Journal of Computational and Graphical Statistics

Nonparametric Estimation of Distributions in Random Effects Models

Jeffrey D. Hart and Isabel Cañette

[a] Jeffrey D. Hart is Professor, Department of Statistics, Texas A&M
University, College Station, TX 77843 hart@stat.tamu.edu. Isabel
Cañette is Senior Statistician, StataCorp LP, College Station, TX
77845 isabel.canette@gmail.com. The work of Professor Hart
was supported by NSF grant DMS-0604801 and by Award no. KUS-
C1-016-04, made by King Abdullah University of Science and
Technology (KAUST).

Available online: 01 Jan 2012

PLEASE SCROLL DOWN FOR ARTICLE

Nonparametric Estimation of Distributions in Random Effects Models

Jeffrey D. HART and Isabel CAÑETTE

We propose using minimum distance to obtain nonparametric estimates of the distributions of components in random effects models. A main setting considered is equivalent to having a large number of small datasets whose locations, and perhaps scales, vary randomly, but which otherwise have a common distribution. Interest focuses on estimating the distribution that is common to all datasets, knowledge of which is crucial in multiple testing problems where a location/scale invariant test is applied to every small dataset. A detailed algorithm for computing minimum distance estimates is proposed, and the usefulness of our methodology is illustrated by a simulation study and an analysis of microarray data. Supplemental materials for the article, including R-code and a dataset, are available online.

Key Words: Characteristic function; Identifiability; Minimum distance estimation; Quantile function.

1. INTRODUCTION

A common problem in modern statistics is to have a large number, say p, of small datasets. In principle, the distributions of data in different datasets could differ in an arbitary manner. Often, however, it is reasonable to assume that there is some degree of commonality among these distributions. A possible model for such commonality is the following:

$$X_{ij} = \mu_i + \sigma_i \epsilon_{ij}, \qquad i = 1, \ldots, p, j = 1, \ldots, n, \qquad (1.1)$$

where X_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, n$, are real-valued observations, and the following assumptions are made:

A1. The pairs (μ_i, σ_i), $i = 1, \ldots, p$, are independent and identically distributed unknown parameters.

A2. The unobserved errors ϵ_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, n$, are independent and identically distributed, with cumulative distribution function (cdf) F. Each ϵ_{ij} has mean 0 and variance 1.

A3. The parameters (μ_i, σ_i), $i = 1, \ldots, p$, are independent of ϵ_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, n$.

Model (1.1) will be referred to as the *location-scale random effects*, or LSRE, model. Such a model has been used, for example, in microarray analyses, wherein the index i denotes different genes and X_{ij}, $j = 1, \ldots, n$, are observations, usually expression levels, made on the ith gene, $i = 1, \ldots, p$. In the LSRE model the distributions for different datasets differ only with respect to location and scale. The current article is mainly concerned with the following question: "How well can F in model (1.1) be *nonparametrically* estimated when n is quite small but p is large?" Although this article is more concerned with practical matters than theory, an asymptotic analysis appropriate for our setting would keep n bounded as p tends to ∞.

Our interest in the question just posed is motivated by the canonical multiple hypotheses testing problem in microarray analyses. In model (1.1), suppose that X_{ij} is the difference between observations obtained from gene i of subject j before and after the subject is given a treatment. Then it is of interest to test the hypotheses

$$H_{0i} : \mu_i = 0, i = 1, \ldots, p,$$

each of which corresponds to the hypothesis of no treatment effect for a given gene. If one applies p standard t-tests, a common practice, then it is crucial to have knowledge of F since each test is based on a small number (n) of observations. One might argue that a rank or permutation test could be used that does not require knowledge of F. However, another benefit of inferring F is that this knowledge could be used to construct a test that is more powerful than a rank or permutation test.

A special case of the LSRE model is when the distribution of σ_i is degenerate at some positive constant σ. We will refer to this as the *location random effects*, or LRE, model. There exists a modest literature on estimation of F and the distribution of μ_i, call it G, in the LRE model. Reiersøl (1950) proved the important result that, under quite general conditions, both F and G are identifiable in the LRE model *when n is as small as* 2. Together with results of Wolfowitz (1957), this implies that both F and G can be estimated consistently using nonparametric methods when $p \to \infty$ and n is fixed at a value that is at least 2.

A large literature exists on the deconvolution problem corresponding to the LRE model with $n = 1$ and a *known* error distribution F. Much of this literature is referenced by Carroll and Hall (2004). Much less work has been done on the LRE model with both F and G unknown. A long gap in this work transpired after the early articles of Reiersøl (1950) and Wolfowitz (1957). To our knowledge the gap was not broken until the article of Horowitz and Markatou (1996), who considered a model for panel data that includes the location model as a special case. Horowitz and Markatou (1996) proposed nonparametric estimators of f and g (the densities of F and G in the location model) that are consistent and

attain optimal rates of convergence. These estimators are similar in construction to ones that are popular in the deconvolution model, that is, they plug estimates of a characteristic function into a Fourier inversion formula. We will call such estimators "deconvolution estimators" to distinguish them from the minimum distance estimators to be considered in the current article. The method of Horowitz and Markatou (1996) targets error densities f that are symmetric about 0. Li and Vuong (1998) also investigated estimators of deconvolution type but were able to avoid the assumption of a symmetric error distribution. Hall and Yao (2003) further weakened consistency conditions and also proposed minimum distance histogram estimators of f and g. Neumann (2007) proposed minimum distance type estimators of F and G in the LRE model and showed them to be strongly consistent, doing so under weaker conditions on F and G than in the aforementioned articles. In the LRE model with F unknown, Delaigle, Hall, and Meister (2008) identified conditions under which deconvolution estimators of the μ_i-density achieve the same rate of convergence as in the case of a known error density. Work related to ours but in the context of nonlinear modeling has been done by Schennach (2004).

The contributions of the current article may be summarized as follows.

C1. We propose new algorithms for approximating minimum distance estimators in LRE and LSRE models, and show their effectiveness via simulation studies and a real-data analysis.

C2. We prove that in the LRE model, the error distribution F is generally identifiable from either the joint distribution of $(X_{i1} - X_{i2}, X_{i1} - X_{i3})$ or the distribution of $X_{i1} - (X_{i2} + X_{i3})/2$. In particular, this result does not require the commonly used assumption that the error distribution is symmetric, as in the article by Delaigle, Hall, and Meister (2008), for example.

C3. The result described in C2 leads to *location-free* methods of estimating F, that is, ones completely free of $\mu_1, \ldots, \mu_p$. Our simulation shows that these location-free methods can yield more efficient estimators of F than a method that simultaneously estimates F and G.

C4. We prove that F is generally identifiable in the LSRE model for n as small as 4. The proof of this result leads to a method of estimating F in the LSRE model.

C5. We formulate a distribution-free rank test of the null hypothesis that the LRE model holds against the alternative of an LSRE model. This test requires only that $n \geq 4$.

The rest of the article proceeds as follows. In the next section we describe the minimum distance method in the context of the LRE model. An algorithm for approximating such estimates is described in Section 3, and simulation results are reported in Section 4. Our ideas are applied to the LSRE model in Section 5 and an example involving microarrays is provided in Section 6. Finally, concluding remarks are made in Section 7.

2. METHODOLOGY FOR THE LOCATION RANDOM EFFECTS MODEL

As indicated previously, our main interest is in estimating F, the error distribution. In this section we consider the LRE model, which is the model investigated by Li and Vuong (1998), Hall and Yao (2003), and Neumann (2007). In Section 2.1 we discuss two new ways of identifying F in the LRE model. These results are valid for n as small as 3 and are completely independent of $\mu_1, \ldots, \mu_p$. In Section 2.2 we describe our minimum distance methodology for estimating F in an LRE model, and in Section 2.3 we discuss consistency of our estimators.

2.1   LOCATION-FREE IDENTIFIABILITY OF F

The seminal result of Reiersøl (1950) shows that in the LRE model, both F and G are identifiable from the joint distribution of (X_{i1}, X_{i2}). This result forms the basis for methods of estimating F and G since the joint distribution is readily estimated by the empirical distribution of (X_{i1}, X_{i2}), $i = 1, \ldots, p$. When G is a nuisance and F the distribution of interest, it is reasonable to ask if there are ways to estimate F that are easier and/or more efficient than existing methods that require estimation of G. A basic question in this regard is the following: "Can F be identified without assuming anything whatsoever about G?" It is well known (see, e.g., Horowitz and Markatou 1996) that if F is symmetric and its characteristic function never vanishes, then it is identifiable from the distribution, call it D_F, of $\epsilon_{i1} - \epsilon_{i2}$. In this case F is estimable from data differences $X_{i1} - X_{i2} = \sigma(\epsilon_{i1} - \epsilon_{i2})$, $i = 1, \ldots, p$. However, if the symmetry assumption is dropped, then F is not identifiable from D_F, as there exist cases where F_1 is different from F_2, but $D_{F_1} \equiv D_{F_2}$.

Suppose now that $n \geq 3$ and define

$$\delta_{ijl} = X_{ij} - X_{il} = \sigma(\epsilon_{ij} - \epsilon_{il}), \qquad i = 1, \ldots, p, j, l = 1, \ldots, n.$$

A simple but important observation is that the two differences δ_{ijl} and δ_{ijm} with j, l, m distinct comprise a special case of the LRE model in which $G(x) = 1 - F(-x)$ for all x. The common random variable ϵ_{ij} in these two differences plays the role of μ_j, while $-\epsilon_{il}$ and $-\epsilon_{im}$ are the error terms. Since $\epsilon_{ij}, \epsilon_{il}, \epsilon_{im}$ are mutually independent, it follows from the work of Reiersøl (1950) that the distribution of ϵ_{ij} is identifiable from the joint distribution of $(\delta_{ijl}, \delta_{ijm})$ as long as the characteristic function (cf) of ϵ_{ij} does not vanish throughout an interval. This identifiability condition is obviously much weaker than the assumption of a real cf that never vanishes. Furthermore, the only cost in weakening the identifiability condition is *one* extra observation in each small dataset.

In the next section we will describe a minimum distance method for estimating F. This method requires a consistent estimator of the joint cf, $\xi(s, t)$, of $(\delta_{ijl}, \delta_{ijm})$. Sufficient for this purpose is the $\sqrt{p}$-consistent estimator

$$\hat{\xi}(s, t) = \frac{2}{n(n-1)(n-2)p} \sum_{j=1}^{p} \sum_{k=1}^{n} \sum_{(l,m) \in \mathcal{S}_{nk}} \exp(is\delta_{jkl} + it\delta_{jkm}),$$

where $\mathcal{S}_{nk} = \{(l, m) : 1 \leq l < m \leq n, l \neq k, m \neq k\}$.

A second method is based on the residuals $\hat{\epsilon}_{ij} = X_{ij} - \bar{X}_{ij}$, $j = 1, \ldots, n$, $i = 1, \ldots, p$, where $\bar{X}_{ij} = \sum_{k \neq j} X_{ik}/(n - 1)$, $j = 1, \ldots, n$, $i = 1, \ldots, p$. Obviously, $\hat{\epsilon}_{ij} = \sigma(\epsilon_{ij} - \bar{\epsilon}_{ij})$, $j = 1, \ldots, n, i = 1, \ldots, p$, with $\bar{\epsilon}_{ij} = \sum_{k \neq j} \epsilon_{ik}/n$, $j = 1, \ldots, n$, $i = 1, \ldots, p$. The following theorem in regard to these residuals is proven in our supplementary materials.

Theorem 1.   *Let the LRE model hold with $n \geq 3$, and suppose that the cf of F does not vanish throughout an interval. Then the distribution of ϵ_{i1} is identifiable from that of $X_{i1} - \bar{X}_{i1}$.*

The nonparametric estimator $\hat{\eta}(t) = \sum_{j=1}^{p} \sum_{k=1}^{n} \exp(it\hat{\epsilon}_{jk})/(np)$ is consistent for the cf η of $\hat{\epsilon}_{ij}$. This estimator is the foundation of a minimum distance method for estimating F. A computational advantage of this method over the first one described in this section is that it involves *univariate*, rather than bivariate, distributions.

2.2   MINIMUM DISTANCE ESTIMATION

For ease of notation the ensuing discussion assumes that σ is known and equal to 1. However, in practice an unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{2pn(n-1)} \sum_{j=1}^{p} \sum_{k=1}^{n} \sum_{l \neq k} (X_{jk} - X_{jl})^2.$$

This estimator is $\sqrt{p}$-consistent so long as ϵ_{ij} has just more than two moments finite. The method below can be modified in an obvious way to incorporate $\hat{\sigma}$.

Initially we describe a method based on the pairs of differences $(\delta_{ijl}, \delta_{ijm})$, as defined in Section 2.1. The joint distribution of $(\delta_{ijl}, \delta_{ijm})$ (with j, l, m distinct) is

$$H(x, y) = \int_{-\infty}^{\infty} [1 - F(z - x)][1 - F(z - y)] \, dF(z), \qquad (2.1)$$

and the joint cf

$$\xi(s, t) = \psi_F(s + t)\psi_F(-s)\psi_F(-t),$$

where ψ_F is the cf of F. Considering that p is assumed to be large in our problem, the cf ξ can be well estimated by the empirical cf $\hat{\xi}$ defined in Section 2.1.

Now, the basic idea of the minimum distance method is straightforward. Try to find a distribution $\hat{F}$ with cf $\psi_{\hat{F}}$ such that

$$\xi_{\hat{F}}(s, t) \equiv \psi_{\hat{F}}(s + t)\psi_{\hat{F}}(-s)\psi_{\hat{F}}(-t)$$

is a good match to $\hat{\xi}(s, t)$. More formally, we may define a metric measuring the discrepancy between $\xi_{\hat{F}}$ and $\hat{\xi}$ and then choose $\hat{F}$ to minimize this discrepancy.

In principle any number of metrics could be used when employing the minimum distance method. However, we have found that *density*-based metrics work much better than ones based on the cdf. An example of the latter type is

$$D_1^2(H_1, H_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [H_1(x, y) - H_2(x, y)]^2 \, dH_2(x, y).$$

Density-based metrics measure the difference between densities rather than cdf's. In the frequency domain, the metric we propose is as follows:

$$D^2(\xi_{\hat{F}}, \hat{\xi}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp[-2b^2(s^2 + t^2)] |\xi_{\hat{F}}(s, t) - \hat{\xi}(s, t)|^2 \, ds \, dt,$$

where b is a small positive number. Using Parseval's formula one can see that this is, indeed, a density-based metric. We have

$$D^2(\xi_{\hat{F}}, \hat{\xi}) = 4\pi^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\hat{h}_{\text{model}}(x, y; b) - \hat{h}_b(x, y))^2 \, dx \, dy, \tag{2.2}$$

where $\hat{h}_{\text{model}}(\cdot; b)$ is the joint density with cf $\exp[-b^2(s^2 + t^2)/2]\xi_{\hat{F}}(s, t)$ and $\hat{h}_b(x, y)$ is a standard kernel density estimate defined by

$$\hat{h}_b(x, y) = \frac{2}{pn(n-1)(n-2)b^2} \sum_{j=1}^{p} \sum_{k=1}^{n} \sum_{(l,m) \in \mathcal{S}_{nk}} \phi\left(\frac{x - \delta_{jkl}}{b}\right) \phi\left(\frac{y - \delta_{jkm}}{b}\right),$$

in which ϕ is the standard normal density. So, the quantity b in (2.2) is actually the bandwidth of a kernel estimate of $h(x, y) = \int_{-\infty}^{\infty} f(z - x) f(z - y) f(z) \, dz$.

Let $\hat{F}$ denote the cdf that gives equal probability to each of the N numbers $\hat{Q}_1 < \hat{Q}_2 < \cdots < \hat{Q}_N$. The cf of $\hat{F}$ is

$$\psi_{\hat{F}}(t) = \frac{1}{N} \sum_{j=1}^{N} \exp(it\hat{Q}_j). \tag{2.3}$$

Using Fourier inversion the estimate $\hat{h}_{\text{model}}(\cdot; b)$ of $h(x, y)$ corresponding to (2.3) is

$$\hat{h}_{\text{model}}(x, y; b) = \frac{1}{b^2 N^3} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \phi\left(\frac{x - \hat{Q}_j + \hat{Q}_k}{b}\right) \phi\left(\frac{y - \hat{Q}_j + \hat{Q}_l}{b}\right).$$

In Section 3 we will describe a random search algorithm that seeks to find (for given N) the quantiles $\hat{Q}_j$, $j = 1, \ldots, N$, that minimize the distance (2.2).

A minimum distance estimate of F based on the residuals $\hat{\epsilon}_{ij}$, $j = 1, \ldots, n$, $i = 1, \ldots, p$, may be defined in an analogous manner. For $n = 3$, the cf of $\hat{\epsilon}_{ij}$ is $\eta_F(t) = \psi_F(t)\psi_F(-t/2)^2$, and we would thus seek $\hat{F}$ to minimize

$$\tilde{D}^2(\eta_{\hat{F}}, \hat{\eta}) = \int_{-\infty}^{\infty} \exp(-2b^2 t^2) |\eta_{\hat{F}}(t) - \hat{\eta}(t)|^2 \, dt$$

$$= 2\pi \int_{-\infty}^{\infty} (\tilde{h}_{\text{model}}(x; b) - \tilde{h}_b(x))^2 \, dx,$$

where $\tilde{h}_b$ is a Gaussian-kernel density estimate based on the np residuals. Representing $\hat{F}$ in terms of quantiles as before, the density $\tilde{h}_{\text{model}}(\cdot; b)$ is

$$\tilde{h}_{\text{model}}(x; b) = \frac{1}{bN^3} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} \phi\left(\frac{x - \hat{Q}_j + (\hat{Q}_k + \hat{Q}_l)/2}{b}\right).$$

2.3 CONSISTENCY OF MINIMUM DISTANCE ESTIMATORS

Neumann (2007) proved that certain minimum distance estimators in the LRE model are consistent under general conditions on F and G. The only way in which Neumann's distance criterion differs from ours is that he multiplied the absolute squared difference between characteristic functions by a *fixed* kernel function, that is, one that does not change with p. Our kernel function, $\exp[-2b^2(s^2 + t^2)]$, depends on the bandwidth b, which will tend to 0 with p if it is chosen by cross-validation. Extending Neumann's result to allow a p-dependent kernel function seems to us a minor technical problem, which in any event is beyond the scope of our article.

The only other issue involved in proving consistency of our estimators involves the fact that our algorithm only *approximates* the actual quantile function that minimizes the distance criterion. However, this is a problem that *any* existing method faces, since, to our knowledge, there is no known closed form for a minimum distance estimator in the context of LRE or LSRE models. Certainly consistency requires that the number N of quantiles increase without bound as p tends to infinity, but, importantly, as shown by Beran and Millar (1994), making the "right" choice of the number of quantiles is not crucial to the good performance of the estimators. It is perfectly acceptable to err on the side of a very large number of quantiles.

3. AN ALGORITHM TO APPROXIMATE MINIMUM DISTANCE ESTIMATES

Our algorithm is similar to that proposed in the context of random coefficient regression by Beran and Millar (1994) in that both are based on searching over quantiles. However, our algorithm is somewhat more detailed, and consists of two types of iterations, that we call *global* and *local*. A great deal of experimentation was done to arrive at the particular algorithm described below, although we do not claim that it is optimal in any sense.

Let $\widehat{\mathbf{Q}}$ denote an N-vector of quantile estimates corresponding to cdf $\hat{F}$. The goodness of this estimate is assessed by computing $D(\xi_{\hat{F}}, \hat{\xi})$ (or $\tilde{D}(\eta_{\hat{F}}, \hat{\eta})$), as defined in Section 2.2. Our algorithm randomly jitters elements of $\widehat{\mathbf{Q}}$ and sorts the jittered elements to produce new quantile estimates $\widehat{\mathbf{Q}}^{\text{new}}$. The quantity $D(\xi_{\hat{F}_{\text{new}}}, \hat{\xi})$ is then computed, where $\hat{F}_{\text{new}}$ is the cdf corresponding to $\widehat{\mathbf{Q}}^{\text{new}}$. If $D(\xi_{\hat{F}_{\text{new}}}, \hat{\xi}) < D(\xi_{\hat{F}}, \hat{\xi})$, then $\hat{F}_{\text{new}}$ is accepted as the currently best estimate of F. Otherwise $\tilde{F}$ remains the currently best estimate.

Global iterations are ones in which *all* elements of $\widehat{\mathbf{Q}}$ are randomly jittered before re-computing the metric D. A local iteration consists of jittering only one element of $\widehat{\mathbf{Q}}$ and then computing D. The algorithm with which we have had the most success starts from an initial estimate (to be discussed below) and does a series of global iterations on $\widehat{\mathbf{Q}}$. When the metric fails to change by more than, say, 1% on a prespecified number k_g of consecutive iterations, then global iterations cease and local ones begin.

One cycle of local iterations consists of N steps. Let $Z_{i1} < Z_{i2} < \cdots < Z_{iN}$ be the quantile estimates just prior to step i, $i = 1, \ldots, N$. Then step i consists of randomly jittering the quantile Z_{ii} and determining if the corresponding change in the quantile function

has made the metric smaller. When k_l cycles through all N quantiles fails to change the metric by a nonnegligible amount, local iterations end.

Our experience is that global iterations are good for getting the quantile estimates quickly headed in the right direction, but are not by themselves sufficient. At some point changing the whole set of quantiles is unlikely to decrease the metric, even if some individual quantiles still need to be moved. At this point switching to local iterations can lead to further decreases in the metric and corresponding improvements in the quantile estimates.

What can be used for an initial estimate of F? We use the uniform distribution having mean 0 and standard deviation 1. Specifically, we take $\widehat{\mathbf{Q}}^0$ to consist of $\sqrt{3}(2i - N - 1)/N$, $i = 1, \ldots, N$.

The number N of quantiles must also be specified. In a related context, Beran and Millar (1994) noted that the value of N should be taken as large as is computationally feasible. The important point here is that N is *not* a smoothing parameter that must be chosen carefully in order for the resulting estimator to be efficient. Obviously, in order to obtain a consistent estimator of F, N should increase without bound as $p \to \infty$, but usually the underlying distribution can be well represented by no more than, say, fifty quantiles. In practice we have had success taking N to be between 30 and 100, even when p is 1000 or more. Again, however, the only reason not to take N *very* large is that it slows our algorithm down considerably.

Let $Z_1 < \cdots < Z_N$ be the currently best set of quantiles. In one global iteration, the jittered quantiles are $Z_i + \eta_i$, $i = 1, \ldots, N$, where $\eta_1, \ldots, \eta_N$ are independent with $\eta_i \sim N(0, s_i^2)$, $i = 1, \ldots, N$. The standard deviations are

$$s_1 = Z_2 - Z_1, \qquad s_i = \frac{Z_{i+1} - Z_{i-1}}{2}, \qquad i = 2, \ldots, N-1, \qquad \text{and}$$

$$s_N = Z_N - Z_{N-1}.$$

A jittered quantile in a local iteration is defined similarly, with the noise variables at different iterations independent of each other.

One needs also to specify a value for the bandwidth b. We propose that this be done by applying least squares cross-validation to the density estimate $\hat{h}_b$ (or $\tilde{h}_b$). Since b has only a second-order effect on the metric D (or $\tilde{D}$), this seems to be more than adequate.

4. LOCATION-SCALE RANDOM EFFECTS MODELS

We now turn attention to LSRE models, which allow for the possibility that both the mean and standard deviation vary from one dataset to the next. In Section 4.1 we give conditions under which the error distribution F is identifiable, and in Section 4.2 describe an algorithm for estimating F. Section 4.3 proposes a rank-based test of the null hypothesis that the data follow an LRE model against the alternative that the LSRE model holds.

4.1 IDENTIFIABILITY OF ERROR DISTRIBUTION

Here we argue that F, the distribution of ϵ_{ij}, is generally identifiable in the LSRE model when $n \geq 4$.

Theorem 2. *Let the LSRE model hold with $n \geq 4$, and suppose that $\log \sigma_i$, ϵ_{i1}, and $\log |\epsilon_{i1} - \epsilon_{i2}|$ are absolutely continuous random variables with characteristic functions having countably many zeroes. Then the distribution of ϵ_{i1} is identifiable from that of $(X_{i1}, X_{i2}, X_{i3}, X_{i4})$.*

We now sketch the proof of this theorem since it is instructive as to our method of estimating F in the LSRE model.

- Let $C = E \log |\epsilon_{i1} - \epsilon_{i2}|$, which is assumed to exist finite. Then the density of $\log \sigma_i + C$ is identifiable from the joint distribution of $(\log |X_{i1} - X_{i2}|, \log |X_{i3} - X_{i4}|)$, this following from the result of Reiersøl (1950). By a simple change of variable it follows that the density of $\exp(C)\sigma_i$ is known.

- The density of $\exp(C)\sigma_i$ can be rescaled to have second moment 1, and hence the density g of $\log \sigma_i$ is determined because $E(X_{i1} - X_{i2})^2 = 2E\sigma_i^2$.

- Let a be the density of $X_{i1} - (X_{i2} + X_{i3} + X_{i4})/3$ and note that for any x

$$e^x a(\pm e^x) = \int_{-\infty}^{\infty} \tilde{f}(\pm e^{x-s}) e^{x-s} g(s) \, ds,$$

where $\tilde{f}$ is the density of $\epsilon_{i1} - (\epsilon_{i2} + \epsilon_{i3} + \epsilon_{i4})/3$. Since a and g are identified, it follows via classic deconvolution that $\tilde{f}$ is as well.

- Finally, Theorem 1 implies that f, the density of ϵ_{ij}, is identifiable from $\tilde{f}$.

In the next section we describe a two-stage estimation procedure that parallels this identifiability argument.

4.2 ALGORITHM FOR ESTIMATING F

Our method requires that $n \geq 4$, and for ease of notation we assume that $n = 4$. In the first stage of our estimation scheme we estimate the characteristic function ψ_g of $\log \sigma_i$. Note that

$$\log |X_{ij} - X_{ik}| = \log \sigma_i + \log |\epsilon_{ij} - \epsilon_{ik}|,$$

and hence $(\log |X_{i1} - X_{i2}|, \log |X_{i3} - X_{i4}|)$, $i = 1, \ldots, p$, comprise an LRE model. Therefore, we may use these data and minimum distance methods to obtain quantile estimates $\hat{q}_1, \ldots, \hat{q}_M$ for the distribution of $\log \sigma_i + C$. Estimates of quantiles for the $\log \sigma_i$ distribution are then given by

$$\hat{Q}_i = \hat{q}_i + \frac{1}{2}(\log \hat{m}_2 - \log S^2), \qquad i = 1, \ldots, M, \tag{4.1}$$

where

$$\hat{m}_2 = \frac{1}{24p} \sum_{i=1}^{p} \sum_{j=1}^{4} \sum_{k=1}^{4} (X_{ij} - X_{ik})^2 \qquad \text{and} \qquad S^2 = \frac{1}{M} \sum_{i=1}^{M} \exp(2\hat{q}_i).$$

Per the proof from the last section, the following two Fourier transforms need to be estimated at the second stage of our estimation scheme:

$$\psi_a^+(t) = \int_{-\infty}^{\infty} e^{itx} a(e^x) e^x \, dx = \int_0^{\infty} e^{it \log y} a(y) \, dy$$

and

$$\psi_a^-(t) = \int_{-\infty}^{\infty} e^{itx} a(-e^x) e^x \, dx = \int_{-\infty}^0 e^{it \log |y|} a(y) \, dy.$$

Letting $e_{jk} = X_{jk} - \sum_{\ell \neq k} X_{j\ell}/3$, $j = 1, \ldots, p$, $k = 1, 2, 3, 4$, define

$$\hat{\psi}_a^+(t) = \frac{1}{4p} \sum_{j=1}^p \sum_{k=1}^4 \exp(it \log |e_{jk}|) I(e_{jk} > 0)$$

and

$$\hat{\psi}_a^-(t) = \frac{1}{4p} \sum_{j=1}^p \sum_{k=1}^4 \exp(it \log |e_{jk}|) I(e_{jk} < 0),$$

where I is an indicator function. Obviously these estimators are unbiased and consistent (as $p \to \infty$) for their respective transforms.

Since $e^x a(e^x)$ is of convolution form, we have $\psi_a^+(t) = \psi_{\tilde{f}}^+(t) \psi_g(t)$, where $\psi_{\tilde{f}}^+(t) = \int_{-\infty}^{\infty} e^{itx} \tilde{f}(e^x) e^x \, dx$, and similarly $\psi_a^-(t) = \psi_{\tilde{f}}^-(t) \psi_g(t)$. The cf ψ_g is estimated by $\hat{\psi}_g(t) = M^{-1} \sum_{j=1}^M \exp(it \hat{Q}_j)$, where $\hat{Q}_1, \ldots, \hat{Q}_M$ are defined in (4.1). Now, any choice of distribution for ϵ_{ij} determines $\tilde{f}$ and hence $\psi_{\tilde{f}}^+$ and $\psi_{\tilde{f}}^-$. At the second stage of our estimation scheme, we thus choose quantiles of ϵ_{ij} to minimize

$$\int_{-\infty}^{\infty} |\hat{\psi}_a^+(t) - \hat{\psi}_{\tilde{f}}^+(t) \hat{\psi}_g(t)|^2 e^{-2b_1^2 t^2} \, dt + \int_{-\infty}^{\infty} |\hat{\psi}_a^-(t) - \hat{\psi}_{\tilde{f}}^-(t) \hat{\psi}_g(t)|^2 e^{-2b_2^2 t^2} \, dt,$$

where b_1 and b_2 are small positive numbers (bandwidths), and $\hat{\psi}_{\tilde{f}}^+$ and $\hat{\psi}_{\tilde{f}}^-$ are the estimates of $\psi_{\tilde{f}}^+$ and $\psi_{\tilde{f}}^-$ corresponding to the chosen quantiles of ϵ_{ij}. One may use the iterative algorithm of Section 3 to arrrive at final estimates for quantiles of ϵ_{ij}'s distribution.

A practical problem with implementing the approach just described is that, unlike the situations addressed previously, there are no explicit expressions for $\psi_{\tilde{f}}^+$ and $\psi_{\tilde{f}}^-$ in terms of the cf of ϵ_{ij}. We address this problem by using simulation.

- Draw a sample of size $4N$ randomly and with replacement from the proposed (discrete) distribution for ϵ_{ij}, where N may be arbitrarily large. Call these values $\epsilon_1^*, \ldots, \epsilon_{4N}^*$.

- Define $r_i = \epsilon_{4i}^* - (\epsilon_{4i-1}^* + \epsilon_{4i-2}^* + \epsilon_{4i-3}^*)/3$, $i = 1, \ldots, N$.

- Define, for example, $\hat{\psi}_{\tilde{f}}^+$ by $\hat{\psi}_{\tilde{f}}^+(t) = N^{-1} \sum_{j=1}^N \exp(it \log(|r_j|)) I(r_j > 0)$.

4.3 A TEST OF HOMOSCEDASTICITY

Of interest is a test that can reveal whether or not there is significant variation in scale parameters. In other words, we desire a test of the null hypothesis that the LRE model holds against the alternative of an LSRE model. If there are at least four replications per small dataset, then a simple rank test can be used. Consider differences $\delta_{ijk} = X_{ij} - X_{ik} = \sigma_i(\epsilon_{ij} - \epsilon_{ik})$ for which $j < k$. Then if $\{j, k\}$ and $\{l, m\}$ are disjoint, δ_{ijk} and δ_{ilm} are independent under the LRE model. On the other hand, if the LSRE model holds and σ_i has a nondegenerate distribution, then the covariance between $|\delta_{ijk}|$ and $|\delta_{ilm}|$ is

$$\text{Cov}(|\delta_{ijk}|, |\delta_{ilm}|) = \text{Var}(\sigma_i^2)E^2|\epsilon_{11} - \epsilon_{12}| > 0.$$

We thus propose the following test:

- From each small dataset, randomly select two differences that are independent of each other under the LRE model. Denote these data $(\delta_{i1}^*, \delta_{i2}^*)$, $i = 1, \ldots, p$.

- Pool the $2p$ absolute differences, rank them from smallest to largest, and let R_{ij} be the rank of $|\delta_{ij}^*|$, $i = 1, \ldots, p$, $j = 1, 2$.

- The test statistic is

$$\hat{\rho} = \frac{2\sum_{i=1}^p (R_{i1} - \bar{R})(R_{i2} - \bar{R})}{\sum_{i=1}^{2p}(i - \bar{R})^2},$$

where $\bar{R}$ is the average of the ranks, or $(2p + 1)/2$.

The distribution of $\hat{\rho}$ is invariant to that of ϵ_{ij} under the null hypothesis of an LRE model, and hence can be arbitrarily well approximated by means of simulation. Ideally, one would use *all* differences from a small dataset in defining a test statistic, but in that case the statistic would not be distribution-free owing to dependence among differences having an error term in common. If n is 8 or more, then two or more independent pairs of differences can be formed, and the statistic can be modified accordingly to account for the extra information.

5. A SIMULATION STUDY

We restrict attention in this study to the LRE model. The main purpose of the simulation is to compare three methods of estimating the error distribution F: the two location-free methods (see Sections 2.1 and 2.2) and a method that simultaneously estimates G and F. The location-free methods based on paired differences and residuals will be referred to as PD and R, respectively, and the other method will be called S (for simultaneous). Our algorithm for method S is identical to the one described in Section 2.3 except that it cycles back and forth between jittering quantiles of the G and F distributions, this being true in the case of both global and local iterations.

We simulated data from the LRE model with $p = 1000$ and $n = 3$. All eight combinations of two choices for G and four choices for F were considered. The choices for G were degenerate at 0 and standard normal, and those for F were standard normal, shifted

Table 1.   Average values of mean absolute error from simulation. Each table value is an average of 200 replications.

G	Method	F			
		Normal	Exponential	t	Bimodal
Degenerate	Simultaneous	0.0728	0.2335	0.1091	0.0902
	Paired differences	0.0882	0.1010	0.1365	0.0861
	Residuals	0.0828	0.0969	0.1319	0.1239
Normal	Simultaneous	0.0832	0.2132	0.1978	0.1517
	Paired differences	0.0902	0.0991	0.1366	0.0880
	Residuals	0.0806	0.0992	0.1351	0.1236

exponential, rescaled t with three degrees of freedom, and the bimodal normal mixture $F(x) = 0.5[\Phi(\sqrt{5}x + 2) + \Phi(\sqrt{5}x - 2)]$, all four of which have mean 0 and variance 1.

Estimates of the quantile function $F^{-1}(u)$ were computed at $u = (j - 1/2)/50$, $j = 1, \ldots, 50$, for each dataset generated from the LRE model. Two hundred replications were performed at each combination of F and G. To save on computing time, a modified version of the estimation algorithm described in Section 3 was used. For each dataset, three sets of 200 global iterations each were performed and the set leading to the smallest distance was chosen. At that point sets of local iterations began and continued until the relative change in the distance from one set to the next changed by less than 0.001.

Results are summarized in Table 1 and Figure 1. As an overall measure of the quality of a quantile estimator $\hat{Q}$, we computed the following mean absolute error:

$$\text{MAE}(\hat{Q}) = \frac{1}{50} \sum_{i=1}^{50} \left| \hat{Q}\left(\frac{i - 1/2}{50}\right) - Q\left(\frac{i - 1/2}{50}\right) \right|,$$

where Q is the true quantile function. Average values of MAE are given in Table 1. First of all, differences between rows 2 and 5 and between 3 and 6 are due entirely to sampling variation, since methods PD and R are completely invariant to G. The average MAE (AMAE) for PD and R was comparable in all cases except for the bimodal error distribution, where the AMAE of R was about 50% more than that of PD. For all but one of the four error distributions, the performance of method S is much better when G is degenerate than when G is normal. This should not be surprising since the noise variable comprises *all* the variation in the data when there is no variation in μ_i. The location-free methods perform much better than S when $G \equiv$ normal in all cases except for F normal. If one were to choose a method on the basis of Table 1, it seems the paired differences approach would be best.

In Figure 1 we provide a visual comparison of methods S and PD when G is normal and F is either exponential or the bimodal normal mixture. We chose these two error distributions since they are the most non-Gaussian of the four choices for F, and hence would seem to be more challenging cases. Method S is substantially biased for both error distributions, whereas the pointwise median quantile estimate for method PD virtually coincides with the true quantile function in both cases. Especially impressive is that the PD method estimates
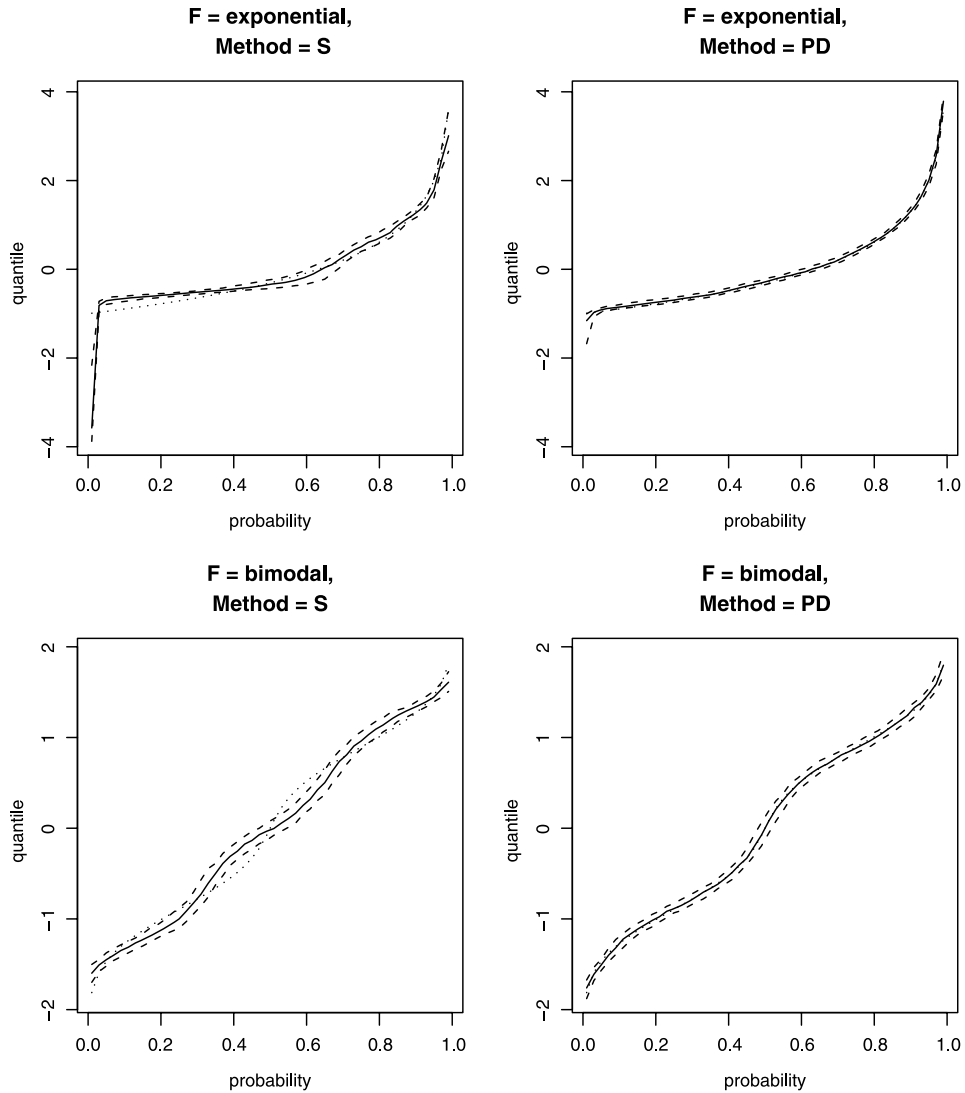
Figure 1. Summary of quantile estimates for the case where G is normal. The dotted line in each case is the true quantile function, and the solid line is the pointwise median of 200 simulated estimates. The dashed lines are pointwise 25th and 75th percentiles of estimates.

the lower endpoint of the exponential so well and effectively captures bimodality of the normal mixture. Obviously, method S can have difficulty estimating F effectively when there is nontrivial variation in the distribution of μ_i. Graphs corresponding to Figure 1 for the other three cases may be found in the supplementary materials.

6. AN ANALYSIS OF MICROARRAY DATA

Here we consider microarray data collected by Robert Chapkin and coworkers at Texas A&M University. An analysis of these data may be found in the article by Davidson et al.

(2004). The data we analyze are only part of a much larger dataset, but provide a good example of our methodology. The data considered are Y_{jk}, $j = 1, \ldots, 8038$, $k = 1, \ldots, 5$, where j indexes genes, k indexes different rats, and Y_{jk} is the logarithm of the expression level for gene j and rat k. The five rats from which these data were collected were all subjected to the same treatment.

We assume the following model for the data:

$$Y_{jk} = R_k + \mu_j + \sigma_j \epsilon_{jk}, \qquad j = 1, \ldots, 8038, k = 1, \ldots, 5,$$

where R_k represents a rat effect, (μ_j, σ_j) a gene effect, and ϵ_{jk} measurement error. Our main goal is to estimate the distribution of ϵ_{jk}. The first step in our analysis is to estimate rat effects by computing the mean of all data for each rat. Defining

$$X_{jk} = Y_{jk} - \frac{1}{8038} \sum_{i=1}^{8038} Y_{ik}, \qquad j = 1, \ldots, 8038, \ k = 1, \ldots, 5,$$

we may say that, to a good approximation, the X_{jk}'s follow either an LRE or LSRE model since each rat effect is estimated by the mean of over 8000 observations.

6.1 TESTING FOR AN LSRE MODEL

As a descriptive device we provide a scatterplot (Figure 2) of sample variances versus sample means for the X-data from all 8038 genes. Datasets with sample variances larger than 3 are not represented in the plot, but there were only 12 such sets, none of which had a sample mean larger than 1.81. There appears to be evidence that σ_j^2 *decreases* with an increase in μ_j. However, this is not necessarily a sound conclusion, as we now argue. In an LSRE model, let $\bar{X}_i$ and S_i^2 be the sample mean and variance, respectively, of $X_{i1}, \ldots, X_{in}$. It is then straightforward to show that

$$\text{Cov}(\bar{X}_i, S_i^2) = \text{Cov}(\mu_i, \sigma_i^2) + \frac{1}{n} E \sigma_i^3 E \epsilon_{ij}^3.$$
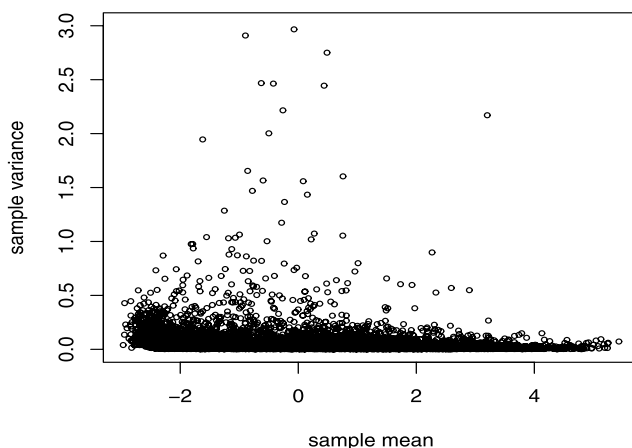

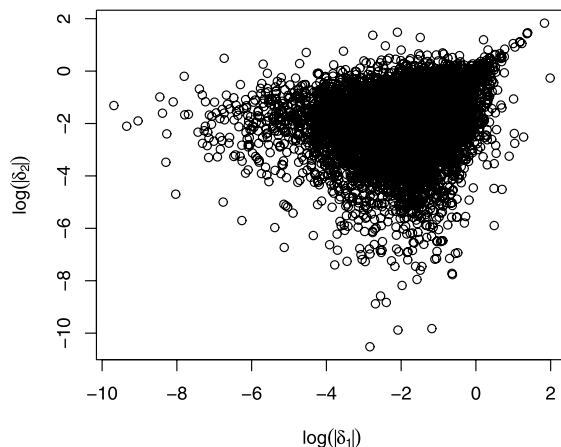Figure 2.   Scatterplot of sample variances versus sample means for 8038 genes.

Figure 3.   Scatterplot of logged absolute differences from the same gene.

So, negative correlation between $\bar{X}_i$ and S_i^2 is not necessarily an indication that μ_i and σ_i^2 are negatively correlated. This could simply be an indication of left skewness in the error distribution. Only when the third moment of ϵ_{ij} is known to be 0 can we conclude a relationship between μ_i and σ_i^2 from a similar relationship between $\bar{X}_i$ and S_i^2. We will return to this point after we have estimated the error distribution.

To formally address the question of whether an LSRE model is more appropriate than an LRE model, we apply the test of Section 4.3. In doing so, we randomly selected, separately for each gene, four rats, and thereby obtained 8038 pairs of differences. The resulting rank correlation $\hat{\rho}$ between differences from the same gene was 0.220. Now, let $j_1, \ldots, j_{2(8038)}$ be a random permutation of the integers from 1 to 2(8038). The null distribution of the test statistic is the same as that of $\hat{\rho}$ with $(R_{i1}, R_{i2}) = (j_{2i-1}, j_{2i})$, $i = 1, \ldots, 8038$. In 100,000 independent permutations, we found no correlation larger than 0.0489 in absolute value, leading to a p-value of less than 0.00001. There is thus strong evidence of differences in scale from one gene to the next. A scatterplot of log-absolute differences from the same gene is shown in Figure 3. Lack of independence between the two differences is evident.

6.2   ESTIMATION OF ERROR DISTRIBUTION IN LSRE MODEL

Since the LSRE model appears to be more tenable than the LRE, we applied the method described in Section 4.2 to estimate the error distribution. In doing so we also obtained an estimate of the marginal distribution of σ_j. One hundred quantiles for each of the error and σ_j distributions were estimated. The bandwidths required at each of the two stages of the estimation scheme were chosen by cross-validation. Two different initial estimates, uniform and normal, for the error distribution were used, and both led to very similar estimates at the end of iterations. The normal initial estimate yielded the smaller of the two final discrepancy measures. Plots of the estimated quantile functions for σ_j and ϵ_{ij} are shown in Figures 4 and 5, respectively. The error quantiles are remarkably close to uniform, which of course is a symmetric distribution. Recalling our comments in Section 6.1, it thus seems reasonable to conclude that the negative correlation seen in Figure 2 is not an
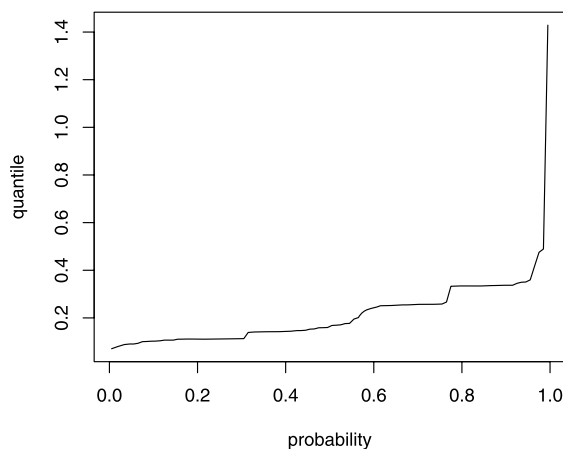
Figure 4.   Minimum distance estimate of quantile function of σ_i.

artifact of the error distribution, but rather a real indication of negative correlation between μ_i and σ_i^2.

It is also noteworthy that the uniform distribution is *short-tailed*. When applying location tests on a per-gene basis, as discussed in Section 1, knowing that the error distribution is short- rather than long-tailed could potentially lead to important differences in conclusions. Knowledge of the error distribution, as provided by our methodology, could lead to tests that are more powerful than, say, a t-test. For example, for the data just analyzed it would be reasonable to use a linear signed rank test with scores designed for short-tailed densities; see, for example, the work of Randles and Wolfe (1979, pp. 323–324).
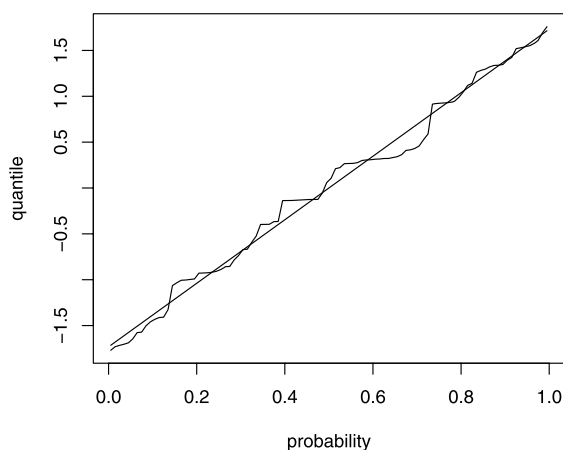

Figure 5.   Uniform quantile function and minimum distance estimate of quantile function of ϵ_{ij}. Both the quantile estimate and uniform distribution have mean 0 and variance 1.

7. DISCUSSION

A number of interesting questions have arisen during the course of our research. We end our article with a discussion of a few of these.

7.1 RELATIVE EFFICIENCY OF MINIMUM DISTANCE AND DECONVOLUTION

Of considerable interest is a comprehensive comparison of minimum distance estimators and those of deconvolution type, that is, estimators based on explicit inversion of cf's. One advantage of minimum distance is that it can be applied more generally than deconvolution. Deconvolution requires that the cf of the target distribution is expressible in terms of an observable cf, and this is not always the case. In cases where deconvolution *can* be applied, it is of interest to know how its efficiency compares with that of minimum distance. Our simulation results, and those of Hall and Yao (2003), suggest that minimum distance is more efficient than deconvolution, at least in the LRE model.

7.2 IDENTIFIABILITY VERSUS CURSE OF DIMENSIONALITY

As noted in Section 2.1, both distributions in the location model are identified from the joint distribution of (X_{ij}, X_{ik}) so long as the cf's of the two distributions never vanish throughout an interval. This assumption does not seem overly restrictive, but one wonders whether it can be weakened. Let $H_{n,F,G}$ be the joint distribution of $(X_{i1}, \ldots, X_{in})$ in the LRE model when ϵ_{ij} has cdf F and μ_i has cdf G. Now define the class $\mathcal{F}_n$ of pairs of distributions (F, G) as follows: $(F, G) \in \mathcal{F}_n$ if and only if there exists no other pair of distributions $(\bar{F}, \bar{G})$ such that $H_{n,F,G} \equiv H_{n,\bar{F},\bar{G}}$. Intuitively it seems plausible that these classes have the property $\mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots$. If this is true, then for $n > 2$ there may be an advantage to using an observable distribution of higher dimension than 2. From an efficiency standpoint, one must cope with the curse of dimensionality when applying minimum distance to a higher dimensional distribution, but if substantially more distributions become identifiable, then perhaps the trade-off is worthwhile.

7.3 CORRELATION ACROSS SMALL DATASETS

Our random effects models are such that data within the same small dataset are correlated, but different small datasets are independent. The latter assumption can be relaxed so long as the correlation among different datasets is not extremely strong. An appealing feature of minimum distance estimates is that they depend on the data only through an empirical distribution function (edf), and edf's are robust to dependence of mixing type (see, e.g., Gastwirth and Rubin 1975).

SUPPLEMENTARY MATERIALS

Appendix: A proof of Theorem 1 and plots illustrating simulation results are provided. (appendix.pdf)

R-code: R-code used to produce the simulation results and the data analysis in Section 6. (Rcode.jcgs, text)

Rat Data: Data analyzed in Section 6. (rat.data, text)

ACKNOWLEDGMENTS

REFERENCES

Beran, R., and Millar, P. W. (1994), "Minimum Distance Estimation in Random Coefficient Regression Models," *The Annals of Statistics*, 22, 1976–1992. [467,468]

Carroll, R. J., and Hall, P. (2004), "Low Order Approximations in Deconvolution and Regression With Errors in Variables," *Journal of the Royal Statistical Society, Ser. B*, 66, 31–46. [462]

Davidson, L. A., Nguyen, D. V., Hokanson, R. M., Callaway, E. S., Isett, R. B., Turner, N. D., Dougherty, E. R., Wang, N., Lupton, J. R., Carroll, R. J., and Chapkin, R. S. (2004), "Chemopreventive *n*-3 Polyunsaturated Fatty Acids Reprogram Genetic Signatures During Colon Cancer Initiation and Progression in the Rat," *Cancer Research*, 64, 6797–6804. [473,474]

Delaigle, A., Hall, P., and Meister, A. (2008), "On Deconvolution With Repeated Measurements," *The Annals of Statistics*, 36, 665–685. [463]

Gastwirth, J. L., and Rubin, H. (1975), "The Asymptotic Distribution Theory of the Empiric cdf for Mixing Stochastic Processes," *The Annals of Statistics*, 3, 809–824. [477]

Hall, P., and Yao, Q. (2003), "Inference in Components of Variance Models With Low Replication," *The Annals of Statistics*, 31, 414–441. [463,464,477]

Horowitz, J. L., and Markatou, M. (1996), "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145–168. [462-464]

Li, T., and Vuong, Q. (1998), "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139–165. [463,464]

Neumann, M. H. (2007), "Deconvolution From Panel Data With Unknown Error Distribution," *Journal of Multivariate Analysis*, 98, 1955–1968. [463,464,467]

Randles, R. H., and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Statistics*, New York: Wiley. [476]

Reiersøl, O. (1950), "Identifiability of a Linear Relation Between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389. [462,464,469]

Schennach, S. (2004), "Estimation of Nonlinear Models With Measurement Error," *Econometrica*, 72, 33–75. [463]

Wolfowitz, J. (1957), "The Minimum Distance Method," *Annals of Mathematical Statistics*, 28, 75–88. [462]