

while

$$\mathrm{RD}_0(d_n, f_\theta) = \lim_{r \to 0} \mathrm{RD}_r(d_n, f_\theta) = \mathrm{KLD}(d_n, f_\theta),$$

where LD and KLD are as defined in Equations (2.8) and (2.9). The case $r = 1/2$ generates (four times) the Bhattacharyya distance.

The connection of the family of Rényi divergences for $r \in (0, 1)$ with that of the members of the generalized Hellinger distance family

$$\mathrm{GHD}(d_n, f_\theta) = \frac{1}{r(1-r)} \left(1 - \sum_x d_n^r(x) f_\theta^{1-r}(x) \right), \ r \in (0, 1), \tag{2.32}$$

(Simpson 1989a, and Basu, Basu and Chaudhuri, 1997) can be easily observed by comparing (2.31) and (2.32).

For the rest of the book, unless otherwise qualified, the term Hellinger distance will refer to the measure on the right-hand side of Equation (2.10).

2.3.3 The Minimum Distance Estimator and the Estimating Equations

We begin this section by demonstrating an important fact. The class of minimum distance estimators based on disparities contains the maximum likelihood estimator as a special case.

The minimum distance estimator $\hat{\theta}$ of θ, based on the disparity ρ_C, is defined by the relation

$$\rho_C(d_n, f_{\hat{\theta}}) = \min_{\theta \in \Theta} \rho_C(d_n, f_\theta) \tag{2.33}$$

provided such a minimum exists.

Consider the estimation setup described earlier in this section. We have

a discrete parametric model $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$, and suppose that a random sample $X_1, \ldots, X_n$ is available from the true distribution G with which we wish to estimate θ; the support of the random variables is assumed, without loss of generality, to be $\mathcal{X} = \{0, 1, 2, \ldots, \}$. Determination of the maximum likelihood estimator of θ corresponds to the maximization of

$$\log \prod_{i=1}^{n} f_\theta(X_i) = \sum_{i=1}^{n} \log f_\theta(X_i). \tag{2.34}$$

However, if we write the sum on the right-hand of the above equation in terms of the elements of the sample space, rather than the index i for the observation of the random sample, then this sum equals

$$n \sum_{x=0}^{\infty} d_n(x) \log f_\theta(x). \tag{2.35}$$

Dropping the dummy variable x, maximizing the above with respect to θ is equivalent to minimizing

$$-\sum d_n \log f_\theta,$$

and hence equivalent to minimizing

$$\sum d_n \log(d_n / f_\theta) \tag{2.36}$$

with respect to θ. Notice that this is the likelihood disparity between d_n and f_θ. Thus, the likelihood disparity is minimized by the maximum likelihood estimator of θ, which shows that the class of minimum distance estimators based on disparities includes the maximum likelihood estimator under discrete models.

As already described, the minimum distance estimator $\hat{\theta}$ based on the disparity ρ_C is obtained through the minimization described in Equation (2.33). In the present subsection we will describe the geometric structure of the disparities and their gradients to have a better insight on the robustness and the efficiency properties of these estimators. Under differentiability of the model, the minimum distance estimator of θ based on the disparity ρ_C is obtained by solving the estimating equation

$$-\nabla \rho_C(d_n, f_\theta) = \sum (C'(\delta)(\delta + 1) - C(\delta)) \nabla f_\theta = 0, \tag{2.37}$$

where ∇ represents the gradient with respect to θ. Letting

$$A(\delta) = C'(\delta)(\delta + 1) - C(\delta), \tag{2.38}$$

the estimating equation for θ has the form

$$-\nabla \rho_C(d_n, f_\theta) = \sum A(\delta) \nabla f_\theta = 0. \tag{2.39}$$

The function $A(\delta)$ can be suitably standardized, without changing the estimating properties of the minimum disparity estimators, so that it satisfies

$$A(0) = 0 \text{ and } A'(0) = 1. \tag{2.40}$$

These properties are automatic when the corresponding C function satisfies the conditions (2.16) and (2.27). The function $A(\delta)$, when thus standardized, is called the residual adjustment function (RAF) of the disparity. The residual adjustment function will be one of our key components in studying the properties of our minimum distance estimators.

The estimating equations of the different minimum distance estimators represented by Equation (2.39) differ only in the form of the residual adjustment function $A(\delta)$. Thus, the different properties of the minimum disparity estimators must be governed by the form of the function $A(\delta)$. Notice that $A'(\delta) = (\delta + 1)C''(\delta)$, and as $C(\cdot)$ is a strictly convex function, $A'(\delta) > 0$ for $\delta > -1$; hence $A(\cdot)$ is a strictly increasing function on $[-1, \infty)$.

As our primary motivation in proposing these minimum distance estimators is in developing a class of estimators, which has full asymptotic efficiency coupled with strong robustness properties, we will now proceed to describe the role of the RAF in determining the robustness properties of the estimators. For this purpose, we will consider a probabilistic – rather than a geometric – characterization of the outliers in a data set; however, these concepts often coincide. An element x of the sample space with a large positive value of the Pearson residual $\delta(x)$ represents a large outlier in relation to the parametric model in the sense that the actual observed proportion is much larger here than what is predicted by the model. We note that Davies and Gather (1993) also defined outliers in terms of their position relative to the model that most of the observations follow. In robust estimation, our aim is to downweight observations having large positive values of δ. This is achieved by such disparities for which the RAF $A(\delta)$ exhibits a severely dampened response to increasing δ. For a qualitative description, we will take the RAF of the likelihood disparity as the basis for comparison. Notice that the likelihood disparity is minimized by the maximum likelihood estimator. The estimating equation of the maximum likelihood estimator (the likelihood equation) can be expressed as

$$-\nabla \mathrm{LD}(d_n, f_\theta) = \sum \delta \nabla f_\theta = 0. \tag{2.41}$$

Thus, one gets $A_{\mathrm{LD}}(\delta) = \delta$, and the RAF is linear for the likelihood disparity, passing through the origin at 45 degrees to the X-axis or the δ-axis (as well as to the Y-axis). Hence the comparison of other minimum distance estimators with the maximum likelihood estimator must focus on how the other RAFs depart from linearity. The set of conditions given by (2.40) guarantees that all RAFs are tangential to the line $A_{\mathrm{LD}}(\delta) = \delta$ at the origin ($\delta = 0$).

In Figure 2.1, we have presented the graphs of the residual adjustment functions of the five common distances within the class of disparities. Notice that the RAFs for the HD, the KLD and the NCS all provide strong

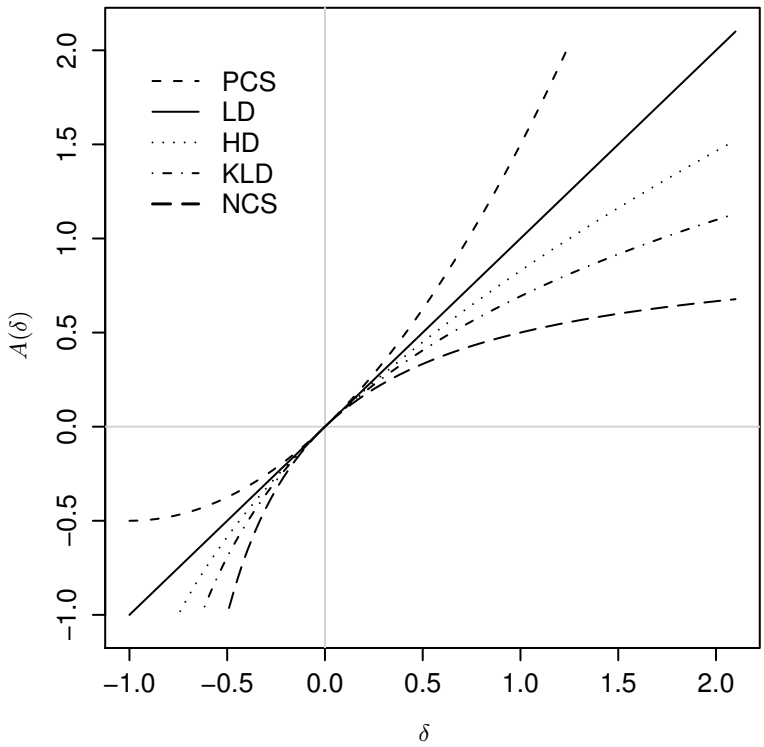

FIGURE 2.1
Residual Adjustment Functions for five common distances.

downweighting for large δ (relative to the likelihood disparity), with the NCS providing the strongest downweighting. On the other hand, the PCS actually magnifies the effect of large δ outliers rather than shrinking them. As a result, the minimum distance estimators based on the Pearson's chi-square distance are expected to be even worse than the maximum likelihood estimator in terms of robustness.

Expanding the estimating equation given by (2.39) in a Taylor series around $\delta = 0$, we get

$$-\nabla\rho_C(d_n, f_\theta) = \sum A(\delta)\nabla f_\theta = \sum \left\{\delta + \frac{A_2}{2}\delta^2 + \ldots\right\}\nabla f_\theta = 0. \qquad (2.42)$$

Comparing with Equation (2.41) we see that the leading term in the estimating function of any disparity is the same as that of the likelihood disparity which gives some intuitive justification of the asymptotic equivalence of all minimum distance estimators based on disparities and the maximum likelihood estimator. As we will see later, the quantity $A_2 = A''(0)$ in the above equation turns

out to be a key player in describing the properties of the minimum distance estimator.

2.3.4 The Estimation Curvature

A dampened response to increasing positive δ will imply that the RAF shrinks the effect of large outliers as δ increases. Under the standardizations given in Equation (2.28), all residual adjustment functions satisfy $A(0) = 0$ and $A'(0) = 1$. For the likelihood disparity, which has a linear residual adjustment function, the second and all successive derivatives of the RAF evaluated at $\delta = 0$ (or anywhere else) are equal to zero. In comparison, RAFs for which $A_2 = A''(0)$ is positive, curve locally upwards (in comparison to $A_{\mathrm{LD}}(\delta)$) at $\delta = 0$, while the reverse is observed when A_2 is negative. Thus, A_2 can be used as a measure of local robustness, with negative values of A_2 being preferred. A negative value for the A_2 parameter is achieved, for example, for the HD, the KLD, the NCS (see Figure 2.1), and all other members of the Cressie–Read family with $\lambda < 0$. Similarly this is achieved for all members of the BWHD and BWCS families with $\alpha > 1/3$. This is also true for all members of the GNED family with $\lambda > 1$, and all members of the GKL family with $\tau > 0$.

The estimation curvature A_2 also has some relevance in the context of the second-order efficiency of these minimum distance estimators. In case of multinomial models, the concept of second-order efficiency of Rao (1961) can be directly applied. The second-order efficiency E_2 of an estimator (the smaller E_2 is, the better the efficiency) T is measured by finding the minimum asymptotic variance of $U_n(\theta) - \alpha(\theta)[T - \theta] - \lambda(\theta)[T - \theta]^2$ over α and λ, where $U_n(\theta)$ is the score function in an independent and identically distributed sample. The minimum value is determined by the model and the parametrization. In the multinomial model this is minimized by the MLE (Rao, 1961, 1962). The deficiency of a minimum distance estimator within the class of disparities is a simple function of the estimation curvature A_2 and a nonnegative quantity D which depends on the model, but not on $A(\delta)$.

We present the following theorem in the context of models having a finite support linking the maximum likelihood estimator to the concept of second-order efficiency. See Rao et al. (1983), Read and Cressie (1988) and Lindsay (1994) for further discussion including a proof of the theorem.

Theorem 2.3. *Suppose the sample space $\mathcal{X} = \{0, 1, \ldots, K\}$ is finite ($K < \infty$). The second-order efficiency E_2 of a minimum distance estimator* (MDE) *based on a disparity having residual adjustment function $A(\delta)$ and estimation curvature A_2 is given by*

$$E_2(\mathrm{MDE}) = E_2(\mathrm{MLE}) + A_2^2 D,$$

where D is a nonnegative quantity depending on the model but not on the residual adjustment function $A(\delta)$.

For the PD family, the result was established by Read and Cressie (1988).

However, it turns out that the calculations presented therein depend only on the first and second derivatives of the estimating functions with respect to δ (at $\delta = 0$). If two RAFs have the same value of the estimation curvature, then they have the same first and second derivatives of the RAF at $\delta = 0$. Thus the above theorem holds generally for the class of minimum distance estimators based on disparities.

Thus, for models such as the multinomial, disparities satisfying $A_2 = 0$ generate second-order efficient estimators; apart from the likelihood disparity, this is achieved by the BWHD and BWCS families for $\alpha = 1/3$, and for the GNED for $\lambda = 1$. The latter disparity is called the negative exponential disparity (NED).

Notice that $A_2 = 0$ implies that the residual adjustment function of the corresponding distance has a second-order contact with that of the likelihood disparity at the origin. In this case, the right-hand side of Equation (2.42) is equivalent to the expression for the likelihood disparity up to the second order. In this sense, our working definition of second-order efficiency of a minimum distance estimator based on disparities is that the corresponding estimation curvature A_2 equals zero.

As mentioned in Section 1.3.3, the concept of second-order efficiency is not as simple as that of first-order efficiency, and is also subject to some controversy. Berkson (1980) has questioned the significance of this measure. Read and Cressie (1988) show that the MLE is no longer the optimal within the class of minimum distance estimators when considering the Hodges–Lehmann deficiency of the estimators.

2.3.5 Controlling of Inliers

Unlike the outliers and the large positive values of the Pearson residual generated by them, observations with fewer data than expected will generate negative values of δ, and such observations will be denoted as *inliers*. If one requires the RAF to shrink both positive and negative residuals (outliers and inliers) relative to maximum likelihood, the RAF should have the property

$$|A(\delta)| \leq |\delta| \tag{2.43}$$

for all δ. Such RAFs must cross the $A_{\mathrm{LD}}(\delta) = \delta$ line at $\delta = 0$, and hence should satisfy $A''(0) = 0$. The corresponding estimators must therefore be automatically second-order efficient. However, to satisfy the condition (2.43) the third derivative $A'''(0)$ must be negative (unless it is itself zero in which case one has to consider the derivative of the next higher order). The negative exponential disparity is an example of disparities satisfying (2.43), and it has remarkable robustness properties in spite of its second-order efficient behavior (Bhandari, Basu and Sarkar, 2006).

However, the negative exponential disparity is a rather special distance, and practically all the other robust distances within the class of disparities (including the Hellinger distance) magnify the effect of inliers while shrinking

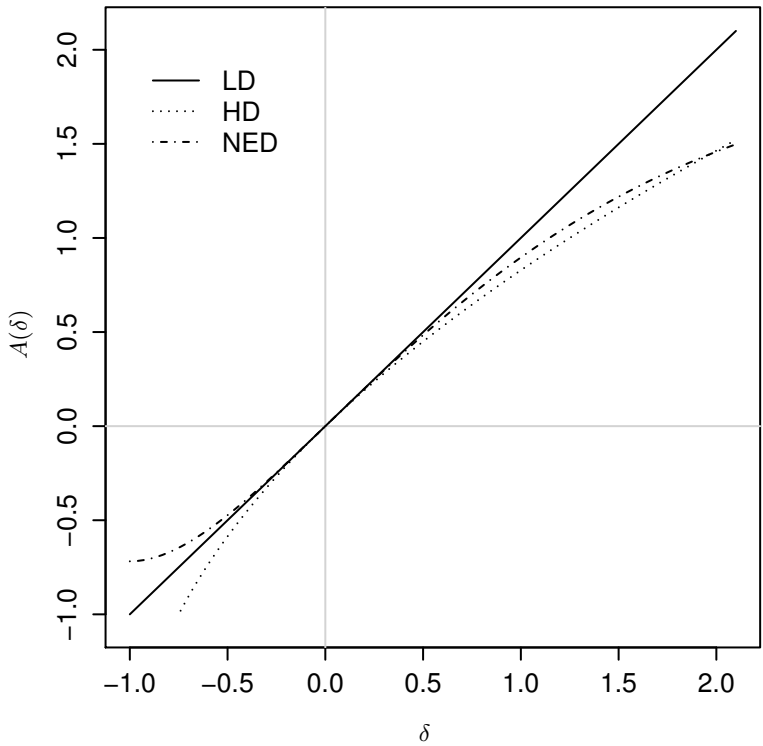

FIGURE 2.2
Residual Adjustment Functions for LD, HD and NED.

the effect of outliers. But it appears that the proper controlling of inliers is one of the keys for good small sample efficiency; as a result the estimators and tests based on these robust distances can be substantially less efficient compared to the likelihood based methods in small samples. In Chapter 6 we will consider some strategies to tackle the inlier problem, without compromising the robustness properties of these distances.

2.3.6 The Robustified Likelihood Disparity

The theory and interpretation of the minimum distance techniques used in this book are based on the two key functions; the first is the disparity generating function $C(\cdot)$, and the second is the residual adjustment function $A(\cdot)$. In the direct approach to estimation, we normally start with a distance function, and obtain the residual adjustment function by taking a derivative of the distance measure with respect to the parameter and suitably standardizing it. Sometimes we also take the reverse approach. In this case, we first define a residual adjustment function with the right properties, and then re-

construct the distance. This is done by solving the differential Equation (2.38) which recovers the disparity generating function $C(\cdot)$. Given any differentiable and increasing function $A(\delta)$, the corresponding disparity generating function obtained through the above procedure has the form

$$C(\delta) = \int_0^\delta \int_0^t A'(s)(1+s)^{-1} ds dt. \tag{2.44}$$

The *strictly increasing* condition may be replaced by *nondecreasing* over parts of the domain bounded away from zero without any technical difficulty. Lack of differentiability at one or two points at such regions of the domain are also entirely fixable if the function is continuous.

In the previous sections of this chapter we have seen that most of the features of the minimum distance estimators are governed by the smoothness properties of the residual adjustment function $A(\delta)$ and the magnitude of its derivatives at $\delta = 0$. In particular when the estimation curvature parameter A_2 equals zero, the residual adjustment function $A(\delta)$ of the disparity has a second-order contact with the residual adjustment function $A_{\mathrm{LD}}(\delta)$ of the likelihood disparity. If successive derivatives of the residual adjustment function $A(\delta)$ at $\delta = 0$ continue to be zero up to the k-th order, $A(\delta)$ has a k-th order contact with $A_{\mathrm{LD}}(\delta)$. A comparison with Equation (2.42) reveals that the estimating equation for the disparity in question is equivalent to that of the likelihood disparity up to the k-th order in that case.

In an attempt to develop a structure where the estimating equations are completely equivalent to the likelihood equations in a neighborhood of $\delta = 0$, but have powerful outlier downweighting properties for large values of δ, Chakraborty, Basu and Sarkar (2001) proposed the residual adjustment function

$$A_{\alpha,\alpha^*}(\delta) = \begin{cases} \alpha & \text{for } -1 \le \delta \le \alpha \\ \delta & \text{for } \alpha < \delta < \alpha^* \\ \alpha^* & \text{for } \delta \ge \alpha^*. \end{cases} \tag{2.45}$$

with numbers α^* and α satisfying $-1 \le \alpha < 0 < \alpha^* < \infty$. That the function will limit the impact of large δ outliers is obvious. The additional tuning parameter α provides further flexibility to the estimation procedure. The function A_{α,α^*} has the shape of the Huber's ψ function, and is constant over parts of its domain.

The corresponding disparity generating function obtained by solving the differential Equation (2.38) has the form

$$C_{\alpha,\alpha^*}(\delta) = \begin{cases} (\delta+1)\log(\alpha+1) - \alpha & \text{for } -1 \le \delta \le \alpha \\ (\delta+1)\log(\delta+1) - \delta & \text{for } \alpha < \delta < \alpha^* \\ (\delta+1)\log(\alpha^*+1) - \alpha^* & \text{for } \delta \ge \alpha^*. \end{cases} \tag{2.46}$$

Clearly, $C_{\alpha,\alpha^*}(\delta)$ is strictly convex on (α, α^*) and linear on $(-\infty, \alpha]$ and $[\alpha^*, \infty)$ with the slopes at $\delta = \alpha$ and $\delta = \alpha^*$ well defined. Hence $C_{\alpha,\alpha^*}(\delta)$

TABLE 2.1
Forms of C functions and RAFs for common disparities. For BWHD and BWCS, $\bar{\alpha} = 1 - \alpha$; for GKL, $\bar{\tau} = 1 - \tau$. For GNED, the forms are given for $\lambda > 0$ ($\lambda = 0$ case is the PCS).

Disparity	C function	RAF
LD	$(\delta+1)\log(\delta+1) - \delta$	δ
HD	$2((\delta+1)^{1/2} - 1)^2$	$2((\delta+1)^{1/2} - 1)$
PCS	$\frac{\delta^2}{2}$	$\delta + \frac{\delta^2}{2}$
NCS	$\frac{\delta^2}{2(\delta+1)}$	$1 - \frac{1}{\delta+1}$
KLD	$\delta - \log(\delta+1)$	$\log(\delta+1)$
PD	$\frac{(\delta+1)^{\lambda+1} - (\delta+1)}{\lambda(\lambda+1)} - \frac{\delta}{\lambda+1}$	$\frac{(\delta+1)^{\lambda+1} - 1}{\lambda+1}$
BWHD	$\frac{1}{2}\frac{\delta^2}{[\alpha(\delta+1)^{1/2} + \bar{\alpha}]^2}$	$\frac{\delta}{[\alpha(\delta+1)^{1/2} + \bar{\alpha}]^2} + \frac{\bar{\alpha}\delta^2}{2[\alpha(\delta+1)^{1/2} + \bar{\alpha}]^3}$
BWCS	$\frac{1}{2}\frac{\delta^2}{[\alpha(\delta+1) + \bar{\alpha}]}$	$\frac{\delta}{1+\alpha\delta} + \frac{\bar{\alpha}}{2}\left[\frac{\delta}{1+\alpha\delta}\right]^2$
SCS	$\frac{\delta^2}{\delta+2}$	$\frac{\delta(3\delta+4)}{(\delta+2)^2}$
NED	$e^{-\delta} - 1 + \delta$	$2 - (2+\delta)e^{-\delta}$
GNED	$\frac{e^{-\lambda\delta} - 1 + \lambda\delta}{\lambda^2}$	$\frac{(\lambda+1) - ((\lambda+1) + \lambda\delta)e^{-\lambda\delta}}{\lambda^2}$
GKL	$\frac{\delta+1}{\bar{\tau}}\log(\delta+1) - \frac{\tau\delta+1}{\tau\bar{\tau}}\log(\tau\delta+1)$	$\frac{1}{\tau}\log(\tau\delta+1)$
RLD	$\begin{cases} (\delta+1)\log\bar{\alpha} + \alpha \\ (\delta+1)\log(\delta+1) - \delta \\ (\delta+1)\log(1/\bar{\alpha}) - \alpha/\bar{\alpha} \end{cases}$	$\begin{cases} -\alpha & : \quad \delta < -\alpha \\ \delta & : \ -\alpha \le \delta < \alpha/\bar{\alpha} \\ \alpha/\bar{\alpha} & : \quad \delta \ge \alpha/\bar{\alpha} \end{cases}$

is convex on the entire interval $[-1, \infty)$. We will refer to the disparity generated by C_{α,α^*} as the robustified likelihood disparity (with tuning parameters α and α^*). The minimizer of the robustified likelihood disparity $\text{RLD}_{\alpha,\alpha^*}(d_n, f_\theta)$ over $\theta \in \Theta$ will be called the robustified likelihood estimator. The robustified likelihood estimators will have full asymptotic efficiency, and good robustness properties depending on the tuning parameters. The RLD has particular relevance in the context of weighted likelihood estimators, as we will see in Chapter 7.

A convenient one parameter formulation of RLD may by obtained by choosing α to be any number between $(0, 1)$, and letting $\alpha^* = \alpha/\bar{\alpha}$, where

$\bar{\alpha} = 1 - \alpha$. It is easy to see that one would recover the residual adjustment function of the likelihood disparity from (2.45) for the limiting case $\alpha \to 1$. Also, smaller values of α will expand the range over which the proposed down-weighting will be applied. It is this one parameter version of RLD that is presented in Table 2.1, where the disparity generating function $C(\cdot)$ and the residual adjustment function $A(\cdot)$ of several common disparities and families of disparities are provided.

2.3.7 The Influence Function of the Minimum Distance Estimators

As indicated in Chapter 1, the influence function of an estimator is a useful indicator of its asymptotic efficiency, as well as of its classical first-order robustness. Consider a generic disparity ρ_C, and let A represent its residual adjustment function. To find the influence function of our minimum distance estimators, we consider the ϵ contaminated version of the true density g given by

$$g_\epsilon(x) = (1-\epsilon)g(x) + \epsilon\chi_y(x).$$

Similarly $G_\epsilon(x) = (1-\epsilon)G(x) + \epsilon \wedge_y (x)$. Here $\chi_y(x)$ and $\wedge_y(x)$ are as defined in Section 1.1. Consider the minimum distance functional $T(G)$ representing the minimizer of $\rho_C(g, f_\theta)$. Let $\theta_\epsilon = T(G_\epsilon)$ be the functional obtained via the minimization of $\rho_C(g_\epsilon, f_\theta)$, which satisfies

$$\sum_x A(\delta_\epsilon(x))\nabla f_{\theta_\epsilon}(x) = 0, \tag{2.47}$$

where $\delta_\epsilon(x) = g_\epsilon(x)/f_{\theta_\epsilon}(x) - 1$. Then the influence function $\phi_G(y) = T'(y)$ of the functional T at the distribution G is the first derivative of θ_ϵ evaluated at $\epsilon = 0$. The form of the influence function of our minimum distance estimators is derived in the following theorem.

Theorem 2.4. [Lindsay (1994, Proposition 1)]. *For a disparity $\rho(\cdot,\cdot)$ associated with a corresponding estimating equation*

$$\sum_x A(\delta(x))\nabla f_\theta(x) = 0, \tag{2.48}$$

the influence function of the minimum distance functional T at G has the form

$$T'(y) = D^{-1}N,$$

where

$$N = A'(\delta(y))u_{\theta^g}(y) - E_g\Big[A'(\delta(X))u_{\theta^g}(X)\Big]$$

and

$$D = E_g\Big[u_{\theta^g}(X)u_{\theta^g}^T(X)A'(\delta(X))\Big] - \sum_x A(\delta(x))\nabla_2 f_{\theta^g}(x)$$

for $\theta^g = T(G)$ *and* $\delta(x) = g(x)/f_{\theta^g}(x) - 1$.

Proof. Direct differentiation of Equation (2.47) gives

$$\frac{\partial}{\partial \epsilon}\theta_\epsilon = D_\epsilon^{-1} N_\epsilon, \tag{2.49}$$

where

$$N_\epsilon = A'(\delta_\epsilon(y))u_{\theta_\epsilon}(y) - \sum A'(\delta_\epsilon(x))u_{\theta_\epsilon}(x)g(x)$$

and

$$D_\epsilon = \sum u_{\theta_\epsilon}(x)u_{\theta_\epsilon}^T(x)A'(\delta_\epsilon(x))g_\epsilon(x) - \sum A(\delta_\epsilon(x))\nabla_2 f_{\theta_\epsilon}(x).$$

When evaluated at $\epsilon = 0$ we get $\theta_\epsilon = \theta^g$ and $\delta_\epsilon(\cdot) = \delta(\cdot)$. Replacing these in Equation (2.49) we get the required result. □

The following is a direct corollary of the above theorem, which indicates the asymptotic efficiency of all the minimum distance estimators based on disparities at the model.

Corollary 2.5. *Consider the conditions of Theorem 2.4. When the true distribution* G *belongs to the parametric model, so that the density* $g(x) = f_\theta(x)$ *for some* $\theta \in \Theta$*, we get* $\theta^g = \theta$*,* $\delta(x) = 0$ *for all* x*, and the minimum distance estimator corresponding to the estimating equation* $\sum A(\delta(x))\nabla f_\theta(x) = 0$ *has influence function* $T'(y) = I^{-1}(\theta)u_\theta(y)$*, where* $I(\theta)$ *is the Fisher information matrix at* θ*.*

Observe that the above theorem and corollary could have been phrased simply in terms of the estimating equation, rather than linking them directly to a disparity. The influence function described in Theorem 2.4 describes all estimators obtained as the solution of estimating equations of the type (2.48) without any reference to a minimum distance problem. However, given any differentiable and increasing function $A(\delta)$, one can construct a corresponding disparity measure ρ_C having residual adjustment function $A(\delta)$ using relation (2.44). Thus, we prefer to present the above as properties of minimum distance estimators based on disparities, rather than as a class of estimators obtained as solutions of appropriate estimating equations. However, we will also consider the influence function of weighted likelihood estimators later on (Chapter 7), where the functionals are obtained as solutions of appropriate estimating equations, and may not directly correspond to the optimization of an objective function.

The most striking revelation of the above corollary is that all the minimum distance estimators based on disparities have the same influence function at the model as the maximum likelihood estimator, as is necessary if these estimators are to be asymptotically fully efficient. Thus, the influence functions

of these minimum distance estimators are not useful indicators for describing their robustness. We will demonstrate in Chapter 4 that a distinction can be made in respect to the higher order influence terms which give a better description of the stability of the minimum distance estimators.

2.3.8 ϕ-Divergences

We have mentioned in Section 2.3.1 that in this book we will follow the approach of Lindsay (1994) and develop the minimum distance estimation procedure in terms of disparities; this is done primarily to exploit the geometry of the method, and to describe the tradeoff between the robustness and the efficiency of the procedure in terms of the residual adjustment function $A(\cdot)$ of the disparity. For the sake of completeness, here we briefly describe the structure in terms of the ϕ-divergences, as presented by Csiszár (1963, 1967a,b), Ali and Silvey (1966), Pardo (2006) and others.

Consider the densities d_n and f_θ under the notation and setup of Section 2.3.1. The ϕ-divergence measure between these densities is given by

$$D_\phi(d_n, f_\theta) = \sum_{x=0}^{\infty} \phi\left(\frac{d_n(x)}{f_\theta(x)}\right) f_\theta(x), \tag{2.50}$$

where the function ϕ is a convex function defined on all nonnegative real values such that $\phi(1) = 0$. The function ϕ is also required to satisfy the conditions $0\phi(0/0) = 0$ and $0\phi(p/0) = p \lim_{u\to\infty} \phi(u)/u$. Comparing Equation (2.50) with Equation (2.7) we see that the definitions produce equivalent distances for $C(u-1) = \phi(u)$. As in the case with disparities, establishing the asymptotic properties of the minimum ϕ-divergence estimator will require additional smoothness assumptions on the function; in particular, the function ϕ will be required to be thrice continuously differentiable.

Just as the representation of the disparity in terms of the disparity generating function C is not unique, the function ϕ can also be appropriately modified to guarantee that it satisfies additional useful properties without changing either the value of the divergence or the essential properties of the function. Thus, given any such function ϕ, one can define the function

$$\phi^*(u) = \frac{\phi(u) - \phi'(1)(u-1)}{\phi''(1)}$$

in analogy with Equation (2.28). The function ϕ^* satisfies $\phi^{*'}(1) = 0$ and $\phi^{*''}(1) = 1$. When such a ϕ^* is used in the representation of the ϕ-divergence, each term in its summand is nonnegative.

A list of some specific cases of ϕ-divergences is provided in Pardo (2006, Section 1.2). There is some overlap of this list with the list of disparities presented in Table 2.1 of this book. As the other divergences in Pardo's (2006) list are of peripheral interest to us, we refer the reader to Pardo for a more expanded discussion of these divergences rather than repeating them here.

However, particular mention may be made here of the Bhattacharyya distance and the class of Rényi divergences; these distances have already been introduced in Section 2.3.2. Except for some special cases, these distances do not belong to the class of ϕ-divergences according to the definition presented here. However, in some cases the distance measure can be written in the form

$$D_\phi^h(d_n, f_\theta) = h(D_\phi(d_n, f))$$

where h is a real, increasing, differentiable function on the range of ϕ, where ϕ satisfies the usual properties. Such divergences have been called (h, ϕ) divergences by Menéndez et al. (1995), who have studied the properties of the corresponding estimators. The Bhattacharyya distance, the family of Rényi divergences, and the family of Sharma and Mittal divergences (Sharma and Mittal, 1977), belong to the class of (h, ϕ) divergences. See Pardo (2006, Section 1.2) for a definition of the Sharma and Mittal divergence.

Consider the problem of minimum distance estimation based on the disparity ρ_C and let G be the true, data generating distribution. Throughout the rest of the book, we will denote θ^g to be the *best fitting* value of the parameter if θ^g minimizes $\rho_C(g, f_\theta)$ over $\theta \in \Theta$. When $G = F_{\theta_0}$, so that the true distribution belongs to the model, θ_0 will be referred to as the *true* value of the parameter.

2.4 Minimum Hellinger Distance Estimation: Discrete Models

Beran (1977) considered the problem of parametric estimation based on the minimum Hellinger distance method. He established the asymptotic distribution of the minimum Hellinger distance estimator under the assumption that the distributions have densities with respect to the Lebesgue measure. This was a path-breaking paper which significantly influenced future research in this area. The existence and consistency results of the minimum Hellinger distance estimator will be discussed in this section following Beran's approach. However, Beran's asymptotic normality results for the minimum Hellinger distance estimator will be discussed in Chapter 3, where the development of the method in continuous models is described. Some of the robustness indicators considered by Beran will be discussed in Chapter 4.

Considerable simplifications over Beran's approach may be possible in the derivation of the asymptotic properties of the minimum Hellinger distance estimator in discrete models, and subsequent authors have further extended Beran's result in many ways. In this section we will briefly review the preliminary results of Beran, which establishes the consistency of the minimum Hellinger distance functional under appropriate conditions for general models. We will follow this up by presenting the additional results of Simpson (1987)

and Tamura and Boos (1986) which enhance Beran's consistency results. Finally, we present Simpson's proof of the asymptotic normality of the estimator under discrete models.

2.4.1 Consistency of the Minimum Hellinger Distance Estimator

We consider the parametric setup and notation of Section 2.3.1. Let $\mathcal{G}$ be the class of all distributions having densities with respect to the dominating measure. The minimum Hellinger distance functional $T(G)$ is defined on $\mathcal{G}$ by the requirement that for every G in $\mathcal{G}$,

$$\text{HD}(g, f_{T(G)}) = \inf_{\theta \in \Theta} \text{HD}(g, f_\theta), \tag{2.51}$$

where g is the density function corresponding to G, provided such a minimum exists.

Definition 2.2. We define a parametric model $\mathcal{F}$ as described in Section 1.1 to be identifiable, if for any $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$ implies $f_{\theta_1}(x) \neq f_{\theta_2}(x)$ on a set of positive dominating measure, where f_θ represents the density function of F_θ.

We now present Beran's proof of the consistency of the minimum Hellinger distance functional. Since the consistency result of Beran applies to a general model, here we will represent the Hellinger distance between the densities g and f as

$$\text{HD}(g, f) = 2 \int (g^{1/2}(x) - f^{1/2}(x))^2 dx, \tag{2.52}$$

with the integrals being replaced by sums when we are specifically dealing with the discrete model.

Lemma 2.6. [Beran (1977, Theorem 1)]. *Suppose that the model family is identifiable, and Θ is a compact subset of $\mathbb{R}^p$. Also suppose that $f_\theta(x)$ is continuous in θ for almost all x. Then,*

(i) for all $G \in \mathcal{G}$, there exists a $T(G)$ satisfying (2.51).

(ii) if $T(G)$ is unique, the functional T is continuous at G in the Hellinger topology (i.e., $T(G_n) \to T(G)$ whenever $g_n \to g$ in the Hellinger metric, where g_n is the density of G_n).

(iii) $T(F_\theta) = \theta$, uniquely, for every $\theta \in \Theta$.

Proof. (i) Consider a sequence of parameter values t_n such that $t_n \to t$. From the definition of the distance and Cauchy-Schwarz inequality, it follows that

$$\begin{aligned} |\text{HD}(g, f_{t_n}) - \text{HD}(g, f_t)| &= 4 \left| \int [f_{t_n}^{1/2}(x) - f_t^{1/2}(x)] g^{1/2}(x) dx \right| \\ &\leq 4 \left(\int \left[f_{t_n}^{1/2}(x) - f_t^{1/2}(x) \right]^2 dx \right)^{1/2}. \end{aligned}$$

The last term converges to zero from the pointwise continuity assumption and a generalized version of the dominated convergence theorem. The above shows that $h(t) = \text{HD}(g, f_t)$ is a continuous function of its argument. By the compactness of Θ, $h(t)$ achieves a minimum over $t \in \Theta$.

(ii) Suppose that the sequence $\{G_n\}$ converges to G in the Hellinger topology, i.e., $\text{HD}(g_n, g) \to 0$ as $n \to \infty$. Let $h_n(t) = \text{HD}(g_n, f_t)$. By writing $\theta = T(G)$ and $\theta_n = T(G_n)$, will show that $h(\theta_n) \to h(\theta)$.

Note that

$$\begin{aligned} |h_n(t) - h(t)| &= 4 \left| \int [g_n^{1/2}(x) - g^{1/2}(x)] f_t^{1/2}(x) dx \right| \\ &\leq 4 \left(\int \left[g_n^{1/2}(x) - g^{1/2}(x) \right]^2 dx \right)^{1/2}. \end{aligned}$$

Since g_n converges to g in the Hellinger metric, the right-hand side in the above equation converges to zero. Thus

$$\lim_{n \to \infty} \sup_t |h_n(t) - h(t)| = 0. \tag{2.53}$$

Now if $h(\theta) \geq h_n(\theta_n)$, then

$$h(\theta) - h_n(\theta_n) \leq h(\theta_n) - h_n(\theta_n),$$

and if $h_n(\theta_n) \geq h(\theta)$, then

$$h_n(\theta_n) - h(\theta) \leq h_n(\theta) - h(\theta).$$

Thus, we have

$$\begin{aligned} |h_n(\theta_n) - h(\theta)| &\leq |h_n(\theta_n) - h(\theta_n)| + |h_n(\theta) - h(\theta)| \\ &\leq 2 \sup_t |h_n(t) - h(t)|. \end{aligned} \tag{2.54}$$

Combining (2.53) and (2.54), we get

$$\lim_{n \to \infty} h(\theta_n) = h(\theta). \tag{2.55}$$

We will show that $\theta_n \to \theta$ is necessarily implied by the above. If not, by compactness of Θ there exists a subsequence $\{\theta_m\} \subset \{\theta_n\}$ such that $\theta_m \to \theta_1 \neq \theta$. By continuity of h, this implies $h(\theta_m) \to h(\theta_1)$, and by (2.55), $h(\theta_1) = h(\theta)$, which contradicts the uniqueness of the functional $T(G)$. Thus, $\theta_n \to \theta$, and the functional T is continuous in the Hellinger topology.

(iii) As the parametric family is identifiable in the sense of Definition 2.2, $\text{HD}(f_\theta, f_t)$ assumes the value zero at $t = \theta$, which uniquely minimizes $\text{HD}(f_\theta, f_t)$ over Θ. Thus, $T(F_\theta) = \theta$, uniquely. Thus, under the assumption that the identifiability condition holds, the existence of the minimum Hellinger distance functional $T(G)$, where G belongs to the model family, is automatic. □

Remark 2.1. There is nothing special about the Hellinger distance in the proof of Lemma 2.6 (iii). At a model element, the existence of any minimum distance estimator based on a disparity satisfying the disparity conditions is automatic under the identifiability condition, and $T(F_\theta) = \theta$ uniquely for the corresponding functional $T(\cdot)$.

Remark 2.2. The compactness assumption for the parameter space in Lemma 2.6 is somewhat restrictive. Beran (1977) argued that the result also applies if the parameter space Θ can be embedded in a compact space $\bar{\Theta}$, provided the distance $\mathrm{HD}(g, f_\theta)$, viewed as a function of θ, can be extended to a continuous function on $\bar{\Theta}$. We illustrate this point with the location-scale family

$$\left\{ \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma} \right) \right\}.$$

The parameter space $(-\infty, \infty) \times [0, \infty)$ is not compact by itself. However, consider the transformation $\mu = \tan(\beta_1)$, $\sigma = \tan(\beta_2)$, and the parameter space for (β_1, β_2) equals $(-\pi/2, \pi/2) \times (0, \pi/2)$. Therefore h can be extended to a continuous function on

$$\bar{\Theta} = [\pi/2, \pi/2] \times [0, \pi/2],$$

which is compact and the extended function attains a minimum in $\bar{\Theta}$. However, the minimum must occur in the interior of θ, since otherwise one must have $h(t) = 4$ for all $t \in \theta$, which is clearly impossible. Therefore the conclusions of the theorem remain valid for this location-scale model.

Although the above efficiently validates the use of Beran's technique for a much wider class of models than those where the parameter space is restricted to a compact set, there are examples where this structure fails to hold, particularly in multiparameter situations.

In this connection we present the following lemma under the additional conditions of Simpson (1987), which extends Beran's existence and continuity result. Under the existing setup and notation of this section, let $\mathcal{G}$ now denote the class of distributions G having densities with respect to the dominating measure, and for which

$$\inf_{\theta \in \Theta - H} \mathrm{HD}(g, f_\theta) > \mathrm{HD}(g, f_{\theta^*}) \tag{2.56}$$

for some compact $H \subset \Theta$ and some $\theta^* \in H$. If Θ is compact, then $H = \Theta$, and Lemma 2.6 applies with $\mathcal{G}$ containing all distributions having densities that are not singular to the model distribution.

Lemma 2.7. *Suppose that $f_\theta(x)$ is continuous in θ for each x. Then we have*

(i) For all $G \in \mathcal{G}$, $T(G)$ exists.

(ii) If $T(G)$ is unique, then $T(G_n) \to T(G)$ as $n \to \infty$ when the sequence of corresponding densities g_n converge to the density g in the Hellinger metric.

(iii) Let the parametric family $\mathcal{F}$ be identifiable. If $G = F_\theta$ for some $\theta \in \Theta$, $T(G_n) \to \theta$ as $n \to \infty$ for any sequence g_n converging to $g = f_\theta$ in the Hellinger metric.

Proof. Part (i) of this Lemma is obvious from Lemma 2.6 (i) and condition (2.56). For part (ii), note that if $\mathrm{HD}(g_n, g)$ converges to zero, $G_n \in \mathcal{G}$ eventually; consequently $T(G_n)$ exists and eventually belongs to H. Then an application of Lemma 2.6 (ii) with the parameter space restricted to H establishes the result.

For part (iii) note that by the identifiability of the parametric family, $T(G) = T(F_\theta) = \theta$, uniquely, and

$$\inf_{t \in \Theta - H} \mathrm{HD}(g, f_t) = \inf_{t \in \Theta - H} \mathrm{HD}(f_\theta, f_t) > 0$$

for any compact subset H of Θ containing θ as an interior point. Then the result follows from part (ii). □

Let the parametric family $\mathcal{F}$ be identifiable and be supported on $\mathcal{X} = \{0, 1, \ldots, \}$. Let the true data generating distribution G be a count distribution, and suppose that the parameter space is either compact, or condition (2.56) is satisfied. It then follows that $\sum_x |d_n(x) - g(x)| \to 0$ almost surely (see Devroye and Gyorfi, 1985, p. 10) where d_n is the density estimate defined in Section 2.3.1. By Lemmas 2.6 and 2.7, a minimizer of $\mathrm{HD}(g, f_\theta)$ exists. Suppose that this minimizer $T(G)$ is unique. Since

$$\mathrm{HD}(d_n, g) = 2 \sum_x (d_n^{1/2}(x) - g^{1/2}(x))^2 \leq 2 \sum_x |d_n(x) - g(x)| \to 0, \quad (2.57)$$

the minimum Hellinger distance estimator (the minimizer of $\mathrm{HD}(d_n, f_\theta)$ over Θ) converges to $T(G)$ in probability.

The following example, presented by Simpson (1987), provides a case in question where the extension of Beran's continuity condition fails and Lemma 2.7 has to be appealed to for establishing the consistency of the minimum Hellinger distance estimator.

Example 2.1. Consider the two parameter negative binomial case. Here the model is given by

$$f_\theta(x) = \frac{\Gamma(x + c^{-1})}{x! \Gamma(c^{-1})} \left(\frac{cm}{1 + cm} \right)^x \left(\frac{1}{1 + cm} \right)^{c^{-1}}, \; x = 0, 1, \ldots,$$

where $\theta = (m, c)$, $0 < m < \infty$, and $0 \leq c < \infty$. The above model generates the Poisson density with mean m for $c = 0$; as $m \to \infty$ with $c = 0$, the model eventually becomes singular with any fixed G, and $\mathrm{HD}(g, f_\theta) \to 4$. But it can be shown that $f_\theta(0) \to 1$ as $m \to \infty$ with m/c fixed, so $\mathrm{HD}(g, f_\theta) \to 4 - 4g^{1/2}(0)$ in this case. Hence as a function of θ, $\mathrm{HD}(g, f_\theta)$ does not extend continuously to the limit points of Θ, and a compaction of Θ via its conformal mapping

onto a sphere would fail to satisfy the conditions under which Lemma 2.6 can be applied.

On the other hand, it can be shown that

$$\lim_{n\to\infty} \inf_{\Theta - H_n} \mathrm{HD}(g, f_\theta) = 4 - 4g^{1/2}(0),$$

with $H_n = \{\theta = (m, c) : n^{-1} \le m \le n, 0 \le c \le n\}$. Thus, in this case, all one has to show is $\mathrm{HD}(g, f_{\theta^*}) < 4 - 4g^{1/2}(0)$ for some θ^*. If g is the model family, then the condition reduces to $f_\theta(0) < 1$, which is satisfied for all θ, and one can apply Lemma 2.7 to establish the consistency of the minimum Hellinger distance estimator under the appropriate convergence and uniqueness results. In particular, if G is a count distribution, $T(G)$ is unique, and d_n is the density estimate defined in Section 2.3.1, the functional T is consistent. ‖

Let $||\cdot||_2$ represent the L_2 norm, and suppose that the true distribution belongs to the model family. In this case, Tamura and Boos provide a simple proof of the consistency of the minimum Hellinger distance estimator under a slightly stronger version of the identifiability condition in Definition 2.2. The result is presented below.

Lemma 2.8. *Suppose that the true distribution belongs to the model and let θ represent the true value of the parameter. Consider a sequence of densities $\{g_n\}$ such that*

$$||g_n^{1/2} - f_\theta^{1/2}||_2 \to 0 \tag{2.58}$$

almost surely. Let $\{G_n\}$ be the corresponding sequence of distributions. Also, for any sequence $\{\theta_n : \theta_n \in \Theta\}$, suppose that

$$||f_{\theta_n}^{1/2} - f_\theta^{1/2}||_2 \to 0$$

implies $\theta_n \to \theta$. In addition, suppose that the minimum Hellinger distance estimator $T(G_n)$ exists for n sufficiently large. Then $T(G_n) \to \theta$ as $n \to \infty$.

Proof. Let $\theta_n = T(G_n)$. Using the triangle inequality we get

$$||f_{\theta_n}^{1/2} - f_\theta^{1/2}||_2 \le ||f_{\theta_n}^{1/2} - g_n^{1/2}||_2 + ||g_n^{1/2} - f_\theta^{1/2}||_2. \tag{2.59}$$

But from the definition of the minimum Hellinger distance

$$||f_{\theta_n}^{1/2} - g_n^{1/2}||_2 \le ||g_n^{1/2} - f_\theta^{1/2}||_2,$$

so that Equation (2.59) reduces to

$$||f_{\theta_n}^{1/2} - f_\theta^{1/2}||_2 \le 2||g_n^{1/2} - f_\theta^{1/2}||_2, \tag{2.60}$$

so that the result follows from the given conditions. □

2.4.2 Asymptotic Normality of the Minimum Hellinger Distance Estimator

To derive the asymptotic normality of the minimum Hellinger distance estimator (MHDE), we impose smoothness conditions on the model. For notational simplicity, let $s_\theta = f_\theta^{1/2}$. Suppose that for θ in the interior of Θ, s_θ is twice differentiable in L_2. These conditions may be expressed as

$$||s_t - s_\theta - \dot{s}_\theta^T(t-\theta)||_2 = o(|t-\theta|) \tag{2.61}$$

and

$$\frac{\dot{s}_t - \dot{s}_\theta - \ddot{s}_\theta(t-\theta)}{|t-\theta|} \to 0 \tag{2.62}$$

componentwise in L_2 as $|t-\theta| \to 0$. Here $\dot{s}_\theta$ $(p \times 1)$ and $\ddot{s}_\theta$ $(p \times p)$ are the indicated first and second derivatives which are in L_2, and $|a| = \max(|a_1|, |a_2|, \ldots, |a_p|)$.

We now present Simpson's proof of the asymptotic normality of the minimum Hellinger distance estimator. First we provide a preliminary result which will be a useful tool in our future calculations.

Lemma 2.9. *For any $x \in \mathcal{X}$ with $0 < g(x) < 1$, $n^{1/4}(d_n^{1/2}(x) - g^{1/2}(x)) \to 0$ with probability 1.*

Proof. Since $0 < g(x) < 1$, we also have $0 < g(x)(1-g(x)) < 1$. Thus, by the strong law of larger numbers, $d_n(x) - g(x) \to 0$ with probability 1.

From Theorem 3 of Feller (1971, page 239), we have

$$n^{\frac{1}{2}-\epsilon}(d_n(x) - g(x)) \to 0$$

for $\epsilon > 0$. In particular for $\epsilon = 1/4$ one gets

$$n^{1/4}(d_n(x) - g(x)) \to 0 \tag{2.63}$$

with probability 1. By a Taylor series expansion we then get

$$n^{1/4}(d_n^{1/2}(x) - g^{1/2}(x)) = n^{1/4}(d_n(x) - g(x))\frac{1}{2g^{1/2}(x)} + o(n^{1/4}|d_n(x) - g(x)|),$$

so that the result follows by using Equation (2.63). □

The next theorem, presented by Simpson (1987), is the primary component of the asymptotic normality proof of the minimum Hellinger distance estimator.

Theorem 2.10. *Assume the setup, conditions and definitions of Section 2.3.1, and let $X_1, \ldots, X_n$ be n independent and identically distributed observations from the true distribution G. Let the model $\mathcal{F}$ and the true distribution G be supported on $\mathcal{X} = \{0, 1, \ldots\}$. Let θ be a zero of $\nabla \mathrm{HD}(g, f_t)$ (i.e.,*

θ solves the equation $\nabla \mathrm{HD}(g, f_t) = 0$), and suppose that condition (2.61) holds at θ. If $\dot{s}_\theta \in L_1$, then

$$-\nabla \mathrm{HD}(d_n, f_\theta) \quad = \quad n^{-1}\left[2\sum_{i=1}^{n} \dot{s}_\theta(X_i) g^{-1/2}(X_i)\right] + o_p(n^{-1/2}). \quad (2.64)$$

Proof. Let

$$\begin{aligned} R_n &= -\nabla \mathrm{HD}(d_n, f_\theta) - n^{-1}\left[2\sum_{i=1}^{n} \dot{s}_\theta(X_i) g^{-1/2}(X_i)\right] \\ &= -\nabla \mathrm{HD}(d_n, f_\theta) - 2\sum_{x=0}^{\infty} \dot{s}_\theta(x) g^{-1/2}(x) d_n(x). \end{aligned}$$

We will show that $n^{1/2} R_n = o_p(1)$, which will establish the result. Since θ is a zero of $\nabla \mathrm{HD}(g, f_t)$, we have

$$-4\sum_{x=0}^{\infty} g^{1/2}(x) \dot{s}_\theta(x) = 0.$$

Then some simple algebra shows that

$$R_n = -2\sum_{x=0}^{\infty} \dot{s}_\theta(x) g^{-1/2}(x) \left[d_n^{1/2}(x) - g^{1/2}(x)\right]^2.$$

If R_{ni} denotes the ith component of R_n, and $\dot{s}_{i\theta}(x)$ denotes the ith component of $\dot{s}_\theta(x)$, we get

$$E\left\{n^{1/2}|R_{ni}|\right\} \le 2\sum_{x=0}^{\infty} |\dot{s}_{i\theta}(x)| g^{-1/2}(x) \times n^{1/2} E\{d_n^{1/2}(x) - g^{1/2}(x)\}^2. \quad (2.65)$$

Let $H_n(x) = n^{1/4}\left(d_n^{1/2}(x) - g^{1/2}(x)\right)$. From Lemma 2.9, $H_n(x) \to 0$ as $n \to \infty$, and by continuity the same convergence holds for

$$H_n^2(x) = n^{1/2}(d_n^{1/2}(x) - g^{1/2}(x))^2.$$

Since $(a^{1/2} - b^{1/2})^2 \le |a - b|$ for $a, b \ge 0$,

$$\begin{aligned} E\{d_n^{1/2}(x) - g^{1/2}(x)\}^2 &\le E|d_n(x) - g(x)| \\ &\le [E\{d_n(x) - g(x)\}^2]^{1/2} \\ &= n^{-1/2}[g(x)(1 - g(x))]^{1/2}, \end{aligned} \quad (2.66)$$

and hence the summand in (2.65) is dominated by $|\dot{s}_{i\theta}(x)|$. Since $\dot{s}_\theta \in L_1$, it will follow that $E\left\{n^{1/2}|R_{ni}|\right\} \to 0$ as $n \to 0$ for each i if we can show that $EH_n^2(x) \to 0$ as $n \to 0$. Now

$$E|H_n(x)|^{2(1+\epsilon)} \le [g(x)(1 - g(x))]^{(1+\epsilon)/2} < \infty$$

for $0 < \epsilon < 1$. Thus, $\{H_n^2(x)\}$ is uniformly integrable (Serfling, 1980, p. 14), and $EH_n^2(x) \to 0$ as $n \to \infty$.

The convergence of $n^{1/2}R_n$ to 0 follows from the convergence of $E[n^{1/2}|R_n|]$ by Markov's inequality. □

We need one more small result before establishing the normality proof. This is given in the lemma below.

Lemma 2.11. *Suppose conditions (2.61) and (2.62) hold. Then*

$$\nabla_2 \mathrm{HD}(d_n, f_\theta) = \nabla_2 \mathrm{HD}(g, f_\theta) + o_p(1),$$

where θ is a zero of $\nabla \mathrm{HD}(g, f_t)$.

Proof. We have $\nabla_2 \mathrm{HD}(d_n, f_\theta) = -4 \sum_x d_n^{1/2}(x) \ddot{s}_\theta(x)$. Thus,

$$|\nabla_2 \mathrm{HD}(d_n, f_\theta) - \nabla_2 \mathrm{HD}(g, f_\theta)| = \left| \sum (d_n^{1/2}(x) - g^{1/2}(x)) \ddot{s}_\theta(x) \right|. \quad (2.67)$$

Our result will be proved if we can show that the right-hand side the above equation tends to zero in probability. By Cauchy Schwarz inequality, the right-hand side of (2.67) is bounded by $M\{\sum_x (d_n^{1/2}(x) - g^{1/2}(x))^2\}^{1/2}$, where M is the maximum of componentwise L_2 norms of $\ddot{s}_\theta$. But by Equation (2.57)

$$\sum_x (d_n^{1/2}(x) - g^{1/2}(x))^2 \leq \sum_x |d_n(x) - g(x)| \to 0$$

as $n \to \infty$, so that the right-hand side of (2.67) converges to zero in probability. □

With this background we are now ready to state and prove the asymptotic normality of the minimum Hellinger distance estimator.

Theorem 2.12. [Simpson (1987, Theorem 2)]. *Let the true distribution G and the model $\mathcal{F}$ be supported on $\mathcal{X} = \{0, 1, \ldots, \}$. Let $X_1, \ldots X_n$ be independent and identically distributed observations from G. Suppose that (2.61) and (2.62) hold, and that $\nabla \mathrm{HD}(g, f_t)$ has a zero θ in the interior of Θ, $\nabla_2 \mathrm{HD}(g, f_\theta)$ is nonsingular, and $\dot{s}_\theta \in L_1$. Then the weak convergence of the minimum Hellinger distance estimator $\hat{\theta}_n$ to θ implies that $n^{1/2}(\hat{\theta}_n - \theta)$ has an asymptotic p-dimensional multivariate normal distribution with mean vector zero and variance $V_\theta = [\nabla_2 \mathrm{HD}(g, f_\theta)]^{-1} I(\theta) [\nabla_2 \mathrm{HD}(g, f_\theta)]^{-1}$. In particular, if $G = F_\theta$ for some $\theta \in \Theta$, $\nabla_2 \mathrm{HD}(g, f_\theta) = I(\theta)$, so that $V_\theta = I^{-1}(\theta)$.*

Proof. Under the given conditions,

$$\nabla \mathrm{HD}(d_n, f_{\hat{\theta}_n}) = 0 = \nabla \mathrm{HD}(d_n, f_\theta) + \nabla_2 \mathrm{HD}(d_n, f_\theta)(\hat{\theta}_n - \theta) + o(|\hat{\theta}_n - \theta|).$$

Together with Lemma 2.11, this immediately gives

$$(\hat{\theta}_n - \theta) = -[\{\nabla_2 \mathrm{HD}(g, f_\theta)\}^{-1} + o_p(1)] \nabla \mathrm{HD}(d_n, f_\theta)$$

Then Theorem 2.10 and a multivariate central limit theorem yields the result. When the true distribution $G = F_\theta$ belongs to the model, direct straightforward calculations show that $V_\theta = I^{-1}(\theta)$. □

2.5 Minimum Distance Estimation Based on Disparities: Discrete Models

Consider the setup of Section 2.3.1. Let ∇_j represent the gradient with respect to θ_j, the j-th component of the parameter vector θ. Similarly let ∇_{jk} and ∇_{jkl} represent the joint partial derivatives with respect to the corresponding parameters.

In an obvious extension of the notation of Equation (1.2), we define

$$u_{j\theta}(x) = \nabla_j(\log f_\theta(x))$$
$$u_{jk\theta}(x) = \nabla_{jk}(\log f_\theta(x))$$
$$u_{jkl\theta}(x) = \nabla_{jkl}(\log f_\theta(x)).$$

Let $\delta_n(x)$ be as defined in Equation (2.6). Let g represent the density of the true, data generating distribution G. The Pearson residual $\delta_g(x)$ corresponding to $g(x)$ will be defined by

$$\delta_g(x) = \frac{g(x) - f_\theta(x)}{f_\theta(x)}. \tag{2.68}$$

Also suppose that given the true density $g(x)$ and a disparity ρ_C, there exists a unique θ^g which minimizes the disparity $\rho_C(g, f_\theta)$, and that it solves the minimum disparity estimating equation

$$\sum_x A(\delta_g(x))\nabla f_\theta(x) = 0.$$

Assumption 2.1. Assume that the parametric family $f_\theta(x), \theta \in \Theta \subseteq \mathbb{R}^p$ has support $\mathcal{X}$ which is independent of θ, and $f_\theta(x) > 0$ for all $x \in \mathcal{X}$ and all $\theta \in \Theta$. Let the true density $g(x)$ also have the same support $\mathcal{X}$, where $g(x) > 0$ for all $x \in \mathcal{X}$.

Definition 2.3. The residual adjustment function $A(\delta)$ will be called regular, if it is twice differentiable and $A'(\delta)$ and $A''(\delta)(1+\delta)$ are bounded on $[-1, \infty)$, where $A'(\cdot)$ and $A''(\cdot)$ represent the first and second derivatives of $A(\cdot)$ with respect to its argument.

For parts of the argument in proving the consistency and the multivariate normality of the minimum distance estimator based on disparities, it helps to use the Hellinger residuals rather than the Pearson residuals. We define the Hellinger residuals as

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

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

Let $Y_n(x) = n^{1/2}(\Delta_n(x) - \Delta_g(x))^2$. Then the following lemma provides some bounds which are useful later in the main theorem of this chapter.

Lemma 2.13. *For any $k \in [0, 2]$ we have*

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

(ii) $E[|\delta_n(x) - \delta_g(x)|] \leq \frac{2g(x)(1-g(x))}{f_\theta(x)}.$

Proof. (i) For $a, b \geq 0$, we get $(a^{1/2} - b^{1/2})^2 \leq |a - b|$. Therefore,

$$\begin{aligned} E[Y_n^k(x)] &= n^{k/2} E\left[\left(\frac{d_n^{1/2}(x)}{f_\theta^{1/2}(x)} - \frac{g^{1/2}(x)}{f_\theta^{1/2}(x)}\right)^2\right]^k \\ &\leq n^{k/2} E\left[\left|\frac{d_n(x)}{f_\theta(x)} - \frac{g(x)}{f_\theta(x)}\right|\right]^k \\ &= n^{k/2} E\left[|\delta_n(x) - \delta_g(x)|\right]^k. \end{aligned}$$

To prove the second inequality of (i), we use the Lyapounov's inequality

$$E[|X|^\alpha]^{1/\alpha} \leq E[|X|^\beta]^{1/\beta}$$

for $\alpha < \beta$. Thus

$$\begin{aligned} E\left[|\delta_n(x) - \delta_g(x)|\right]^k &\leq [E(\delta_n(x) - \delta_g(x))^2]^{k/2} \\ &= \frac{1}{f_\theta^k(x)}[E(d_n(x) - g(x))^2]^{k/2} \\ &= \frac{1}{f_\theta^k(x)}\left[\frac{g(x)(1-g(x))}{n}\right]^{k/2}, \end{aligned}$$

and the result follows.

(ii) By definition,

$$\begin{aligned} E[|\delta_n(x) - \delta_g(x)|] &= \frac{1}{f_\theta(x)} E\left[\left|\frac{1}{n}\sum_{i=1}^n \chi(X_i = x) - g(x)\right|\right] \\ &\leq \frac{1}{f_\theta(x)}\frac{1}{n}\sum_{i=1}^n E\left[|\chi(X_i = x) - g(x)|\right] \\ &= \frac{2g(x)(1-g(x))}{f_\theta(x)}. \end{aligned}$$

The last equality follows from the fact that for a Bernoulli random variable X with parameter p, $E|X - p| = 2p(1-p)$. □

Next we prove the limiting result for the expectation of Y_n^k for values of k in $[0, 2)$.

Lemma 2.14. $\lim_n E[Y_n^k(x)] = 0$, *for* $k \in [0, 2)$.

Proof. By Lemma 2.9, $n^{1/4}(d_n^{1/2}(x) - g^{1/2}(x)) \to 0$ with probability 1 for each $x \in \mathcal{X}$. As $f(x) = x^2$ is a continuous function $n^{1/2}(d_n^{1/2}(x) - g^{1/2}(x))^2 \to 0$ with probability 1. Dividing by $f_\theta(x)$ we see that $Y_n(x)$ goes to zero with probability 1 for each x. Now by Lyapounov's inequality $E[Y_n^k(x)] \leq E[Y_n^2(x)]^{k/2}$ for $k \in [0, 2)$. But

$$\sup_n E[Y_n^2(x)] \leq \frac{g(x)(1 - g(x))}{f_\theta^2(x)}$$

from Lemma 2.13 (i), so that $\sup_n E[Y_n^k(x)]$ is bounded. The result then follows from Theorem 4.5.2 of Chung (1974). □

Lemma 2.15. [Lindsay (1994, Lemma 25)]. *Suppose that* $A(\delta)$ *is a regular RAF as described in Definition 2.3. Then there exists a finite* $B > 0$ *such that for all positive* c *and* d, *we have*

$$|A(c^2 - 1) - A(d^2 - 1) - (c^2 - d^2)A'(d^2 - 1)| \leq B(c - d)^2.$$

Proof. Consider a second-order Taylor series – in c, around d – of the function within the absolute values. Note that the function is zero at $c = d$, and its first derivative with respect to c is $2cA'(c^2 - 1) - 2cA'(d^2 - 1)$, which is also zero at $c = d$. Thus, all that is needed is to show that the second derivative of the function within absolute values is bounded. This second derivative equals $4c^2A''(c^2 - 1) + 2A'(c^2 - 1) - 2A'(d^2 - 1)$, which is bounded since $A(\delta)$ is a regular residual adjustment function. □

Let

$$a_n(x) = A(\delta_n(x) - A(\delta_g(x)) \quad \text{and} \quad b_n(x) = (\delta_n(x) - \delta_g(x))A'(\delta_g(x)).$$

Later we will need the limiting distribution of $S_{1n} = n^{1/2} \sum_x a_n(x) \nabla f_\theta(x)$, and Lemma 2.16 below shows that it is the same as the limiting distribution of $S_{2n} = n^{1/2} \sum_x b_n(x) \nabla f_\theta(x)$.

Assumption 2.2. We assume that $\sum_x g^{1/2}(x)|u_{j\theta}(x)|$ is finite for all $j = 1, 2, \ldots, p$.

Lemma 2.16. $E|S_{1n} - S_{2n}| \to 0$ *as* $n \to \infty$.

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

$$\begin{aligned} E\,|S_{1n} - S_{2n}| &= En^{1/2} \left| \sum_x (a_n(x) - b_n(x)) \nabla f_\theta(x) \right| \\ &\leq \sum_x E(\tau_n(x))|\nabla f_\theta(x)|. \end{aligned} \tag{2.69}$$

But by Lemma 2.15, $\tau_n(x) \leq Bn^{1/2}(\Delta_n(x) - \Delta_g(x))^2 = BY_n(x)$. By Lemma 2.14, $E(\tau_n(x)) \to 0$. Also by Lemma 2.13 (i),

$$E(\tau_n(x)) \leq BE(Y_n(x)) \leq B\frac{g^{1/2}(x)}{f_\theta(x)},$$

so that $Bg^{1/2}(x)|u_\theta(x)|$ bounds the summand on the right-hand of Equation (2.69). The required result then follows from Assumption 2.2. An application of Markov's inequality shows that $S_{1n} - S_{2n} \to 0$ in probability. □

Corollary 2.17. *Assume that the RAF $A(\delta)$ is regular, and Assumption 2.2 holds. Then, if*

$$V = \mathrm{Var}_g[A'(\delta_g(X))u_\theta(X)] \tag{2.70}$$

is finite, we get

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

Proof. The quantity in question is $n^{1/2}\sum_x a_n(x)\nabla f_\theta(x)$. By Lemma 2.16, the asymptotic distribution of this is the same as that of $n^{1/2}\sum_x b_n(x)\nabla f_\theta(x)$, which can be written as

$$\begin{aligned}
& n^{1/2}\sum_x(\delta_n(x) - \delta_g(x))A'(\delta_g(x))\nabla f_\theta(x) \\
= \ & n^{1/2}\sum_x(d_n(x) - g(x))A'(\delta_g(x))u_\theta(x) \\
= \ & n^{1/2}\frac{1}{n}\sum_{i=1}^n\sum_x[\chi(X_i = x) - g(x)]A'(\delta_g(x))u_\theta(x) \\
= \ & n^{1/2}\frac{1}{n}\sum_{i=1}^n\left[A'(\delta_g(X_i))u_\theta(X_i) - E_g(A'(\delta_g(X))u_\theta(X))\right].
\end{aligned}$$

The required result then follows from a simple application of the central limit theorem. □

When all the relevant expressions are evaluated at $\theta = \theta^g$, one gets

$$\begin{aligned}
S_{1n} &= n^{1/2}\sum(A(\delta_n^g(x)) - A(\delta_g^g(x)))\nabla f_{\theta^g}(x) \\
&= n^{1/2}\sum A(\delta_n^g(x))\nabla f_{\theta^g}(x) \\
&= -n^{1/2}\nabla\rho_C(d_n, f_\theta)|_{\theta=\theta^g},
\end{aligned} \tag{2.71}$$

where $\delta_n^g(x) = d_n(x)/f_{\theta^g}(x) - 1$ and $\delta_g^g(x) = g(x)/f_{\theta^g}(x) - 1$. But from

Lemma 2.16, $S_{1n} = S_{2n} + o_p(1)$, so that

$$\begin{aligned} &- n^{1/2}\nabla\rho_C(d_n, f_\theta)|_{\theta=\theta^g} \\ &\quad = S_{2n} + o_p(1) \\ &\quad = n^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[A'(\delta_g^g(X_i))u_{\theta^g}(X_i) - E_g(A'(\delta_g^g(X))u_{\theta^g}(X))\right]\right\} + o_p(1). \end{aligned} \tag{2.72}$$

On the other hand, when the true distribution belongs to the model, $G = F_\theta$ for some $\theta \in \Theta$; in this case, $\theta^g = \theta$, $A'(\delta_g^g(x)) = A'(0) = 1$, so that under f_θ, Equation (2.72) reduces to

$$\begin{aligned} -n^{1/2}\nabla\rho_C(d_n, f_\theta) &= n^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n} u_\theta(X_i)\right] + o_p(1) \\ &= Z_n(\theta) + o_p(1) \end{aligned} \tag{2.73}$$

where

$$Z_n(\theta) = n^{-1/2}\sum_{i=1}^{n} u_\theta(X_i). \tag{2.74}$$

In particular for the likelihood disparity – leading to the likelihood equation – the above relation is exact (the additional $o_p(1)$ term is absent). The relation (2.73) gives partial indication of the asymptotic efficiency of the minimum distance estimator based on the disparity ρ_C.

Going back to Lemma 2.16, the following corollary obtains the representation of the minimum Hellinger distance estimator in Theorem 2.10 of Section 2.4 as the special case of Lemma 2.16. When evaluated at the true parameter θ^g, we get, from the above lemma,

$$\begin{aligned} -\nabla\rho_C(d_n, f_\theta)|_{\theta=\theta^g} &= n^{-1/2}S_{1n}|_{\theta=\theta^g} \\ &= n^{-1/2}S_{2n}|_{\theta=\theta^g} + o_p(n^{-1/2}) \\ &= \sum_x (\delta_n^g(x) - \delta_g^g(x))A'(\delta_g^g(x))\nabla f_{\theta^g}(x) + o_p(n^{-1/2}). \end{aligned} \tag{2.75}$$

We then have the following corollary.

Corollary 2.18. *Suppose $\rho_C \equiv \mathrm{HD}$ is the Hellinger distance in (2.10), and let θ^g be the best fitting parameter which solves $\nabla\mathrm{HD}(g, f_\theta) = 0$. Then*

$$\begin{aligned} -\nabla\rho_C(d_n, f_\theta)|_{\theta=\theta^g} &= -\nabla\mathrm{HD}(d_n, f_\theta)|_{\theta=\theta^g} \\ &= \frac{1}{n}\sum_{i=1}^{n} 2\dot{s}_{\theta^g}(X_i)g^{-1/2}(X_i) + o_p(n^{-1/2}), \end{aligned}$$

where $s_t = f_t^{1/2}$, and $\dot{s}_t$ represents the first derivative of s_t with respect to t.

Proof. By (2.75) we have

$$-\nabla \mathrm{HD}(d_n, f_\theta)|_{\theta=\theta^g} = \sum_x (\delta_n^g(x) - \delta_g^g(x)) A'_{\mathrm{HD}}(\delta_g^g(x)) \nabla f_{\theta^g}(x) + o_p(n^{-1/2}). \tag{2.76}$$

By replacing the values of δ_n^g, δ_g^g, $A'_{\mathrm{HD}}(\cdot)$, and by observing that $\sum_{x=0}^{\infty} \dot{s}_{\theta^g}(x) g^{1/2}(x) = 0$, we get

$$\sum_x (\delta_n^g(x) - \delta_g^g(x)) A'_{\mathrm{HD}}(\delta_g^g(x)) \nabla f_{\theta^g}(x) = \sum_x \dot{s}_{\theta^g}(x) g^{-1/2}(x) d_n(x).$$

Writing the term on the right-hand side of the above equation as a sum of the observation index i, and replacing the same in Equation (2.76), the required result follows. Thus, the representation presented by Simpson in (2.64) is a special case of the representation in (2.75) for the Hellinger distance. □

Definition 2.4. Suppose that the densities in the parametric family $\{f_\theta : \theta \in \Theta\}$ have common support K^*. Then the true density g is called compatible with the family $\{f_\theta\}$ if K^* is also the support of g.

Suppose $X_1, \ldots, X_n$ are n independent and identically distributed observations from a discrete distribution G modeled by $\mathcal{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ and let $\mathcal{X} = \{0, 1, \ldots, \}$. Let g and $\{f_\theta\}$ represent the corresponding densities. Consider a disparity $\rho_C(d_n, f_\theta)$, where C is the disparity generating function, and let $A(\cdot)$ be the associated residual adjustment function. Let θ^g be the best fitting value of the parameter. We make the following assumptions for the proof of our main theorem.

(A1) The model family $\mathcal{F}$ is identifiable in the sense of Definition 2.2.

(A2) The probability density functions f_θ of the model distributions have common support so that the set $\mathcal{X} = \{x : f_\theta(x) > 0\}$ is independent of θ. Also, the true distribution g is compatible with the model family $\{f_\theta\}$ of densities in the sense of Definition 2.4.

(A3) There exists an open subset ω of Θ for which the best fitting parameter θ^g is an interior point and for almost all x the density $f_\theta(x)$ admits all third derivatives of the type $\nabla_{jkl} f_\theta(x)$ for all $\theta \in \omega$.

(A4) The matrix

$$J_g = E_g[u_{\theta^g}(X) u_{\theta^g}^T(X) A'(\delta_g^g(X))] - \sum A(\delta_g^g(x)) \nabla_2 f_{\theta^g}(x)$$

is positive definite where $\delta_g^g(x) = g(x)/f_{\theta^g}(x) - 1$ and $\nabla_2 f_\theta(x)$ is the $p \times p$ matrix of second derivatives of $f_\theta(x)$ having $\nabla_{ij} f_\theta(x)$ as its (i, j)th element. Notice that J_g is the same as the matrix D defined in Theorem 2.4.

(A5) The quantities

$$\sum_x g^{1/2}(x)|u_{j\theta}(x)| \ , \sum_x g^{1/2}(x)|u_{j\theta}(x)||u_{k\theta}(x)| \text{ and } \sum_x g^{1/2}(x)|u_{jk\theta}(x)|$$

are bounded for all j and k and all $\theta \in \omega$.

(A6) For almost all x there exist functions $M_{jkl}(x), M_{jk,l}(x), M_{j,k,l}(x)$ that dominate, in absolute value,

$$u_{jkl\theta}(x), \quad u_{jk\theta}(x)u_{l\theta}(x) \quad \text{and} \quad u_{j\theta}(x)u_{k\theta}(x)u_{l\theta}(x)$$

for all j, k, l, and that are uniformly bounded in expectation with respect to g and f_θ for all $\theta \in \omega$.

(A7) The RAF $A(\delta)$ is regular in the sense of Definition 2.3, and K_1 and K_2 represent the bounds of $A'(\delta)$ and $A''(\delta)(1+\delta)$ respectively.

Theorem 2.19. [Lindsay (1994, Theorem 33)]. *Suppose that Assumptions (A1)–(A7) hold. Then there exists a consistent sequence θ_n of roots to the minimum disparity estimating equations in (2.39). Also the asymptotic distribution of $n^{1/2}(\theta_n - \theta^g)$ is p-dimensional multivariate normal with mean vector 0 and covariance matrix $J_g^{-1}V_gJ_g^{-1}$, where V_g is the quantity defined in (2.70) evaluated at $\theta = \theta^g$.*

Proof. To prove consistency, we will use the arguments of Lehmann (1983, page 430). Consider the behavior of $\rho_C(d_n, f_\theta)$ on a sphere Q_a which has radius a and center at θ^g. We will show that for a sufficiently small a, the probability tends to 1 that $\rho_C(d_n, f_\theta) > \rho_C(d_n, f_{\theta^g})$ for θ on the surface of Q_a, so that the disparity has a local minimum with respect to θ in the interior of Q_a. At a local minimum the estimating equations must be satisfied. Therefore, for any $a > 0$ sufficiently small, the minimum disparity estimating equations have a solution θ_n within Q_a with probability tending to 1 as $n \to \infty$.

Taking a Taylor series expansion of $\rho_C(d_n, f_\theta)$ about $\theta = \theta^g$ we get

$$\begin{aligned}
&\rho_C(d_n, f_{\theta^g}) - \rho_C(d_n, f_\theta) \\
&\qquad = -\Big[\sum_j (\theta_j - \theta_j^g)\nabla_j \rho_C(d_n, f_\theta)\Big|_{\theta=\theta^g} \\
&\qquad\quad + \frac{1}{2}\sum_{j,k}(\theta_j - \theta_j^g)(\theta_k - \theta_k^g)\nabla_{jk}\rho_C(d_n, f_\theta)\Big|_{\theta=\theta^g} \\
&\qquad\quad + \frac{1}{6}\sum_{j,k,l}(\theta_j - \theta_j^g)(\theta_k - \theta_k^g)(\theta_l - \theta_l^g)\nabla_{jkl}\rho_C(d_n, f_\theta)\Big|_{\theta=\theta^*}\Big] \\
&\qquad = S_1 + S_2 + S_3 \quad \text{(say)}, \qquad (2.77)
\end{aligned}$$

where θ^* lies on the line segment joining θ^g and θ; θ_j and θ_j^g represent the j-th component of θ and θ^g respectively. We will inspect the linear, quadratic, and cubic terms one by one and determine their proper limits.

For the linear term S_1 in (2.77), we have

$$\nabla_j \rho_C(d_n, f_\theta)|_{\theta=\theta^g} = -\sum_x A(\delta_n^g(x))\nabla_j f_{\theta^g}(x) \tag{2.78}$$

where $\delta_n^g(x)$ is $\delta_n(x)$ evaluated at $\theta = \theta^g$, and we will show that the right-hand side converges to $-\sum_x A(\delta_g^g(x))\nabla_j f_{\theta^g}(x)$ by showing that the difference converges to 0 in probability. Since $A'(\delta)$ is bounded by K_1, the absolute value of the difference is bounded by

$$K_1 \sum_x \left|\delta_n^g(x) - \delta_g^g(x)\right| |\nabla_j f_{\theta^g}(x)|. \tag{2.79}$$

We will show that the expected value of the above quantity goes to zero. By Lemma 2.13 (i), $E[|\delta_n^g(x) - \delta_g^g(x)|]$ is bounded above by $n^{-1/2}$ times a finite quantity. Thus, $E[|\delta_n^g(x) - \delta_g^g(x)|] \to 0$ as $n \to \infty$. Again from the bound in Lemma 2.13 (ii), we get

$$E[K_1 \sum_x |\delta_n^g(x) - \delta_g^g(x)||\nabla f_{\theta^g}(x)|] \le 2K_1 \sum_x g^{1/2}(x)|u_{\theta^g}(x)|,$$

and the right-hand side is finite by assumption. Thus, by the dominated convergence theorem, the expectation of the quantity in (2.79) goes to zero. By Markov's inequality, the quantity in (2.79) itself goes to zero in probability. Thus,

$$\sum_x A(\delta_n^g(x))\nabla f_{\theta^g}(x) \to \sum_x A(\delta_g^g(x))\nabla f_{\theta^g}(x), \tag{2.80}$$

in probability. But the quantity on the right-hand side of (2.80) is zero by the definition of the minimum disparity estimating equations. Thus, with probability tending to 1, $|S_1| < pa^3$, where p is the dimension of θ and a is the radius of the sphere Q_a.

For the quadratic term S_2 in (2.77), we have

$$\nabla_{jk}\rho_C(d_n, f_\theta)_{\theta=\theta^g} = -[\nabla_k \sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^g}].$$

We will show that $\nabla_k \sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^g}$ converges to $-J_g^{jk}$, the negative of the (j,k)-th term of J_g, with probability tending to 1 so that $2S_2$ converges to a negative definite quadratic form. The first term of $\nabla_k \sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^g}$ equals

$$-\sum_x A'(\delta_n^g(x))(1+\delta_n^g(x))u_{j\theta^g}(x)u_{k\theta^g}(x)f_{\theta^g}(x) \tag{2.81}$$

which will be shown to converge to the term

$$-\sum_x A'(\delta_g^g(x))(1+\delta_g^g(x))u_{j\theta^g}(x)u_{k\theta^g}(x)f_{\theta^g}(x). \tag{2.82}$$

Consider the absolute difference between the two terms. By doing a one term Taylor series expansion of the difference

$$\left|A'(\delta_n^g(x))(1+\delta_n^g(x)) - A'(\delta_g^g(x))(1+\delta_g^g(x))\right|$$

in δ_n around δ_g we see that

$$(K_1+K_2)\sum_x |\delta_n^g(x)-\delta_g^g(x)||u_{j\theta^g}(x)||u_{k\theta^g}(x)|f_{\theta^g}(x)$$

bounds the absolute difference between (2.81) and (2.82). Since by assumption

$$\sum_x g^{1/2}(x)|u_{j\theta^g}(x)||u_{k\theta^g}(x)| < \infty,$$

the absolute difference between (2.81) and (2.82) goes to zero in probability by an argument similar to that of the linear term.

Next we will show that

$$\sum_x A(\delta_n^g(x))\nabla_{jk} f_{\theta^g}(x) \text{ converges to } \sum_x A(\delta_g^g(x))\nabla_{jk} f_{\theta^g}(x).$$

The absolute difference is bounded by $K_1\sum_x |\delta_n^g(x)-\delta_g^g(x)||\nabla_{jk}f_{\theta^g}(x)|$. But note that $\frac{\nabla_{jk}f_\theta(x)}{f_\theta(x)} = u_{jk\theta}(x) + u_{j\theta}(x)u_{k\theta}(x)$. Hence by assumption this difference goes to zero in probability by a similar argument. Thus, $\nabla_k \sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^g}$ converges to $-J_g^{jk}$. Therefore

$$\begin{aligned} 2S_2 = &\sum_{j,k}\Big\{[\nabla_k \sum_x A(\delta_n(x))\nabla_j f_{\theta^g}(x)] - [-J_g^{jk}]\Big\}(\theta_j-\theta_j^g)(\theta_k-\theta_k^g) \\ &+\sum_{j,k}\Big\{-J_g^{jk}(\theta_j-\theta_j^g)(\theta_k-\theta_k^g)\Big\}. \end{aligned}$$

The absolute value of the first term is less than p^2a^3 with probability tending to 1. The second term is a negative definite quadratic form in the variables $(\theta_j-\theta_j^g)$. Letting λ_1 be the largest eigenvalue of J_g, the quadratic form is less than $\lambda_1 a^2$. Combining the two terms, we see that there exists $c>0$ and $a_0>0$, such that for $a<a_0$, $S_2 < -ca^2$ with probability tending to 1.

For the cubic term S_3 in (2.77), we have

$$\nabla_{jkl}\rho_C(d_n,f_\theta)|_{\theta=\theta^*} = -\nabla_{kl}\sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^*}.$$

In this case, the terms are calculated at θ^* (not θ^g). We will show that the quantities are bounded in absolute value. Let us look at the terms of $\nabla_{kl}\sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^*}$ one by one. We use the notation

$$\delta_n^*(x) = \frac{d_n(x)-f_{\theta^*}(x)}{f_{\theta^*}(x)}.$$

$\sum_x A''(\delta_n^*(x))(\delta_n^*(x)+1)^2 u_{j\theta^*}(x)u_{k\theta^*}(x)u_{l\theta^*}(x)f_{\theta^*}(x)$: The sum will be bounded in absolute value by a constant times $|\delta_n^*(x)+1|M_{j,k,l}(x)f_{\theta^*}(x) = d_n(x)M_{j,k,l}(x)$. By the central limit theorem, the sum converges to the expectation of $M_{j,k,l}(X)$ (with respect to the density g). Thus, by assumption this term is bounded.

$\sum_x A'(\delta_n^*(x))(\delta_n^*(x)+1)u_{j\theta^*}(x)u_{k\theta^*}(x)u_{l\theta^*}(x)f_{\theta^*}(x)$: The sum will again be bounded in absolute value by a constant times $|\delta_n^*(x)+1|M_{j,k,l}(x)f_{\theta^*}(x) = d_n(x)M_{j,k,l}(x)$. The boundedness of this term therefore follows as in the case of the previous term.

$\sum_x A'(\delta_n^*(x))(\delta_n^*(x)+1)u_{jk\theta^*}(x)u_{l\theta^*}(x)f_{\theta^*}(x)$: The sum will be bounded in absolute value by a constant times $|\delta_n^*(x)+1|M_{jk,l}(x)f_{\theta^*}(x) = d_n(x)M_{jk,l}(x)$. Its boundedness follows similarly.

$\sum_x A(\delta_n^*(x))\nabla_{jkl}f_{\theta^*}(x)$: We can write

$$|A(\delta)| = |\int_0^\delta A'(x)dx| \leq K_1|\delta|,$$

so that

$$\begin{aligned}|A(\delta_n^*(x)| &\leq K_1|d_n(x)/f_{\theta^*}(x) - 1| \\ &\leq K_1|d_n(x)/f_{\theta^*}(x) + 1| = \frac{K_1}{f_{\theta^*}(x)}(d_n(x) + f_{\theta^*}(x)).\end{aligned}$$

Also note that

$$\begin{aligned}\left|\frac{\nabla_{jkl}f_\theta(x)}{f_\theta(x)}\right| =& u_{jkl\theta}(x) + u_{jk\theta}(x)u_{l\theta}(x) + u_{jl\theta}(x)u_{k\theta}(x) \\ &+ u_{j\theta}(x)u_{kl\theta}(x) + u_{j\theta}(x)u_{k\theta}(x)u_{l\theta}(x).\end{aligned}$$

Thus, the summand is bounded by $K_1|d_n(x) + f_{\theta^*}(x)|M(x)$, where

$$M(x) = M_{jkl}(x) + M_{jk,l}(x) + M_{jl,k}(x) + M_{j,kl}(x) + M_{j,k,l}(x),$$

so that this sum is also bounded with probability tending to 1. Hence we have $|S_3| < ba^3$ on the sphere Q_a with probability tending to 1.

Combining the three inequalities, we see that

$$\max(S_1 + S_2 + S_3) < -ca^2 + (b+p)a^3,$$

which is less than zero for $a < c/(b+p)$.

Thus, for any sufficiently small a there exists a sequence of roots $\theta_n = \theta_n(a)$ to the minimum disparity estimating equations such that $P(||\theta_n - \theta^g|| < a)$ converges to 1, where $||\cdot||_2$ represents the L_2 norm. It remains to show that we can determine such a sequence independently of a. Let θ_n^* be the root which is closest to θ^g. This exists because the limit of a sequence of roots is again a root by the continuity of the disparity as a function of the parameter. This completes the proof of the consistency part.

For the multivariate normality, let us expand $\sum_x A(\delta_n(x))\nabla_j f_\theta(x)$ about $\theta = \theta^g$ to obtain

$$\begin{aligned}
&\sum_x A(\delta_n(x))\nabla_j f_\theta(x) \\
&= \sum_x A(\delta_n^g(x))\nabla_j f_{\theta^g}(x) \\
&\quad + \sum_k (\theta_k - \theta_k^g)\nabla_k \sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^g} \\
&\quad + \frac{1}{2}\sum_{k,l}(\theta_k - \theta_k^g)(\theta_l - \theta_l^g)\nabla_{kl}\sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta'}
\end{aligned}$$

where $\theta = \theta'$ is a point on the line segment connecting θ and θ^g. Next we will replace θ by θ_n where θ_n is a solution of the minimum disparity estimating equation, which can be assumed to be consistent by the previous part. The left-hand side of the above equation then becomes zero and the equation can be rewritten as

$$\begin{aligned}
&- n^{1/2}\sum_x A(\delta_n^g(x))\nabla_j f_{\theta^g}(x) \\
&= n^{1/2}\sum_k (\theta_{nk} - \theta_k^g)\Big[\nabla_k \sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta^g} \\
&\qquad\qquad + \frac{1}{2}\sum_l (\theta_{nl} - \theta_l^g)\nabla_{kl}\sum_x A(\delta_n(x))\nabla_j f_\theta(x)|_{\theta=\theta'}\Big]
\end{aligned} \tag{2.83}$$

But

$$n^{1/2}\sum_x A(\delta_n^g(x))\nabla_j f_{\theta^g}(x) = n^{1/2}\sum_x \big\{A(\delta_n^g(x)) - A(\delta_g^g(x))\big\}\nabla_j f_{\theta^g}(x)$$

has a multivariate normal distribution with mean zero and variance V_g, by Corollary 2.17. As argued in the consistency part, the first term within the bracketed quantity in the right-hand side of (2.83) converges to J_g with probability tending to 1, while the second term within the brackets is an $o_p(1)$ term. Then it follows from Lehmann (1983, Lemma 4.1) that the asymptotic distribution of $n^{1/2}(\theta_n - \theta^g)$ is multivariate normal with mean zero and covariance matrix $J_g^{-1}V_g J_G^{-1}$. □

Corollary 2.20. *Assume the conditions of Theorem 2.19. In addition, suppose that the true distribution belongs to the model ($G = F_\theta$ for some $\theta \in \Theta$). If θ_n represents the minimum distance estimator corresponding to a disparity satisfying the conditions in Definition 2.1, then $n^{1/2}(\theta_n - \theta)$ has an asymptotic normal distribution with mean vector 0 and covariance matrix $I^{-1}(\theta)$, where $I(\theta)$ is the Fisher information about θ in f_θ.*

Proof. When $G = F_\theta$, we get $\theta^g = T(G) = T(F_\theta) = \theta$. In this case, $\delta_g(x) = 0$, $A(\delta_g(x)) = 0$, $A'(\delta_g(x)) = 1$ and $J_g = I(\theta)$. Also $V_g = I(\theta)$, so that one gets

$$J_g^{-1} V_g J_g^{-1} = I^{-1}(\theta)$$

and the result holds. □

When the model is true, i.e., $G = F_\theta$ for some $\theta \in \Theta$, Equations (2.71), (2.72) and (2.73) show that the left-hand side of Equation (2.83) equals $Z_n(\theta) + o_p(1)$. In addition, since the bracketed quantity on the right-hand side of Equation (2.83) now converges in probability to $I(\theta)$, Equation (2.83) leads to the relation

$$\theta_n = \theta + n^{-1/2} I^{-1}(\theta) Z_n(\theta) + o_p(n^{-1/2}), \tag{2.84}$$

as one would get when the estimator θ_n is first-order efficient, where $Z_n(\theta)$ is as in Equation (2.74).

Corollary 2.21. *The influence function approximation (1.14) is valid for the minimum disparity functionals.*

Proof. From Equation (2.71) and Lemma 2.16, we get

$$-n^{1/2} \nabla \rho_C(d_n, f_\theta)|_{\theta=\theta^g} = n^{12} \sum (\delta_n^g - \delta_g^g) A'(\delta_g^g) \nabla f_{\theta^g} + o_p(1).$$

Using Equation (2.83), it then follows,

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta^g) &= -n^{1/2} J_g^{-1} \nabla \rho_C(d_n, f_\theta)|_{\theta=\theta^g} + o_p(1). \\ &= n^{1/2} J_g^{-1} \sum (\delta_n^g - \delta_g^g) A'(\delta_g^g) \nabla f_{\theta^g} + o_p(1). \\ &= n^{1/2} J_g^{-1} \sum (d_n - g) A' \left(g/f_{\theta^g} - 1\right) u_{\theta^g} + o_p(1). \\ &= n^{1/2} \left[\frac{1}{n} \sum_{i=1}^{n} T'(X_i) \right] + o_p(1) \end{aligned} \tag{2.85}$$

where

$$T'(y) = J_g^{-1} \left[A'(\delta_g^g(y)) u_{\theta^g}(y) - E_g(A'(\delta_g^g(X)) u_{\theta^g}(X)) \right],$$

and J_g and δ_g^g are as defined in Assumption (A4) of this section. Note that the form of $T'(y)$ above is exactly same as the one obtained in Theorem 2.4. Thus, the linearization of the minimum distance estimators based on the influence function approximation (1.14) holds. □

Remark 2.3. The most important component of the consistency and the asymptotic normality proofs of Theorem 2.19 are the convergences of the linear, quadratic, and cubic terms of the derivatives of the distance in its Taylor series expansion. In fact, once the three convergences have been established, the remaining steps in the proof of the theorem above are routine. In the

subsequent chapters when we undertake consistency and normality proofs of different variants of our minimum distance estimator, we will simply work out the proofs of these three components. When the true distribution g belongs to the model, i.e., $g = f_{\theta_0}$ and $\theta^g = \theta_0$ for some $\theta_0 \in \Theta$, these convergences may be stated, under the assumptions and notation of Theorem 2.19, as:

1. In case of the linear term one has the convergence result

$$\nabla_j \rho_C(d_n, f_\theta)|_{\theta=\theta_0} = -\sum_x d_n(x) u_{j\theta_0}(x) + o_p(n^{-1/2}). \tag{2.86}$$

   This provides the key distributional result in the asymptotic normality of the minimum distance estimator. This result has been encountered several times in this chapter; see, for example, Equation (2.73). Theorem 2.19 proves the more general version of the result for an arbitrary density g not necessarily in the model.

2. For the quadratic term, one has the convergence

$$\nabla_{jk} \rho_C(d_n, f_\theta)|_{\theta=\theta_0} = \sum_x f_{\theta_0}(x) u_{j\theta_0}(x) u_{k\theta_0}(x) + o_p(1) = I_{jk}(\theta_0) + o_p(1), \tag{2.87}$$

   so that the $p \times p$ matrix of second derivatives of the disparity converges to the Fisher information matrix $I(\theta_0)$. Theorem 2.19 proves the more general convergence to J_g, which reduces to $I(\theta_0)$ under model conditions.

3. In case of the cubic term there exists a finite positive constant γ such that, with probability tending to 1,

$$\left|\nabla_{jkl} \rho_C(d_n, f_\theta)|_{\theta=\theta^*}\right| < \gamma, \tag{2.88}$$

   where θ^* lies on the line segment joining θ_0 and $\hat{\theta}_n$, the minimum disparity estimator.

2.6 Some Examples

Example 2.2. Here we consider a chemical mutagenicity experiment. These data were analyzed previously by Simpson (1987). The details of the experimental protocol are available in Woodruff et al. (1984). In a sex linked recessive lethal test in Drosophila (fruit flies), the experimenter exposed groups of male flies to different doses of a chemical to be screened. Each male was then mated with unexposed females. Sampling 100 daughter flies from each male (roughly), the number of daughters carrying a recessive lethal mutation on

TABLE 2.2
Fits of the Poisson model to the Drosophila data using several estimation methods: First experimental run.

	Recessive lethal count						
	0	1	2	3	4	≥ 5	$\hat{\theta}$
Observed	23	3	0	1	1	0	
LD	19.59	7.00	1.25	0.15	0.01	-	0.357
LD + D	24.95	2.88	0.17	0.01	-	-	0.115
HD	24.70	3.09	0.19	0.01	-	-	0.125
$PD_{-0.9}$	26.17	1.77	0.06	-	-	-	0.068
PCS	13.89	9.74	3.42	0.80	0.14	0.02	0.701
NED	24.79	3.02	0.18	0.01	-	-	0.122
$BWHD_{1/3}$	21.44	5.73	0.76	0.07	-	-	0.267
SCS	24.87	2.95	0.18	0.01	-	-	0.119
$BWCS_{0.2}$	24.30	3.45	0.24	0.01	-	-	0.142
$GKL_{1/3}$	24.73	3.07	0.19	0.01	-	-	0.124
$RLD_{1/3}$	24.92	2.90	0.17	0.01	-	-	0.117

TABLE 2.3
Fits of the Poisson model to the Drosophila data using several estimation methods: Second experimental run.

	Recessive lethal count						
	0	1	2	3	4	≥ 5	$\hat{\theta}$
Observed	23	7	3	0	0	1 (91)	
LD	1.60	4.88	7.47	7.61	5.82	6.62	3.0588
LD + D	22.93	9.03	1.78	0.23	0.02	-	0.3939
HD	23.63	8.59	1.56	0.19	0.02	-	0.3637
$PD_{-0.9}$	25.79	7.13	0.98	0.09	0.01	-	0.2763
PCS	-	-	-	-	-	34	32.5649
NED	22.85	9.08	1.80	0.24	0.02	-	0.3973
$BWHD_{1/3}$	22.99	9.00	1.76	0.23	0.02	-	0.3913
SCS	23.24	8.84	1.68	0.21	0.02	-	0.3805
$BWCS_{0.2}$	22.58	9.24	1.89	0.26	0.03	-	0.4094
$GKL_{1/3}$	23.22	8.85	1.69	0.21	0.02	-	0.3813
$RLD_{1/3}$	23.75	8.52	1.53	0.18	0.02	-	0.3588

the X chromosome was noted. The data set consisted of the observed frequencies of males having $0, 1, 2, \ldots$ recessive lethal daughters. For our purpose, we consider two specific experimental runs, those on day 28 and the second run of day 177. In this example, we will refer to them as the first and the second experimental runs. The data are presented in Tables 2.2 and 2.3.

Poisson models are fitted to the data for both experimental runs using several different methods of parameter estimation within our minimum dis-

tance class. A quick look at the observed frequencies for the two experimental runs reveals that there are two mild outliers in the first experimental run. In comparison the second experimental run contains a huge outlier – an exceptionally large count – where one male is reported to have produced 91 daughters with the recessive lethal mutation. Thus, between the fitted models of these two experimental runs, all kinds of robust behavior (or lack thereof) of the different minimum distance techniques can be demonstrated.

In each of Tables 2.2 and 2.3, the expected frequencies corresponding to the different methods are provided in the body of the table, while the distances (together with the tuning parameters) and the parameter estimates are described in the first and the last column of the table respectively. Thus, the expected frequencies and the estimator for the LD row are based on full data maximum likelihood; the LD + D row represents the results of fitting the model by the method of maximum likelihood after a qualitative deletion of the outlier(s), i.e., removing the observations at 3 and 4 in case of the first experimental run, and the observation at 91 for the second experimental run. A '-' represents an expected frequency smaller than 0.01. All the abbreviations of the distance names are as described in this chapter.

Several things deserve mention. First we look at Table 2.2 and enumerate the striking observations.

1. The difference between the maximum likelihood estimate (minimum LD estimate) and the outlier deleted maximum likelihood estimate is substantial.

2. Other than the maximum likelihood estimate, the minimum PCS estimate, and, to lesser extent the minimum $\text{BWHD}_{1/3}$ estimate appear to be significantly influenced by the outlying values. In terms of robustness, the minimum PCS estimate is clearly the worst, by far, among all the minimum distance estimates presented in this example.

3. All the other estimates in our list successfully withstand the effect of the outliers. Each of these estimates provide an excellent fit to the first three cells of the observed data while effectively ignoring the large values.

4. The set of estimates which effectively discount the outlying observations include the minimum $\text{BWCS}_{0.2}$ estimate. Interestingly, the latter estimator has a positive estimation curvature ($A_2 = 0.4$) and by itself that would have predicted a nonrobust outcome in this case. That it does not happen demonstrates that while the estimation curvature is a useful local measure, it does not necessarily capture or represent the full global characteristics of a particular distance. In Chapter 4 we will see that the graphical interpretation based on combined residuals can give further insight on the robustness of some of the estimators not captured by the estimation curvature.

5. The minimum $\text{PD}_{-0.9}$ estimate appears to be the most conservative

in our list of robust estimates, and seems to downweight not just the outliers, but some of the more legitimate values as well. While there is no robustness issue here, this is indicative of another problem – that involving inliers – which we will discuss in Chapter 6.

The observations in Table 2.3 are generally similar. However, in this case, the large observation is a wildly discrepant value and not just a mild outlier. All our minimum distance estimators apart from the minimum LD estimator and the minimum PCS estimator are entirely successful in effectively ignoring this observation. These include both the $\text{BWHD}_{1/3}$ and the $\text{BWCS}_{0.2}$ estimators. This demonstrates that while sometimes a robust minimum distance method may provide a tentative treatment for a marginal outlier, an extreme outlier is solidly dealt with. ‖

Example 2.3. The data set for this example involves the incidence of peritonitis for 390 kidney patients. The data, presented in Table 2.4, were provided by Professor Peter W. M. John (personal communication) of the Department of Mathematics, University of Texas at Austin, USA. The observed frequencies resulting from the number of cases of peritonitis are reported at the top of the table. A visual inspection suggests that a geometric distribution with parameter θ (success probability) around 0.5 may fit the data well. We fit a geometric model with parameter θ to this data using several of our minimum distance methods. In this case, the estimates do not show any dramatic outlier effect. The two observations at 10 and 12 are mild to moderate outliers. But unlike Table 2.2, the sample size in this case is substantially higher, so that the relative impact of these moderate outliers are expected to be less severe. Indeed a comparison of the $\hat{\theta}$ column (the estimates for the full data) with the $\hat{\theta}_D$ column (the estimates for the cleaned data after removing the two outliers) shows that the full data estimates and the outlier deleted estimates are fairly close for practically all the methods, and even for the LD and the $\text{BWHD}_{1/3}$ case the impacts are relatively minor. However, the outliers do appear to have a fair influence on the minimum PCS estimate, again outlining its robustness problems. The minimum $\text{PD}_{-0.9}$ estimate, on the other hand, is still highly conservative and appears to drag the estimate the other way, underscoring the inlier problem once again. ‖

TABLE 2.4
Observed frequencies of the number of cases of peritonitis for each of 390 kidney patients, together with the expected frequencies under different estimation methods for the geometric model.

	Number of cases														
	0	1	2	3	4	5	6	7	8	9	10	11	≥ 12	$\hat{\theta}$	$\hat{\theta}_D$
Observed	199	94	46	23	17	4	4	1	0	0	1	0	1		
LD	193.5	97.5	49.1	24.7	12.5	6.3	3.2	1.6	0.8	0.4	0.2	0.1	0.1	0.496	0.509
$PD_{-0.9}$	212.4	96.7	44.1	20.1	9.1	4.2	1.9	0.9	0.4	0.2	0.1	-	-	0.544	0.551
PCS	179.8	96.9	52.2	28.2	15.2	8.2	4.4	2.4	1.3	0.7	0.4	0.2	0.2	0.461	0.501
NED	196.4	97.5	48.4	24.0	11.9	5.9	2.9	1.5	0.7	0.4	0.2	0.1	0.1	0.504	0.507
HD	199.1	97.5	47.7	23.4	11.4	5.6	2.7	1.3	0.7	0.3	0.2	0.1	0.1	0.510	0.518
$BWHD_{1/3}$	194.3	97.5	48.9	24.6	12.3	6.2	3.1	1.6	0.8	0.4	0.2	0.1	0.1	0.498	0.510
$BWCS_{0.2}$	193.0	97.5	49.2	24.9	12.6	6.3	3.2	1.6	0.8	0.4	0.2	0.1	0.1	0.495	0.505
$GKL_{1/3}$	197.1	97.5	48.2	23.8	11.8	5.8	2.9	1.4	0.7	0.3	0.2	0.1	0.1	0.506	0.512
$RLD_{1/3}$	197.8	97.5	48.0	23.7	11.7	5.7	2.8	1.4	0.7	0.3	0.2	0.1	0.1	0.507	0.509
SCS	197.8	97.5	48.0	23.7	11.7	5.8	2.8	1.4	0.7	0.3	0.2	0.1	0.1	0.507	0.511

This page intentionally left blank