The Annals of Statistics
2003, Vol. 31, No. 2, 414–441

INFERENCE IN COMPONENTS OF VARIANCE MODELS WITH LOW REPLICATION

BY PETER HALL AND QIWEI YAO

London School of Economics and Australian National University, and London School of Economics

In components of variance models the data are viewed as arising through a sum of two random variables, representing between- and within-group variation, respectively. The former is generally interpreted as a group effect, and the latter as error. It is assumed that these variables are stochastically independent and that the distributions of the group effect and the error do not vary from one instance to another. If each group effect can be replicated a large number of times, then standard methods can be used to estimate the distributions of both the group effect and the error. This cannot be achieved without replication, however. How feasible is distribution estimation if it is not possible to replicate prolifically? Can the distributions of random effects and errors be estimated consistently from a small number of replications of each of a large number of noisy group effects, for example, in a nonparametric setting? Often extensive replication is practically infeasible, in particular, if inherently small numbers of individuals exhibit any given group effect. Yet it is quite unclear how to conduct inference in this case. We show that inference is possible, even if the number of replications is as small as 2. Two methods are proposed, both based on Fourier inversion. One, which is substantially more computer intensive than the other, exhibits better performance in numerical experiments.

1. Introduction. Problems involving components of variance arise in many areas of sampling and design, including the design and analysis of interlaboratory standardization trials and analysis of reliability of measurements such as blood pressure. The components of variance approach dates from Airy's (1861) work on measurement errors in astronomy and has been used in the contexts of randomized design, population genetics, variability of industrial processes, educational testing and many other fields; see, for example, Tippett (1931), Daniels (1939) and Cornfield and Tukey (1965). Eisenhart (1947) introduced the terms "fixed effects," "random effects" and "random effects of analysis of variance." It is with the latter problem, and the extent to which the distribution of random effects can be estimated or approximated with minimal replication, that we are concerned.

The essence of a simple components of variance model is that variability may be expressed as a sum of two independent random quantities, representing a group

Received May 2001; revised December 2001.

AMS 2000 subject classifications. Primary 62J10; secondary 62E17, 62G05.

Key words and phrases. Analysis of variance, characteristic function, components of variability, curve estimation, deconvolution, hierarchical models, nonparametric curve estimation, random effects, standardization trials.

effect (corresponding roughly to a treatment effect in more conventional settings) and an error, respectively. More generally, in a multilevel setting there can be a group effect for each stratum of variation. By separating out a deterministic location parameter we may assume that the random components have zero mean. Then, under normality assumptions, and in the simplest balanced situations, the distribution of each component is describable solely in terms of its variance. Much recent research under the heading of hierarchical models emphasizes unbalanced data and specific nonnormal distributions, however. In this paper a relatively simple situation is revisited from the very different viewpoint of a wholly nonparametric formulation.

It is not difficult to show that in this general setting the distributions of the group effects and the errors are consistently estimable, provided the number of groups, and the number of replications within each group, diverges without bound. However, the situation in the case of small, fixed numbers of replications is quite unclear.

Solving that problem motivates the present paper. More particularly, the approach we adopt is motivated by three goals: (1) to give general conditions under which, when the number of replications is fixed and as small as 2, the problem of consistently estimating the distributions of group effects and errors can be solved in a nonparametric context (and so is identifiable there); (2) to exhibit two particular estimator types that achieve consistent estimation, one of them not requiring separate choice of smoothing parameter; and (3) to provide a basis on which other techniques can be developed, for example, more descriptive methods based on moments.

We shall show that under very mild side conditions the distributions of group effects and errors are identifiable, provided only that each group contains at least two replications and the number of groups is allowed to diverge. The main regularity condition is that the characteristic function of neither distribution vanishes in an interval.

Once identifiability has been established, the way is open for a range of relatively ad hoc methods to be implemented. In particular, tth moments of the distributions of the group effects and errors may be estimated root-n consistently using relatively simple techniques, such as those based on homogeneous polynomials of degree t in the data. We shall outline our methodology in Section 2.1. The cases $t = 1$, 2 and 3 are straightforward, and $t = 4$ is quite practicable, although $t \geq 5$ presents significantly greater difficulty. Fitting, say, a distribution from the Pearson system [see, e.g., Johnson, Kotz and Balakrishnan (1994), pages 15–25] to the first three or four estimated moments will often provide very useful approximations to the distributions of group effects and errors. Moreover, moment-based estimators can be used as starting values for iterative solution of the likelihood equations, provided a finite-dimensional model is appropriate.

Thus the results in this paper pave the way for a variety of approaches for estimating distributions of group effects and errors. While methods based on low-order moments are arguably the most attractive, from a practical viewpoint, if one's only goal is to acquire an impression of the shape of the sampled distribution, they do not lend themselves to consistent distribution estimation. In addition to the difficulty of estimating moments of order 5 or more, and transiting from moment approximations to distribution approximations, the problem of determining the "smoothing parameter," or, equivalently, how many moments should be fitted, is very difficult to solve. One of our alternative approaches is particularly attractive in this regard, since it involves an empirically chosen smoothing parameter and leads to consistent distribution estimation.

Despite, or perhaps because of, their significant practical interest, components of variance models are not without an element of controversy, not least because it can be argued that general linear models may be developed to accommodate a particularly wide range of sources of variability. See, for example, the proposals of Nelder (1977) and Yates' (1966) interpretation of Eisenhart's (1947) suggestions. But note, too, the discussion of Nelder (1977), and the views of Kempthorne (1975). Variance components analysis has been discussed and surveyed by Plackett (1960), Khuri and Sahai (1985) and Sahai, Khuri and Kapadia (1985). A broad coverage of techniques for inference in variance components models has been provided by Searle, Casella and McCulloch (1992).

A problem that is related, more in the context of mathematical methods than direct statistical motivation, is that of estimating a linear relationship between variables that are observed with error. Early contributions in this setting include those of Reiersøl (1950), Neyman (1951) and Wolfowitz (1952); see the survey paper by Moran (1971). The problem can be treated either parametrically [e.g., Bickel and Ritov (1987)] or nonparametrically [e.g., Spiegelman (1979)]. Methods used for random coefficient regression are also related; see, for example, Beran, Feuerverger and Hall (1996).

2. Methodology.

2.1. *Structural models for components of variance.* A naive model is

$$X_j = \mu + \xi_j + \varepsilon_j, \qquad 1 \le j \le n, \tag{2.1}$$

where μ is a constant, $\mu + \xi_j$ denotes the jth group effect, ε_j represents the observation error associated with the jth group and the random variables ξ_j and ε_j are mutually independent with zero mean. The common distributions F and G of the ξ_j's and ε_j's, respectively, are clearly not identifiable from an infinite sequence of data from the model (2.1). Even if Gaussian models are assumed for F and G the parameters are not identifiable. We shall be primarily concerned with the nonparametric setting, where identification is still more complex.

Suppose, however, that each group is replicated r times:

$$X_{js} = \mu + \xi_j + \varepsilon_{js}, \qquad 1 \le j \le n,\ 1 \le s \le r, \tag{2.2}$$

where X_{js} denotes the sth replicate of the jth group, observed with additive error ε_{js}, and the variables ξ_j and ε_{js} are mutually independent with zero mean. Each ξ_j is assumed to have distribution F, and each ε_{js} to have distribution G. If we allow n and r to diverge together, then we can obviously identify F and G. One approach is via conventional empirical methods, for example, giving convergence rates equal to $\min(n, r)^{-1/2}$. This is true in a nonparametric sense; we do not require more than basic assumptions, such as moment conditions, on the distributions of ξ and ε, and the assumption $E(\xi) = E(\varepsilon) = 0$, which serves to identify the centers of F and G as well as the value of μ.

In contrast, it is unclear whether identification of F and G is even possible if r is small relative to n, in particular, if r is held fixed as $n \to \infty$. We shall introduce and describe the properties of two characteristic function-based, nonparametric methods for inference. Both methods are valid for r as small as 2. One is explicit, and is based on estimating the characteristic functions of F and G and explicitly inverting them. It has features in common with deconvolution. The other is approach implicit, and is founded on fitting histogram-type density and distribution estimators using a goodness of fit measure expressed in terms of characteristic functions. The former method is less computer intensive; the latter requires an algorithm such as simulated annealing, but has somewhat better performance. Using our techniques, and provided the number of groups is large, it is unnecessary to have conducted a large number of replications in order to estimate F and G.

Some insight into the types of regularity conditions needed can be gained by simply calculating the characteristic function of X in formulas such as (2.1) and (2.2). Assuming, without loss of generality, that $\mu = 0$, we find that the characteristic function of X equals the product of the characteristic functions of ξ and ε, and so the characteristic function of ξ (respectively, ε) is not always identifiable if the characteristic function of ε (respectively, ξ) vanishes on an interval. Therefore we should assume the latter does not occur. This argument remains valid if we have only a bounded number of replications, because we can never get close to the particular value of ξ.

Simple estimators of moments of the distributions of ξ and ε can be based on polynomials in the data, for example,

$$\sum_{s_1,\ldots,s_t} a_{s_1\cdots s_t} Y_{js_1} \cdots Y_{js_t},$$

where the coefficients $a_{s_1\cdots s_t}$ are constants, $Y_{js} = X_{js} - \bar{X}_{\cdot\cdot}$, $\bar{X}_{\cdot\cdot}$ denotes the grand mean of the data X_{js} generated by the model (2.2), the sum is over all distinct unordered t-tuples $s_1, \ldots, s_t$, and each s_j lies between 1 and r. The coefficients $a_{s_1\cdots s_t}$ can be chosen such that the estimator is order invariant, is root-n consistent for either $E(\xi^t)$ or $E(\varepsilon^t)$ and has bias equal to $O(n^{-1})$. However, it is difficult to

choose $a_{s_1 \ldots s_t}$ to have good variance properties unless $t \le 4$. Problems such as this preclude a theory of consistent distribution estimation based on moment fitting. Further details are given in Cox and Hall (2002).

2.2. *Estimating characteristic functions.* In this section we suggest estimators of the characteristic functions ϕ and ψ of ξ_j and ε_{js}, respectively, and of simple functionals of those characteristic functions. Put

$$\widehat{\chi}(t \mid a, b) = \frac{1}{nr(r-1)} \sum_{j=1}^{n} \sum\sum_{1 \le s_1, s_2 \le r \,:\, s_1 \ne s_2} \exp\{it(aY_{js_1} + bY_{js_2})\}, \tag{2.3}$$

where $i = \sqrt{-1}$ and a and b are real numbers. Then $\widehat{\chi}(t \mid a, b)$ estimates the characteristic function, $\chi(t \mid a, b)$ say, of $(a+b)\xi + a\varepsilon_1 + b\varepsilon_2$:

$$\chi(t \mid a, b) = \phi\{(a+b)t\}\psi(at)\psi(bt), \tag{2.4}$$

where ξ, ε_1 and ε_2 are mutually independent random variables, ξ being distributed as ξ_j and ε_j distributed as ε_{js} in the model at (2.2).

Observe that, in view of (2.4),

$$\psi(t) = \exp\left[\sum_{j=0}^{\infty} 2^j \{\log \chi(t/2^j \mid 1, 0) - \log \chi(t/2^j \mid \tfrac{1}{2}, \tfrac{1}{2})\}\right], \tag{2.5}$$

assuming neither ϕ nor ψ vanishes. The infinite series on the right-hand side of (2.5) converges provided

$$E|\xi| + E|\varepsilon| < \infty \quad \text{and} \quad E(\xi) = E(\varepsilon) = 0. \tag{2.6}$$

The latter condition serves to identify the centers of the distributions of ξ and ε as well as the value of μ. Note particularly that (2.5) motivates the estimators

$$\hat{\psi}(t) = \exp\left[\sum_{j=0}^{\infty} 2^j \{\log \widehat{\chi}(t/2^j \mid 1, 0) - \log \widehat{\chi}(t/2^j \mid \tfrac{1}{2}, \tfrac{1}{2})\}\right] \tag{2.7}$$

and $\hat{\phi}(t) = \widehat{\chi}(t \mid 1, 0)/\hat{\psi}(t)$ of $\psi(t)$ and $\phi(t)$, respectively. To remove ambiguity about the branch of the logarithm in (2.6) and (2.7), we stipulate that each should be interpreted as the corresponding infinite product.

Our next result shows that these estimators are well defined, in particular, that the infinite series converges.

PROPOSITION 2.1. *Assume $r \ge 2$ and the distributions F and G are continuous. Then for each $t \in (-\infty, \infty)$ the estimators $\hat{\phi}$ and $\hat{\psi}$ are well defined and finite with probability* 1.

Neither $\hat{\phi}$ nor $\hat{\psi}$ is the characteristic function of a proper probability distribution, and, in fact, both will generally fail, for some value of their argument, to satisfy the constraint that they do not exceed 1 in absolute value. To overcome the latter difficulty, we may replace $\hat{\phi}$ and $\hat{\psi}$ by their truncated forms $\hat{\phi}_{\rm tr}$ and $\hat{\psi}_{\rm tr}$, respectively, where

$$\hat{\kappa}_{\rm tr} = \min(1, |\hat{\kappa}|) \exp(i \arg \hat{\kappa}) \tag{2.8}$$

and $\hat{\kappa}$ denotes either $\hat{\phi}$ or $\hat{\psi}$.

Both $\hat{\phi}$ and $\hat{\psi}$ have analogues in cases where the number of replicates, $r = r(j)$, depends on j but nevertheless satisfies $r(j) \geq 2$ for each j, or where the number of values of $j \leq n$ for which $r(j) \geq 2$ diverges to ∞ as $n \to \infty$. For notational convenience we shall not treat such cases explicitly. Our methods do not allow ready inclusion of information from instances where $r(j) = 1$.

There are, however, alternative approaches to inference. It does not seem possible to address the issue of conventional statistical efficiency here, on account of the difficulty of obtaining a limit theory that provides more information than simply rates of convergence. Nevertheless, it is clearly possible to enhance the performance of our estimators, for example, by altering their moduli using a subsidiary method, but retaining our estimators of the args, or phases, of the characteristic functions. The moduli of ϕ and ψ can be estimated relatively precisely as the square roots of the absolute values of the empirical characteristic functions computed from pairwise differences.

2.3. *Explicit characteristic function inversion.* We may invert $\hat{\phi}$ and $\hat{\psi}$ in elementary fashion, obtaining estimators $\hat{f}$ and $\hat{g}$ of the densities f and g of the respective distributions F and G:

$$\begin{aligned} \hat{f}(x) &= (2\pi)^{-1} \Re \int_{|t| \leq t_n} e^{-itx} \hat{\phi}_{\rm tr}(t)\, dt, \\ \hat{g}(x) &= (2\pi)^{-1} \Re \int_{|t| \leq t_n} e^{-itx} \hat{\psi}_{\rm tr}(t)\, dt, \end{aligned} \tag{2.9}$$

where the operator $\Re$ denotes the real part, $t_n > 0$ is a smoothing parameter that regularizes the estimators, and $\hat{\phi}_{\rm tr}$ and $\hat{\psi}_{\rm tr}$ are defined in terms of $\hat{\phi}$ and $\hat{\psi}$ by (2.8).

In practice, the rather sharp truncation of the integrals at points $\pm t_n$, suggested by (2.9), tends to introduce spurious oscillations of Gibbs phenomenon type. A tapering operation can produce more satisfactory results; see Section 3 for a discussion. Provided the characteristic functions ϕ and ψ do not vanish on intervals, consistent estimators of f and g are obtained by allowing t_n to diverge to ∞ sufficiently slowly as n increases; see Section 4.

Distributions are, of course, estimable by integrating the appropriate density estimator. For future reference we give the formula here: if $-\infty < x_1 < x_2 < \infty$, then

$$\hat{F}(x_1, x_2) = (2\pi)^{-1} \Re \int_{|t| \le t_n} \frac{e^{-itx_2} - e^{-itx_1}}{-it} \hat{\phi}_{\mathrm{tr}}(t)\, dt \tag{2.10}$$

estimates the probability $F(x_1, x_2)$ that $\xi \in (x_1, x_2)$, and the estimator $\widehat{G}$ of G is defined analogously. However, it is to be expected that accurate estimation of F, for example, would require a larger value of t_n than would be appropriate for estimating g, and we shall show in Section 4 that this is, in fact, the case.

Neither $\hat{f}(x)$ nor $\hat{g}(x)$ will be positive for all x, and neither $\hat{F}(x_1, x_2)$ nor $\widehat{G}(x_1, x_2)$ will be monotone in either x_1 or x_2. These deficiencies may be overcome by taking the positive parts of $\hat{f}$ and $\hat{g}$ and by monotonizing $\hat{F}(x_1, x_2)$ and $\widehat{G}(x_1, x_2)$ in the standard way (e.g., as functions of x_2 for small, fixed x_1). An alternative approach is to compute estimators that are constrained to be densities, or constrained to be distributions, by fitting them to the characteristic function estimator $\widehat{\chi}(t \mid a, b)$ defined in Section 2.2. This is the method suggested in the next section.

2.4. *Histogram-based estimators.* Recall the definition of $\widehat{\chi}(t \mid a, b)$ at (2.3) and note that $\widehat{\chi}(t \mid u, 1-u) = \widehat{\chi}_1(t, u) + i\,\widehat{\chi}_2(t, u)$, where $\widehat{\chi}_1$ and $\widehat{\chi}_2$ are real-valued functions,

$$\widehat{\chi}_j(t \mid u) = \frac{1}{nr(r-1)} \sum_{j=1}^{n} \sum\sum_{1 \le s_1, s_2 \le r\,:\, s_1 \ne s_2} \mathrm{trig}_j\bigl[t\{uY_{js_1} + (1-u)Y_{js_2}\}\bigr],$$

trig_1 denotes the cosine function, and trig_2 is the sine.

Observe too that, by (2.4),

$$\chi(t \mid a, b) = \Bigl\{\int e^{i(a+b)tx} f(x)\, dx\Bigr\}\Bigl\{\int e^{iatx} g(x)\, dx\Bigr\}\Bigl\{\int e^{ibtx} g(x)\, dx\Bigr\}. \tag{2.11}$$

Suppose for the present that f and g are histograms, with heights f_k and g_k, respectively, on intervals (x_k, x_{k+1}) for $-\infty < k < \infty$. These intervals are the histogram bins; for simplicity, we take them to have equal widths. We shall assume the x_k's are given; for example, they might be integer multiples of the common bin width. In the histogram case,

$$\int e^{itx} f(x)\, dx = \frac{i}{t} \sum_{-\infty < k < \infty} e^{itx_k} (f_k - f_{k-1}).$$

From this formula, and its analogue for g, we may deduce an expression for the right-hand side of (2.11): when $a = u$ and $b = 1 - u$ the right-hand side

has the form $K_1(t, u \mid \mathbf{f}, \mathbf{g}) + iK_2(t, u \mid \mathbf{f}, \mathbf{g})$, where K_1 and K_2 are real-valued functions, $\mathbf{f}$ and $\mathbf{g}$ denote the sequences of values f_k and g_k, respectively, and

$$K_j(t, u \mid \mathbf{f}, \mathbf{g}) = \frac{(-1)^{j+1}}{t^3 u(1-u)} \sum_{k_1} \sum_{k_2} \sum_{k_3} \operatorname{trig}_j\left[\frac{1}{2}\pi - \{tx_{k_1} + tux_{k_2} + t(1-u)x_{k_3}\}\right] \times (f_{k_1} - f_{k_1-1})(g_{k_2} - g_{k_2-1})(g_{k_3} - g_{k_3-1}).$$

We suggest computing empirical versions $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ of $\mathbf{f}$ and $\mathbf{g}$ by minimizing the distance of $\widehat{\chi}(t \mid u, 1-u)$ from $\chi(t \mid u, 1-u)$ when the latter is defined in the histogram case. Thus, our density estimators will be histograms, and the smoothing parameter will be the common bin width, $h = x_{k+1} - x_k$. More particularly, letting w denote a nonnegative weight function, we suggest estimating $\mathbf{f}$ and $\mathbf{g}$ as the minimizers of

$$J(\mathbf{f}, \mathbf{g}, h) = \int_{-\infty<t<\infty} w(t)\, dt \times \int_0^1 \{|\widehat{\chi}_1(t, u) - K_1(t, u|\mathbf{f}, \mathbf{g})|^2 + |\widehat{\chi}_2(t, u) - K_2(t, u|\mathbf{f}, \mathbf{g})|^2\}\, du, \tag{2.12}$$

subject to

$$\begin{aligned} &\text{(a)} \quad f_k \geq 0 \quad \text{and} \quad g_k \geq 0 \qquad \text{for each } k, \\ &\text{(b)} \quad h\sum_k f_k = h\sum_k g_k = 1, \\ &\text{(c)} \quad \sum_k f_k(x_k + x_{k+1}) = \sum_k g_k(x_k + x_{k+1}) = 0, \\ &\text{(d)} \quad h\sum_k (f_k + g_k)(x_k^2 + x_k x_{k+1} + x_{k+1}^2) \leq 3C(nr)^{-1} \sum_j \sum_s Y_{js}^2, \end{aligned} \tag{2.13}$$

where $C > 1$ is arbitrary. Conditions (a) and (b) require that the histogram densities be nonnegative and integrate to 1, respectively, (c) requires that the distributions corresponding to the densities have zero mean, and (d) requires that the sum of the variances be no more than C times the variance of the dataset $\{X_{js}\}$ generated by the model at (2.2). The latter constraint serves to prevent the algorithm from producing distribution estimates that are too highly variable relative to the empirical variance. In Section 4 we suggest a simple-to-code simulated annealing approach to solving the problem.

We could generalize (2.12) by multiplying the quantities $|\widehat{\chi}_j - K_j|^2$ by respective weights $w_j(u)$, and taking the integral with respect to u over a wider domain than simply the interval [0, 1]. This has the potential to alter efficiency, although it will not change convergence rates of estimators. Additionally, one could incorporate within-group information from other sources and among-group

information. Sieving via histograms is also appropriate; it leads to spline-based estimators.

Our final estimators of f and g are thus

$$\tilde{f}(x) = \sum_{-\infty<k<\infty} \tilde{f}_k \, I\{x \in (x_k, x_{k+1})\},$$

$$\tilde{g}(x) = \sum_{-\infty<k<\infty} \tilde{g}_k \, I\{x \in (x_k, x_{k+1})\},$$

where $\tilde{f}_k$ and $\tilde{g}_k$ are the kth elements of $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$, respectively, and I denotes the indicator function. The corresponding distribution estimators, $\widetilde{F}$ and $\widetilde{G}$ say, are the integrals of $\tilde{f}$ and $\tilde{g}$, respectively. A more sophisticated approach would use more general histosplines on the bins, rather than simply histograms (histosplines of order 0).

We would generally restrict the range of bins for which f_k and g_k were nonzero, for example, by taking them to lie within the range of the data Y_{js}. In the case of distribution estimation there seems no good theoretical reason for restricting the bin width h. In this setting we may interpret J at (2.12) as a function of h as well as of $\mathbf{f}$ and $\mathbf{g}$ and take the minimum over h as well as over the histogram heights f_k and g_k. In computational practice a lower bound on the value of h is generally determined by feasible computational time and occurs well before numerical instabilities arise.

Thus, for distribution estimation using the present method, there is a natural way of selecting the smoothing parameter. Empirical choice of h for density estimation, or (in the case of the method proposed in Section 2.3) choice of t_n for either density or distribution estimation, is more of a problem, however. Neither cross-validation nor substitution methods seem to have attractive counterparts in the present setting.

A third approach to distribution estimation would be to approximate the characteristic functions of F and G by appropriate exponential functions of empirical moments or cumulants, where the number of moments used grew slowly with sample size. Approaches of this type were touched on in Section 2.1.

3. Numerical properties.

3.1. *Introduction and summary.* We report results of a simulation study in two cases, first where F and G are both standard normal distributions, and second where both are exponential. These examples were chosen because they are both potentially difficult for different reasons. In the normal case, the fact that the density is analytic means that its characteristic function has pathologically light tails. That makes it difficult to recover high frequencies by Fourier inversion. As a result, convergence rates in a range of deconvolution problems involving

normal errors are particularly slow; see, for example, Carroll and Hall (1988) and Fan (1991).

The second problem is potentially difficult because the density of the exponential distribution has a marked discontinuity at its finite boundary. This makes density estimation awkward and likewise complicates distribution estimation when (as in the case of our first approach) the distribution estimators are intrinsically very smooth functions.

We also experimented with other distribution combinations, for example, the case of normal F and exponential G, and the opposite combination. Little that was new was learned from such cases, however, so results there will not be detailed here. Out of interest, in the exponential example we present results for the case where F is skewed to the right and G is skewed to the left. Results for exponential F and G skewed in the same direction were similar.

Our second technique, suggested in Section 2.4 and based on implicit histogram approximation, performed strongly. In general, it produced distribution estimators with low variability. In this respect it was preferable to the first method, introduced in Section 2.3 and based on explicit characteristic function inversion. However, implementation of the second method demanded substantially more computing time, and because of its histogram nature it tended to produce rougher, less pleasing density estimates. Since distribution estimation was our main goal, we did not regard the latter difficulty as a major drawback, but the heavy computational labor required by the second method was a problem.

For neither of our methods do we have a satisfactory empirical rule for smoothing parameter choice when the goal is density estimation. Throughout our numerical work we used adaptive methods to choose the amount of smoothing, but in the case of density estimation they appear to undersmooth; they are more appropriate to distribution than to density estimation. However, the density estimation problem is of less direct statistical importance than distribution estimation, and moreover (as we shall show) our empirical smoothing parameter rules for distribution estimation perform well.

3.2. *Explicit characteristic function inversion.* For this method, performance of the estimator depends on choice of the smoothing parameter t_n. One approach to selecting t_n would be to use cross-validation to choose values that would be optimal for estimating the distribution or density of $\xi_j + \varepsilon_{js}$. For this purpose one could adapt methods employed by Sarda (1993) and Bowman, Hall and Prvan (1998) in the case of distribution estimation, and Rudemo (1982) and Bowman (1984) for density estimation. However, we found this method too computer intensive in the present setting. Additionally, its intuitive appeal was significantly diminished by the fact that it targeted only the convolution of the distributions of ξ and ε. Instead we used the following more numerically efficient rule which, it can be proved, guarantees consistent distribution estimation.

When implementing the characteristic function inversion method to estimate distributions, we employed smoothing parameters $t_n^{(1)}$ and $t_n^{(2)}$ when estimating $F(x)$ and $G(x)$, respectively, using tapered versions of the estimators suggested at (2.10) for $x = x_2$ and x_1 large and negative. We chose $t_n^{(1)}$ and $t_n^{(2)}$ to minimize

$$\int_{-\infty}^{\infty} |\widehat{\chi}(t \mid 1, 0) - \hat{\phi}(t \mid t_n^{(1)}) \hat{\psi}(t \mid t_n^{(2)})| e^{-Ct^2}\, dt,$$

where $\widehat{\chi}(t \mid 1, 0)$ was as defined at (2.3), $\hat{\phi}(\cdot \mid t_n)$ and $\hat{\psi}(\cdot \mid t_n)$ were the characteristic functions corresponding to the distribution estimators, and $C > 0$. We chose $C = \frac{1}{2}$, making the weights proportional to the standard normal density. When estimating densities we employed the same procedure, except that $t_n^{(1)}$ and $t_n^{(2)}$ were now used to construct tapered versions $\hat{f}(\cdot \mid t_n)$ and $\hat{g}(\cdot \mid t_n)$ of the estimators at (2.9). These in turn led to $\hat{\phi}(\cdot \mid t_n)$ and $\hat{\psi}(\cdot \mid t_n)$. We constructed $\hat{f}(\cdot \mid t_n)$ as

$$\hat{f}(x \mid t_n) = (2\pi)^{-1} \Re \int_{-\infty}^{\infty} e^{-itx} \hat{\phi}_{\mathrm{tr}}(t) K(t \mid t_n)\, dt,$$

where $K_n(t \mid t_n) = 1$ for $|t| \le t_n$ and $K(t \mid t_n) = \exp\{-C(|t| - t_n)^2\}$ for $|t| > t_n$ and $C > 0$. (Again we chose $C = \frac{1}{2}$.) The tapering used to construct $\hat{F}$ and $\widehat{G}$ was done analogously.

One might expect, from experience with random coefficient regression, tomographic inversion or other regularization methods in statistics, that this problem requires rather large samples if good quantitative performance (as distinct from qualitatively accurate results achievable through fitting low-order moments) is to be achieved. This does indeed appear to be the case. In the (normal, normal) and (exponential, exponential) cases we drew 100 samples with $n = 1000$ and $r = 2$. Figure 1 summarizes simulation results when $t_n^{(1)}$ and $t_n^{(2)}$ were chosen between 1 and 10. Panels (a)–(d) show results for the (normal, normal) case. Each of panels (a) and (c) gives pointwise median curves and pointwise 90% quantile curves computed from the 100 distribution estimates, while panels (b) and (d) show analogous information obtained from the 100 density estimates. Similarly, panels (e)–(h) show results for the (exponential, exponential) case.

The characteristic function inversion method performs well in the (normal, normal) case, but it has more difficulty in the (exponential, exponential) problem. Nevertheless, in the latter setting it captures the shape of the true distribution, even though it shows significant variability. For density estimation in the (exponential, exponential) case, the best that can be said of the method is that it captures skewness reasonably well. However, bearing in mind that the density estimator is constrained to be a smooth curve, and the target density is characterized by a marked discontinuity, density estimation in the (exponential, exponential) case is arguably too difficult a problem for this technique.

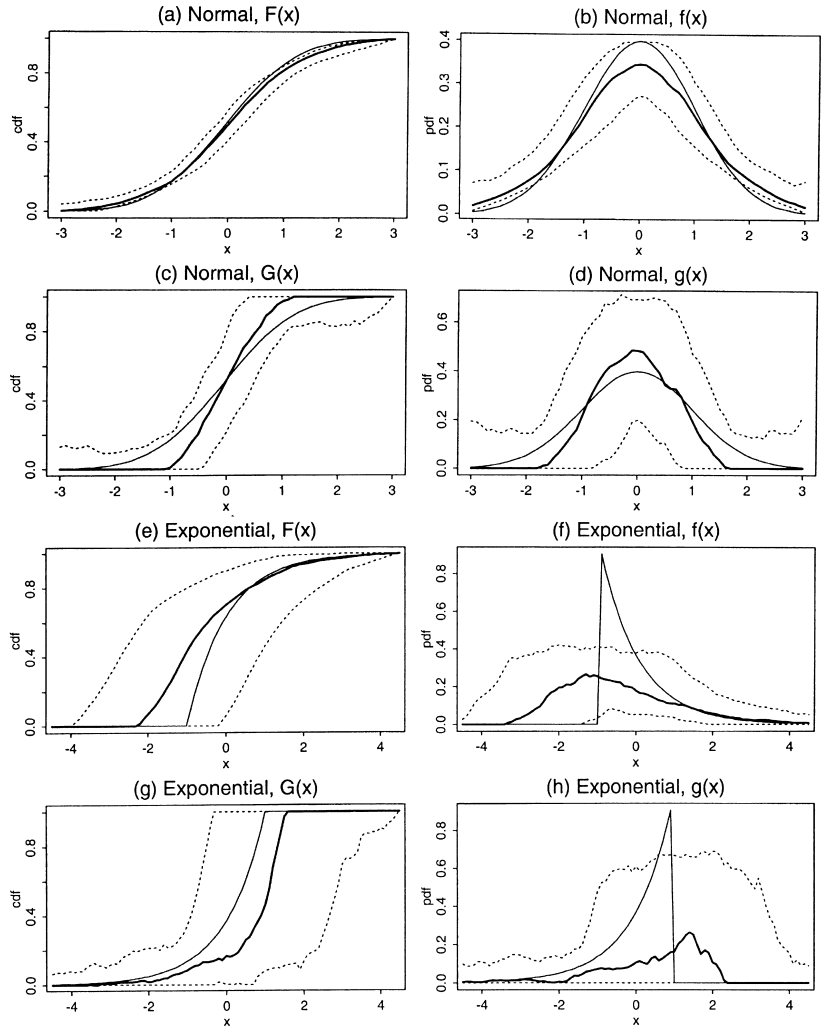

FIG. 1. *Explicit characteristic function inversion. True distribution and density functions are shown by thin unbroken lines, pointwise medians of curve estimates are depicted by thick unbroken lines, and pointwise* 90% *upper and lower quantiles are indicated by dotted lines. The top four panels give results in the* (*normal, normal*) *case, and the lower four panels show results in the* (*exponential, exponential*) *setting. Results for distribution and density function estimates are given in the first and second columns, respectively.*

3.3. *Histogram-based estimators.* This method was proposed in Section 2.4, along with a technique for choosing the smoothing parameter. We used a simulated annealing approach, as follows. Starting with initial estimators $\tilde{f}_0$ and $\tilde{g}_0$, we added an independent random Uniform$(0, \delta)$ perturbation to each of the bin heights of both $\tilde{f}_0$ and $\tilde{g}_0$, where $\delta > 0$ was a small constant. The new estimators $\tilde{f}_1$ and $\tilde{g}_1$ were obtained by standardizing the perturbed histograms, as follows.

We set the new bin height to 0 if its perturbed value was negative, we normalized the histograms so they were proper densities, and we shifted the supports of the histograms so their means were 0. We took $(\tilde{f}_0, \tilde{g}_0) = (\tilde{f}_1, \tilde{g}_1)$ only if $J(\tilde{f}_1, \tilde{g}_1) \le J(\tilde{f}_0, \tilde{g}_0) + \tau$, where $\tau > 0$ was a small constant. [Recall that J was defined at (2.12).]

We iterated this procedure until the minimum value of J, over successive versions of $(\tilde{f}_0, \tilde{g}_0)$, was not reduced after a large predetermined number, N say, of attempts. We then repeated the above procedure m times, using reduced values of δ and τ. The final estimator $(\tilde{f}, \tilde{g})$ was the overall minimizer of J in the search process. In theory the algorithm can converge to a local extremum, but the chance of this occurring is minimized by starting the algorithm in different places and checking that the same limit is achieved.

Throughout we took the number of bins to be 10. In the initial step of the algorithm we took all bin heights to be equal. We set $N = 3000$ and $\delta = 0.5$ and took τ equal to 20% of the value J for the initial estimators. The procedure described in the previous paragraph was repeated $m = 4$ times, each time reducing δ and τ by 70% and increasing N by 1000. The distribution estimators were simply integrals of the density estimators; unlike the case in Section 3.3, no attempt was made to smooth differently in the two problems.

We drew 40 samples with $n = 1000$ and $r = 2$. Each replication took about 4.8 hours using a PC equipped with a Pentium III 1 GHz processor. Results are displayed in Figure 2. Analogously to Figure 1 they show pointwise medians and 90% confidence bounds. The relatively low variation of histogram-based distribution estimators, compared with estimators produced by the first method, is clear on comparing the first columns of Figures 1 and 2. In the (exponential, exponential) case the histogram-based approach also produces more accurate density estimators.

3.4. *A real-data example.* Finally, we apply the histogram-based method to a dataset reported by Heckman (1960). An experiment was conducted to compare two approaches (i.e., $r = 2$) to measuring the calcium content of animal feeds. Data on the percentage calcium content, using either technique, were recorded for $n = 118$ feed samples. Assuming the model (2.2), we estimated the density functions f and g. The range of the data was 7.09, and the difference of the measurements from the two methods was always less than 0.5, so we assumed that ξ_j and ε_{js} were distributed on intervals with lengths 7.09 and 0.5, respectively. Simulated annealing was used in the same manner as in Section 3.3, but now with $N = 100{,}000$ and the increment 10,000 in each of $m = 4$ replications. The histogram estimators, in the cases of 4 and 6 bins, are plotted in Figure 3. They suggest that the distribution of ξ_j might have at least two modes, with the largest mode around -2 and another around 4, and that the distribution of ε_{js} may be unimodal with its mode near 0.1.

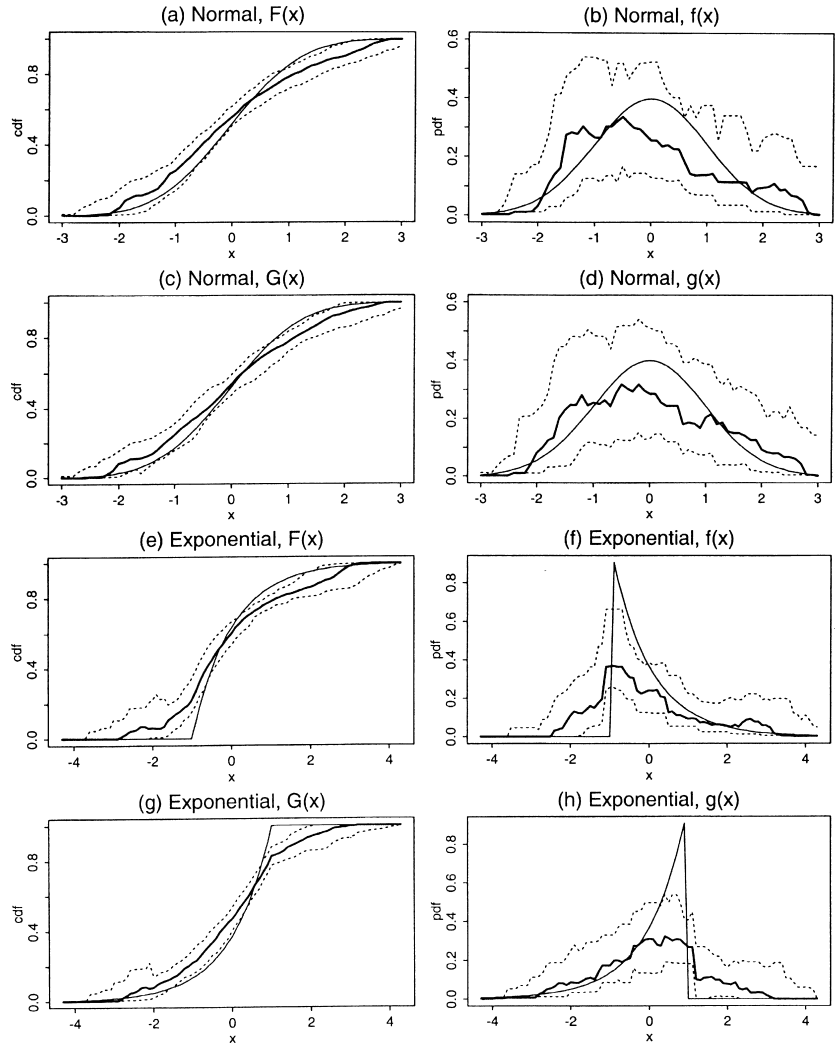

FIG. 2. *Histogram-based estimators. Legend and arrangement of panels is as for Figure* 1.

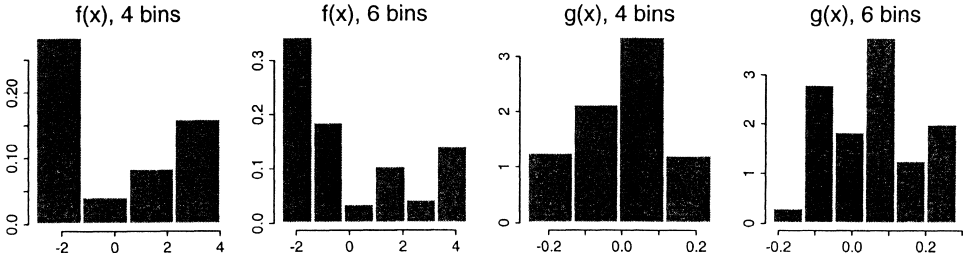

FIG. 3. *Histogram estimators computed for the calcium data example in Section* 3.5.

4. Theoretical properties.

4.1. *Explicit characteristic function inversion.* Here we treat the method suggested in Section 2.3. Our first result shows that under mild regularity conditions the estimators $\hat{\phi}$ and $\hat{\psi}$, and hence $\hat{\phi}_{\rm tr}$ and $\hat{\psi}_{\rm tr}$, are root-n consistent for ϕ and ψ.

THEOREM 4.1. *Assume that the distributions F and G are continuous and have finite moments of order $2+\eta$ for some $\eta > 0$ and that $r \geq 2$ is fixed. If, for a particular t, $\phi(t)$ does not vanish and $\psi(t/2^j)$ does not vanish for any $j \geq 0$, then $\hat{\phi}(t)$ and $\hat{\psi}(t)$ are root-n consistent for $\phi(t)$ and $\psi(t)$, respectively, as $n \to \infty$. Furthermore, if for some $t_0 \in (0, \infty)$ neither ϕ nor ψ vanishes in the interval $[-t_0, t_0]$, then $\hat{\phi}(t)$ and $\hat{\psi}(t)$ are uniformly root-n consistent there*:

$$\sup_{|t| \leq t_0} \{|\hat{\phi}(t) - \phi(t)| + |\hat{\psi}(t) - \psi(t)|\} = O_p(n^{-1/2}). \tag{4.1}$$

Next we show that the density estimators $\hat{f}$ and $\hat{g}$, defined at (2.9), are consistent if the smoothing parameter t_n increases sufficiently slowly. Our proof of Theorem 4.2 will show that under the conditions there, $\hat{\phi}$ and $\hat{\psi}$ converge strongly (i.e., with probability 1) to their respective limits ϕ and ψ at all but at most countably many points.

THEOREM 4.2. *Assume that the distributions F and G have densities f and g, respectively, that their respective characteristic functions ϕ and ψ are absolutely integrable and vanish at no more than a countable number of points, and that both distributions have moments of order $1+\eta$ for some $\eta > 0$. Then there exists a sequence of positive constants τ_n, increasing to ∞, such that, provided $t_n \to \infty$ and $t_n \leq \tau_n$, $\hat{f}(x)$ and $\hat{g}(x)$ converge to $f(x)$ and $g(x)$, respectively, uniformly in x with probability* 1.

The assumption, in Theorem 4.2, that $|\phi|$ and $|\psi|$ are integrable is a mild smoothness condition. It holds if fractional derivatives of f' and g', of arbitrarily small but positive order, exist and are integrable.

Convergence rates of density and distribution estimators depend on the tail behavior of the characteristic functions ϕ and ψ. For simplicity, and to illustrate theoretical arguments that can be used more generally, we shall assume that ϕ and ψ both decrease polynomially fast and that neither function vanishes: for constants $\alpha, \beta > 1$,

$$\begin{aligned} &\text{both } |\phi(t)|(1+|t|)^{\alpha} \text{ and } |\psi(t)|(1+|t|)^{\beta} \text{ are} \\ &\text{bounded away from 0 and } \infty, \text{ uniformly in } t. \end{aligned} \tag{4.2}$$

Examples satisfying (4.2) include gamma distributions. Condition (4.2) is one of smoothness, which increases with the values of α and β. For example, in the

gamma case α and β are identical to the distributions' respective exponents and are increasing functions of the maximum number of derivatives that the densities have on the real line. (A gamma density is infinitely differentiable everywhere except at the origin, and so its smoothness at the origin determines overall differentiability.)

Our methods can also be used to derive convergence rates in many other settings, for example, when (4.2) holds in the characteristic function tails but either ϕ or ψ vanishes at a finite number of points, or when the tails of ϕ and ψ decrease exponentially fast. In addition to (4.2) we shall assume of the smoothing parameter t_n that, for some $\eta > 0$,

$$t_n \to \infty \quad \text{and} \quad t_n^{\gamma+1} = O(n^{(1/2)-\eta}), \tag{4.3}$$

where $\gamma = \alpha + 2\beta$. Define $\hat{F}$ as at (2.10) and define $\widehat{G}$ analogously.

THEOREM 4.3. *Assume the distributions F and G have finite moments of order $2 + \eta$ for some $\eta > 0$ and that* (4.2) *and* (4.3) *hold. Then, for each $x_0 > 0$,*

$$\sup_{-\infty<x<\infty} |\hat{f}(x) - f(x)| = O_p\bigl(t_n^{1-\alpha} + t_n^{\gamma-\alpha+(1/2)} n^{-1/2}\bigr), \tag{4.4}$$

$$\sup_{-\infty<x<\infty} |\hat{g}(x) - g(x)| = O_p\bigl(t_n^{1-\beta} + t_n^{\gamma-\beta+(1/2)} n^{-1/2}\bigr), \tag{4.5}$$

$$\sup_{x_0\le x_1\le x_2\le x_0} |\hat{F}(x_1, x_2) - F(x_1, x_2)| = O_p\bigl(t_n^{-\alpha} + t_n^{\gamma-\alpha-(1/2)} n^{-1/2}\bigr), \tag{4.6}$$

$$\sup_{x_0\le x_1\le x_2\le x_0} |\widehat{G}(x_1, x_2) - G(x_1, x_2)| = O_p\bigl(t_n^{-\beta} + t_n^{\gamma-\beta-(1/2)}\, n^{-1/2}\bigr). \tag{4.7}$$

The first term on the right-hand side of each of (4.4)–(4.7) represents the order of bias, and the second is the order of stochastic error about the mean. The fact that both terms are a little smaller in the case of distribution estimation, that is, in (4.6) and (4.7), reflects the intrinsic relative simplicity of that problem; in particular, a distribution function is smoother than its density. Upper bounds to convergence rates may be derived by equating the two respective terms to obtain an order of magnitude for t_n. For example, taking t_n to equal a constant multiple of $n^{-1/(2\gamma-1)}$, we may show from (4.4) that

$$\sup_{-\infty<x<\infty} |\hat{f}(x) - f(x)| = O_p\bigl(n^{-(\alpha-1)/(2\gamma-1)}\bigr).$$

4.2. *Histogram-based estimators.* Here we treat the implicit method suggested in Section 2.4. It produces estimators $\widetilde{F}$ and $\widetilde{G}$ of F and G, respectively; let the corresponding characteristic functions be $\tilde{\phi}$ and $\tilde{\psi}$. We choose the histograms $\tilde{f}$ and $\tilde{g}$ (the respective densities of $\widetilde{F}$ and $\widetilde{G}$) and the bin width h simultaneously, by minimizing $J(\mathbf{f}, \mathbf{g}, h)$ defined at (2.12). Of course, in theory the minimum of J will occur at the limit as $h \downarrow 0$, and an argument based on limits of subsequences of characteristic functions shows that proper distributions $\widetilde{F}$ and $\widetilde{G}$ arise at this

practically infeasible limit. These distributions may be approximated arbitrarily closely by regimes where $h > 0$.

THEOREM 4.4. *Assume the distributions F and G have uniformly continuous densities, finite variances and characteristic functions that vanish at no more than a countable number of points. Suppose, too, that the weight function w is strictly positive on the whole real line and is bounded and continuous and satisfies $\int (1+t^2)\, w(t)\, dt < \infty$. Then* (a) *both* $\sup|\widetilde{F} - F|$ *and* $\sup|\widetilde{G} - G|$ *converge to* 0 *with probability* 1 *and* (b)

$$\int_{-\infty}^{\infty} w(t)\, dt \int_0^1 \bigl|\phi(t)\psi(tu)\psi\{t(1-u)\} - \tilde{\phi}(t)\tilde{\psi}(tu)\tilde{\psi}\{t(1-u)\}\bigr|^2\, du = O_p(n^{-1}).$$

5. Technical details.

5.1. *Proof of Proposition* 2.1. Since F and G are continuous, then for each t the probability that either $\widehat{\chi}(t \mid 1, 0)$ or $\widehat{\chi}(t \mid \frac{1}{2}, \frac{1}{2})$ vanishes equals 0. Hence the probability that either $\widehat{\chi}(t/2^j \mid 1, 0)$ or $\widehat{\chi}(t/2^j \mid \frac{1}{2}, \frac{1}{2})$ vanishes for some $j \geq 0$ equals 0. Therefore the proposition will follow if we prove that the series on the right-hand side of (2.7) converges. By definition of Y_{js}, $\sum_j \sum_s Y_{js} = 0$, and so, for each real a and b, $\widehat{\chi}(t \mid a, b) = 1 - \frac{1}{2}t^2\hat{\sigma}(a, b)^2 + O(|t|^3)$, with probability 1, as $t \to 0$, where

$$\hat{\sigma}(a, b)^2 = \frac{1}{nr(r-1)} \sum_{j=1}^{n} \sum\sum_{1 \leq s_1, s_2 \leq r\,:\, s_1 \neq s_2} \bigl(aY_{js_1} + bY_{js_2}\bigr)^2 < \infty.$$

Hence, as $j \to \infty$,

$$\begin{aligned} &2^j\bigl\{\log \widehat{\chi}(t/2^j \mid 1, 0) - \log \widehat{\chi}\bigl(t/2^j \mid \tfrac{1}{2}, \tfrac{1}{2}\bigr)\bigr\} \\ &\qquad = 2^{-j-1}t^2\bigl\{\hat{\sigma}\bigl(\tfrac{1}{2}, \tfrac{1}{2}\bigr)^2 - \hat{\sigma}(1, 0)^2\bigr\} + O(2^{-2j}), \end{aligned}$$

with probability 1 as $t \to 0$. It follows that the infinite series at (2.7) converges with probability 1.

5.2. *Proof of Theorem* 4.1. We shall derive only (4.1). For fixed real numbers a and b, let A denote the distribution of $aY_{j1} + bY_{j2}$ and let $\widehat{A}_1$ be the empirical distribution corresponding to the dataset $\{aY_{js_1} + bY_{js_2} : 1 \leq j \leq n, 1 \leq s_1, s_2 \leq r, s_1 \neq s_2\}$. Then, integrating by parts, we may show that, for an absolute constant $C_1 > 0$,

$$\begin{aligned} |\widehat{\chi}(t \mid a, b) - \chi(t \mid a, b)| &\leq \biggl|\int \cos(tx)\, d\bigl\{\widehat{A}_1(x) - A(x)\bigr\}\biggr| \\ &\quad + \biggl|\int \sin(tx)\, d\bigl\{\widehat{A}_1(x) - A(x)\bigr\}\biggr| \\ &\leq C_1|t| \int \min\bigl\{1, (tx)^2\bigr\}\bigl|\widehat{A}_1(x) - A(x)\bigr|\, dx. \end{aligned} \tag{5.1}$$

(Here and below, integrals without specified limits will be assumed to be over the whole real line.) Therefore we may show that, uniformly in $|t| \le t_0$ and for all $\zeta > 0$,

$$S(t \mid a, b) \equiv \sum_{j=0}^{\infty} 2^j \bigl|\{\widehat{\chi}(t/2^j \mid a, b) - \chi(t/2^j \mid a, b)\}\big/\chi(t/2^j \mid a, b)\bigr|$$

$$= O_p\biggl[\sum_{j=0}^{\infty} \int \min\{1, (tx/2^j)^2\} \bigl|\widehat{A}_1(x) - A(x)\bigr|\, dx\biggr] \tag{5.2}$$

$$= O_p\biggl\{\int (1 + |x|^{\zeta}) \bigl|\widehat{A}_1(x) - A(x)\bigr|\, dx\biggr\}.$$

Put $U_{js} = \xi_j + \varepsilon_{js}$ and $\bar{U} = (nr)^{-1} \sum_j \sum_s U_{js}$. Let $\widehat{A}_2$ denote the empirical distribution function of the dataset $\{aU_{js_1} + bU_{js_2} : 1 \le j \le n,\ 1 \le s_1, s_2 \le r,\ s_1 \ne s_2\}$. Then $\widehat{A}_1(x) = \widehat{A}_2(x + \bar{U})$, and so

$$\int (1 + |x|^{\zeta}) \bigl|\widehat{A}_1(x) - A(x)\bigr|\, dx$$

$$\le \int (1 + |x - \bar{U}|^{\zeta}) \bigl|\widehat{A}_2(x) - A(x)\bigr|\, dx$$

$$+ \int (1 + |x|^{\zeta}) \bigl|A(x + \bar{U}) - A(x)\bigr|\, dx.$$

We may write $\widehat{A}_2$ as the mean of $\frac{1}{2}r(r-1)$ empirical distribution functions, each of which is computed from n independent and identically distributed random variables. Arguing in this way, we may prove, using the fact that moments of order $2 + \eta$ are finite for some $\eta > 0$, that if $0 < \zeta < \frac{1}{2}\eta$, then

$$\int (1 + |x|^{\zeta}) \bigl|\widehat{A}_2(x) - A(x)\bigr|\, dx = O_p(n^{-1/2}).$$

Moreover, if $\bar{U} > 0$, then

$$\int_0^{\infty} (1 + |x|^{\zeta}) \bigl|A(x + \bar{U}) - A(x)\bigr|\, dx$$

$$= \int_0^{\bar{U}} (1 + x^{\zeta})\{1 - A(x)\}\, dx$$

$$+ \int_{\bar{U}}^{\infty} \{x^{\zeta} - (x - \bar{U})^{\zeta}\}\{1 - A(x)\}\, dx,$$

which, for sufficiently small $\zeta > 0$, equals $O_p(n^{-1/2})$ since $\bar{U} = O_p(n^{-1/2})$. Analogous results hold if $\bar{U} < 0$ or if the integral on the left is taken over $-\infty < x < 0$. Therefore, provided $\zeta > 0$ is sufficiently small,

$$\int (1 + |x|^{\zeta}) \bigl|A(x + \bar{U}) - A(x)\bigr|\, dx = O_p(n^{-1/2}).$$

Combining the results in the previous paragraph, we deduce that

$$\int (1+|x|^{\zeta})|\widehat{A}_1(x)-A(x)|\,dx = O_p(n^{-1/2}), \tag{5.3}$$

which along with (5.2) implies that $S(t \mid a,b) = O_p(n^{-1/2})$ uniformly in $|t| \le t_0$. Therefore

$$\sum_{j=0}^{\infty} 2^j \bigl|\log\bigl[1+\{\widehat{\chi}(t/2^j \mid a,b)-\chi(t/2^j \mid a,b)\}\chi(t/2^j \mid a,b)^{-1}\bigr]\bigr| = O_p(n^{-1/2}),$$

uniformly in $|t| \le t_0$. This result, along with the fact that

$$\log\{\hat{\psi}(t)/\psi(t)\} = \sum_{j=0}^{\infty} 2^j \log\bigl[\{1+\hat{\delta}(t/2^j \mid 1,0)\}/\{1+\hat{\delta}(t/2^j \mid \tfrac{1}{2},\tfrac{1}{2})\}\bigr], \tag{5.4}$$

where $\hat{\delta}(t \mid a,b) = \{\widehat{\chi}(t \mid a,b)-\chi(t \mid a,b)\}/\chi(t \mid a,b)$, implies that, uniformly in $|t| \le t_0$,

$$\hat{\psi}(t) = \psi(t) + O_p(n^{-1/2}). \tag{5.5}$$

From (5.1), with $(a,b) = (1,0)$, and (5.3), it follows that $\widehat{\chi}(t \mid 1,0) = \chi(t \mid 1,0) + O_p(n^{-1/2})$ uniformly in $|t| \le t_0$. The latter result and (5.5) imply that $\hat{\phi}(t) = \phi(t) + O_p(n^{-1/2})$ uniformly in $|t| \le t_0$, completing the proof of (4.1).

5.3. *Proof of Theorem* 4.2. Let (P_1) denote the property that $\hat{\phi}(t) \to \phi(t)$ with probability 1 for all but a countable number of points t. Observe that

$$2\pi \sup_{-\infty<x<\infty} |\hat{f}(x)-f(x)| \le \int_{|t|\le t_n} |\hat{\phi}_{\mathrm{tr}}(t)-\phi(t)|\,dt + \int_{|t|>t_n} |\phi(t)|\,dt.$$

It follows from this result, and from the fact that $|\hat{\phi}_{\mathrm{tr}}(t)-\phi(t)| \le 2$ for each t, that if (P_1) holds, and if we choose the constants τ_n to diverge so slowly that

$$\int_{|t|\le\tau_n} |\hat{\phi}_{\mathrm{tr}}(t)-\phi(t)|\,dt \to 0,$$

with probability 1, then $\sup|\hat{f}-f| \to 0$ with probability 1 whenever $\hat{f}$ is defined using a sequence $t_n \le \tau_n$ for which $t_n \to \infty$.

An identical argument applies in the case of $\hat{g}$; there we should prove that $\hat{\psi} \to \psi$ with probability 1 at all but countably many points. Let $\mathcal{T}$ denote the set of t such that either ϕ or ψ vanishes at one or more elements of the set $\{t, t/2, t/4, \ldots\}$, and let $\mathcal{T}^{\mathrm{c}}$ denote the complement of $\mathcal{T}$ in the real line. We shall show that $\hat{\psi}(t)$ and $\widehat{\chi}(t \mid 1,0)$ converge to $\psi(t)$ and $\chi(t \mid 1,0)$, respectively, with probability 1 for each $t \in \mathcal{T}^{\mathrm{c}}$; call this property (P_2). Since $\mathcal{T}$ is countable, (P_2) implies (P_1), and so we have proved the theorem.

To establish (P_2), return to step (5.2) of the proof of Theorem 4.1 and note that, using the arguments there, we may show that, provided $t \in \mathcal{T}^c$,

$$S(t \mid a, b) = O\left\{\int (1 + |x|^\zeta)|\widehat{A}_1(x) - A(x)|\,dx\right\} \tag{5.6}$$

for all $\zeta > 0$, with probability 1. Now, for each $x_0 > 0$,

$$\begin{aligned}&\int (1 + |x|^\zeta)|\widehat{A}_1(x) - A(x)|\,dx \\ &\qquad \le 2x_0(1 + x_0^\zeta) \sup_{-\infty < x < \infty} |\widehat{A}_1(x) - A(x)| + \widehat{R}(x_0),\end{aligned} \tag{5.7}$$

where

$$\begin{aligned}\widehat{R}(x_0) = &\int_{x_0}^{\infty} (1 + x^\zeta)\{1 - \widehat{A}_1(x) + 1 - A(x)\}\,dx \\ &+ \int_{-\infty}^{x_0} (1 + |x|^\zeta)\{\widehat{A}_1(x) + A(x)\}\,dx.\end{aligned}$$

In the notation of Section 5.2,

$$|\widehat{A}_1(x) - A(x)| \le |\widehat{A}_2(x + \bar{U}) - A(x + \bar{U})| + |A(x + \bar{U}) - A(x)|. \tag{5.8}$$

Recall from Section 5.2 that $\widehat{A}_2$ can be expressed as an average of a finite number of empirical distribution functions, each computed from n independent and identically distributed random variables having distribution function A. It follows that $\sup|\widehat{A}_2 - A| \to 0$ with probability 1. Since $\bar{U} \to 0$ with probability 1, $\sup_x |A(x + \bar{U}) - A(x)| \to 0$. Combining the results from (5.8) down, we deduce that $\sup|\widehat{A}_1 - A| \to 0$ with probability 1; call this result (P_3). Also, if $\zeta > 0$ is sufficiently small, then

$$\widehat{R}(x_0) \to R(x_0) \equiv 2\int_{x_0}^{\infty} (1 + x^\zeta)\{A(-x) + 1 - A(x)\}\,dx, \tag{5.9}$$

with probability 1, and the value of $R(x_0)$ can be made arbitrarily small by choosing x_0 sufficiently large. [Finiteness of $|R(x_0)|$ follows from the assumption of finite moments of order $1 + \eta$; we require $0 < \zeta \le \eta$.]

Combining (5.6), (5.7), (5.9) and (P_3), we deduce that, for each $t \in \mathcal{T}^c$, $S(t \mid a, b) \to 0$ with probability 1. Analogously to the proof of Theorem 4.1, this is sufficient to imply first that

$$\sum_{j=0}^{\infty} 2^j \left|\log\left[1 + \left\{\widehat{\chi}(t/2^j \mid a, b) - \chi(t/2^j \mid a, b)\right\}\chi(t/2^j \mid a, b)^{-1}\right]\right| \to 0,$$

with probability 1, and then that $\hat{\psi}(t) \to \psi(t)$ with probability 1; compare the argument leading to (5.5). Likewise, taking $(a, b) = (1, 0)$, we may deduce that $\widehat{\chi}(t \mid 1, 0) \to \chi(t \mid 1, 0)$ with probability 1. All these limit properties hold for each $t \in \mathcal{T}^c$, and so we have established (P_2).

5.4. *Proof of Theorem* 4.3. We shall derive only (4.5); results (4.4), (4.6) and (4.7) have similar proofs.

In establishing (4.5) we start with (5.4), from which it follows that, provided

$$\sup_{|t|\le t_n} S(t \mid a, b) = o_p(1) \tag{5.10}$$

for $(a, b) = (1, 0)$ and $(\frac{1}{2}, \frac{1}{2})$, we have, for each $\nu \ge 2$,

$$\hat{\psi}(t) = \psi(t)\big[1 + \Delta_1(t) + Q_{1\nu}(t) + O_p\big\{S(t \mid 1, 0)^{\nu+1} + S\big(t \mid \tfrac{1}{2}, \tfrac{1}{2}\big)^{\nu+1}\big\}\big], \tag{5.11}$$

uniformly in t, where

$$\Delta_1(t) = \sum_{j=0}^{\infty} 2^j \big\{\hat{\delta}(t/2^j \mid 1, 0) - \hat{\delta}\big(t/2^j \mid \tfrac{1}{2}, \tfrac{1}{2}\big)\big\} \tag{5.12}$$

and $Q_{1\nu}(t)$ is a finite linear form in terms

$$T(k_1, k_2, l_1, l_2) = \left\{\sum_{j=0}^{\infty} 2^j \hat{\delta}(t/2^j \mid 1, 0)^{k_1}\right\}^{l_1} \left\{\sum_{j=0}^{\infty} 2^j \hat{\delta}\big(t/2^j \mid \tfrac{1}{2}, \tfrac{1}{2}\big)^{k_2}\right\}^{l_2} \tag{5.13}$$

for $1 \le k_1, k_2 < \infty$, $0 \le l_1, l_2 < \infty$ and $2 \le k_1 l_1 + k_2 l_2 \le \nu$.

To derive (5.10), observe that (4.2) implies, for a constant $C_2 > 0$,

$$|\chi(t \mid 1, 0)| + \big|\chi\big(t \mid \tfrac{1}{2}, \tfrac{1}{2}\big)\big| \ge C_2(1 + |t|)^{-\gamma}. \tag{5.14}$$

Properties (5.1) and (5.14) imply that, for $(a, b) = (1, 0)$ or $(\frac{1}{2}, \frac{1}{2})$ and all t,

$$\begin{aligned} &|\widehat{\chi}(t \mid a, b) - \chi(t \mid a, b)|/|\chi(t \mid a, b)| \\ &\qquad \le C_3|t|(1 + |t|)^{\gamma} \int \min\{1, (tx)^2\} |\widehat{A}_1(x) - A(x)|\, dx. \end{aligned}$$

(Here and below, $C_2, C_3, \ldots$ denote positive constants depending on F and G but not on t or n.) The argument leading to (5.2) now shows that, uniformly in t,

$$\begin{aligned} S(t \mid a, b) &\le C_4 \sum_{j=0}^{\infty} (|t| + 2^{-\gamma j}|t|^{\gamma+1}) \\ &\qquad \times \int \min\{1, (tx/2^j)^2\} |\widehat{A}_1(x) - A(x)|\, dx. \end{aligned} \tag{5.15}$$

Since

$$\sum_{j=0}^{\infty} (|t| + 2^{-\gamma j}|t|^{\gamma+1}) \min\{1, (tx/2^j)^2\} \le C_5 \max(|t|, |t|^{\gamma+1}),$$

then, by (5.15),

$$\begin{aligned} S(t \mid a, b) &\le C_4 C_5 \max(|t|, |t|^{\gamma+1}) \int |\widehat{A}_1(x) - A(x)|\, dx \\ &= O_p\{\max(|t|, |t|^{\gamma+1}) n^{-1/2}\}, \end{aligned} \tag{5.16}$$

uniformly in t. To obtain the last identity, we have used the fact that distributions F and G have finite moments of order $2+\eta$ for some $\eta > 0$.

We may deduce from (4.2), (4.3), (5.11) and (5.16) that

$$\begin{aligned}\int_{|t|\le t_n} e^{-itx}\hat\psi(t)\,dt - \int_{|t|\le t_n} e^{-itx}\psi(t)\,dt \\ = \int_{|t|\le t_n} e^{-itx}\psi(t)\{\Delta_1(t) + Q_{1\nu}(t)\}\,dt \\ + O_p\bigl(t_n^{(\nu+1)(\gamma+1)+1-\beta}\, n^{-(\nu+1)/2}\bigr),\end{aligned} \tag{5.17}$$

uniformly in x. To derive a version of (5.17) in the case of $\hat\phi$ and ϕ, rather than $\hat\psi$ and ψ, note that $\widehat{\chi}(t \mid 1, 0) = \chi(t \mid 1, 0)\{1 + \hat\delta(t \mid 1, 0)\}$, whence we obtain an analogue of (5.11):

$$\begin{aligned}\hat\phi(t) &= \widehat{\chi}(t \mid 1, 0)/\hat\psi(t) \\ &= \phi(t)\bigl[1 + \Delta_2(t) + Q_{2\nu}(t) + O_p\bigl\{S(t \mid 1, 0)^{\nu+1} + S\bigl(t \mid \tfrac{1}{2}, \tfrac{1}{2}\bigr)^{\nu+1}\bigr\}\bigr],\end{aligned}$$

where $\Delta_2(t) = \hat\delta(t \mid 1, 0) - \Delta_1(t)$ and $Q_{2\nu}(t)$ is a finite linear form in terms

$$T(k_1, k_2, l_1, l_2)\hat\delta(t \mid 1, 0)^m$$

for $1 \le k_1, k_2 < \infty$, $0 \le l_1, l_2 < \infty$, $m = 0$ or 1, and $2 \le k_1 l_1 + k_2 l_2 + m \le \nu$, with $T(k_1, k_2, l_1, l_2)$ defined as at (5.13). Thus we obtain the following analogue of (5.17):

$$\begin{aligned}\int_{|t|\le t_n} e^{-itx}\hat\phi(t)\,dt - \int_{|t|\le t_n} e^{-itx}\phi(t)\,dt \\ = \int_{|t|\le t_n} e^{-itx}\phi(t)\{\Delta_2(t) + Q_{2\nu}(t)\}\,dt \\ + O_p\bigl(t_n^{(\nu+1)(\gamma+1)+1-\alpha} n^{-(\nu+1)/2}\bigr).\end{aligned} \tag{5.18}$$

We shall derive the rate of convergence of $\hat g$ to g, given at (4.5), starting from (5.17). An analogous argument would give the rate for $\hat f$ to f, starting from (5.18). We shall prove that

$$\int_{|t|\le t_n} e^{-itx}\psi(t)\Delta_1(t)\,dt = O_p\bigl(t_n^{\gamma-\beta+(1/2)} n^{-1/2}\bigr), \tag{5.19}$$

uniformly in x. A similar argument will show that

$$\int_{|t|\le t_n} e^{-itx} Q_{2\nu}(t)\,dt = O_p\bigl(t_n^{\gamma-\beta+(1/2)}\, n^{-1/2}\bigr).$$

In fact, the left-hand side immediately above is of smaller order; the quantities $\Delta_1(t)$ and $Q_{2\nu}(t)$ denote in effect linear and higher order terms, respectively, and the latter make a contribution of lower order than do the former.

Condition (5.1) implies that

$$(2\pi)^{-1}\int_{|t|\le t_n} e^{-itx}\psi(t)\,dt = f(x) - (2\pi)^{-1}\int_{|t|>t_n} e^{-itx}\psi(t)\,dt$$
$$= f(x) + O(t_n^{1-\beta}),$$

uniformly in x. Combining (5.17) with the results from (5.19) down, we deduce that

$$\hat g(x) = (2\pi)^{-1}\int_{|t|\le t_n} e^{-itx}\hat\psi(t)\,dt = g(x) + O_p\bigl(t_n^{1-\beta} + t_n^{\gamma-\beta+(1/2)}n^{-1/2}\bigr),$$

as claimed at (4.5).

Finally, we establish (5.19). Recall from (5.12) that $\Delta_1 = \hat\delta_2 - \hat\delta_1$, where

$$\hat\delta_1(t) = \sum_{j=0}^{\infty} 2^j\bigl\{\widehat\chi\bigl(t/2^j \mid \tfrac12,\tfrac12\bigr) - \chi\bigl(t/2^j \mid \tfrac12,\tfrac12\bigr)\bigr\}$$

and $\hat\delta_2$ has the same form except that $(\frac12,\frac12)$ is replaced by $(1,0)$. We shall derive the version of (5.19) in which Δ_1 is replaced by $\hat\delta_1$:

$$\int_{|t|\le t_n} e^{-itx}\psi(t)\hat\delta_1(t)\,dt = O_p\bigl(t_n^{\gamma-\beta+(1/2)}n^{-1/2}\bigr), \tag{5.20}$$

uniformly in x. The case of $\hat\delta_2$ is similar, although in that case the order of magnitude on the right-hand side is smaller, since the quantity $\chi(t \mid a,b)$ appearing in the denominators is of smaller order, as $|t|\to\infty$, if $(a,b) = (\frac12,\frac12)$ than it is if $(a,b) = (1,0)$.

Put $r_0 = \frac12 r(r-1)$ and let $(s_1(k), s_2(k))$, for $1\le k\le r_0$, denote an enumeration of the r_0 pairs (s_1,s_2) such that $1\le s_1 < s_2 \le r$. Recall the definitions of U_{js} and $\bar U$ in Section 5.2, and for $1\le k\le r_0$ put

$$\widehat\chi_k(t) = \frac1n\sum_{j=1}^n \exp\Bigl\{\frac12 it\bigl(U_{js_1(k)} + U_{js_2(k)}\bigr)\Bigr\}, \qquad \bar U_k = \frac{1}{2n}\sum_{j=1}^n\bigl(U_{js_1(k)} + U_{js_2(k)}\bigr).$$

Let $\widehat\chi_0 = r_0^{-1}\sum_k \widehat\chi_k$ and $\chi_0 = \chi(\cdot \mid \frac12,\frac12)$ and observe that $\bar U = r_0^{-1}\sum_k \bar U_k$, $\widehat\chi(t \mid \frac12,\frac12) = e^{-it\bar U}\widehat\chi_0(t)$ and

$$\widehat\chi\bigl(t \mid \tfrac12,\tfrac12\bigr) - \chi_0(t) = \widehat\chi_0(t) - (1+it\bar U)\chi_0(t)$$
$$+ O_p\bigl\{t^2\bar U^2 + \bigl|\widehat\chi_0(t) - (1+it\bar U)\chi_0(t)\bigr|^2\bigr\}, \tag{5.21}$$

uniformly in t. Standard moment methods, based on the linear structure of $\widehat\chi_0$ and $\bar U$, may be used to prove that

$$\sum_{j=0}^{\infty} 2^j\bigl\{(t/2^j)^2 E(\bar U^2) + E\bigl|\widehat\chi_0(t/2^j) - (1+it2^{-j}\bar U)\chi_0(t/2^j)\bigr|^2\bigr\} = O(t^2/n),$$

uniformly in t. Note, too, that $|\chi_0(t/2^j)|^{-1} = O(1 + |t|^\gamma)$ uniformly in t and in $j \geq 0$ and that if we define $\hat\delta_3 = r_0^{-1} \sum_k \hat\delta_{3k}$ where

$$\hat\delta_{3k}(t) = \sum_{j=0}^{\infty} 2^j \{\widehat\chi_k(t/2^j) - (1 + it2^{-j}\bar U_k)\chi_0(t/2^j)\}/\chi_0(t/2^j), \tag{5.22}$$

then $\hat\delta_3$ is also given by (5.22) if $\widehat\chi_k$ and $\bar U_k$ on the right-hand side are replaced by $\widehat\chi_0$ and $\bar U$, respectively. Combining the results from (5.21) down, we deduce that $\hat\delta_1(t) = \hat\delta_3(t) + O_p(|t|^{\gamma+2}/n)$ uniformly in t. Hence

$$\int_{|t|\leq t_n} e^{-itx}\psi(t)\hat\delta_1(t)\,dt = r_0^{-1}\sum_{k=1}^{r_0} S_{nk}(x) + O_p(|t_n|^{\gamma-\beta+3}/n), \tag{5.23}$$

uniformly in x, where $S_{nk}(x) = \int_{|t|\leq t_n} e^{-itx}\psi(t)\hat\delta_{3k}(t)\,dt$.

The real and imaginary parts of $S_{nk}(x)$ are both expressible as sums of n independent and identically distributed random variables with zero means, and so relatively conventional methods may be used to compute the variances of those quantities. To illustrate the argument, we treat only one of the terms that arises; it is

$$\begin{aligned} T_n(x) = \int_{|t|\leq t_n} &\cos(tx)(1+|t|)^{-\beta}\kappa_1(t) \\ &\times \Bigg[\frac{1}{n}\sum_{l=1}^{n}\sum_{j=0}^{\infty} 2^j\{\cos(tV_l/2^j) - E\cos(tV_l/2^j)\} \\ &\qquad\times (1+|t/2^j|)^\gamma \kappa_2(t/2^j)\Bigg]\,dt, \end{aligned}$$

where, here and below, κ_j denotes a real-valued function whose absolute value is uniformly bounded and $V_1, \ldots, V_n$ are independent and identically distributed as $\frac{1}{2}(U_{11} + U_{12})$. (More generally, S_{nk} is a linear form in terms like T_n.) Now,

$$\begin{aligned} &nE\{T_n(x)^2\} \\ &\quad= \iint_{|t_1|,|t_2|\leq t_n} (1+|t_1|+|t_2|+|t_1t_2|)^{-\beta}\kappa_3(x,t_1,t_2) \\ &\qquad\times\Bigg[\sum_{j_1=0}^{\infty}\sum_{j_2=0}^{\infty} 2^{j_1+j_2}\{\tfrac{1}{2}\lambda(t_1 2^{-j_1} + t_2 2^{-j_2}) \\ &\qquad\qquad + \tfrac{1}{2}\lambda(t_1 2^{-j_1} - t_2 2^{-j_2}) - \lambda(t_1/2^{j_1})\lambda(t_2/2^{j_2})\} \\ &\qquad\qquad\times (1 + |t_1/2^{j_1}| + |t_2/2^{j_2}| + |t_1/2^{j_1}\cdot t_2/2^{j_2}|)^\gamma \\ &\qquad\qquad\qquad\times \kappa_4(t_1/2^{j_1}, t_2/2^{j_2})\Bigg]\,dt_1\,dt_2, \end{aligned} \tag{5.24}$$

where $\lambda(t) = E\{\cos(tV_1)\} = (1+|t|)^{-\beta}\kappa_5(t)$. The right-hand side of (5.24) equals $O(t_n^{2(\gamma-\beta)+1})$. Combining this result with its counterpart for the other, analogous terms, we deduce (5.20) from (5.23).

5.5. *Proof of Theorem* 4.4. In the proof below we recede an arbitrarily small amount from the $h = 0$ limit, which arises when theoretically minimizing $J(\mathbf{f}, \mathbf{g}, h)$ over h as well as $\mathbf{f}$ and $\mathbf{g}$. Thus we work with an arbitrarily small but positive h.

First we prove part (a) of the theorem, addressing first the matter of whether it is possible to identify the distributions of F and G using our approach. Suppose that, along a subsequence of values of n, the distribution estimators $\widetilde{F}$ and $\widetilde{G}$ converge to subdistributions F_0 and G_0, respectively. Let ϕ_0 and ψ_0 denote the respective characteristic functions and assume that

$$\int_{-\infty<t<\infty} w(t)\,dt \int_0^1 \bigl|\phi(t)\psi(tu)\psi\{t(1-u)\} - \phi_0(t)\psi_0(tu)\psi_0\{t(1-u)\}\bigr|^2\,du = 0. \tag{5.25}$$

Then, since ϕ, ψ, ϕ_0 and ψ_0 are continuous,

$$\phi(t)\psi(tu)\psi\{t(1-u)\} = \phi_0(t)\psi_0(tu)\psi_0\{t(1-u)\} \tag{5.26}$$

for all $-\infty < t < \infty$ and $0 < u < 1$. Taking $t = 0$ in (5.25), we deduce that $\phi_0(0) = \psi_0(0) = 1$, and so the two subdistributions are actually proper distributions.

Let $\mathcal{T}$ be the (countable) set of points t such that $\phi(t/2^j)\psi(t/2^j) = 0$ for some $j \ge 0$ and let $\mathcal{T}^{\mathrm{c}}$ denote the complement of $\mathcal{T}$. If $t \in \mathcal{T}^{\mathrm{c}}$, then, replacing t by $t/2^j$ in (5.26), taking the ratio of both sides of (5.26) in the cases $u = 1$ and $u = \frac{1}{2}$, taking logarithms of both sides of the ratio of equations, multiplying by 2^j and summing from $j = 0$ to $j = k-1 \ge 1$, we deduce that

$$\log\psi(t) - 2^k \log\psi(t/2^k) = \log\psi_0(t) - 2^k \log\psi_0(t/2^k). \tag{5.27}$$

Since the distribution of ε has zero mean, $2^k\psi(t/2^k)$ converges to 0 as $k \to \infty$, and so $2^k\psi_0(t/2^k)$ must also converge. From this result and (5.27) it can be deduced first that the limit, as δ converges to 0, of $\delta^{-1}\log\psi_0(\delta)$ must exist and equal a constant, $i\mu$ say; second that $\psi_0(t) = 1 + i\mu t + o(|t|)$ as $t \to 0$; and third that $\psi(t) = \psi_0(t)\,e^{-i\mu t}$. By considering the behavior in the neighborhood of $t = 0$ we may deduce that μ is real valued and equal to the mean of the distribution with characteristic function ψ_0. But constraint (c) on the histogram density estimators, introduced as part of condition (2.13), implies that the mean of each distribution estimator is 0; and constraint (d), which imposes an upper bound on the distributions' variances, implies that the means of any limit distributions, such as F_0 and G_0, are equal to the limits of their respective means for finite values of n. Therefore the mean of the distribution with characteristic function ψ_0 is 0, and hence $\mu = 0$.

It follows that $\psi(t) = \psi_0(t)$ for each $t \in \mathcal{T}^c$ and thus, by the continuity of the characteristic functions, that $\psi \equiv \psi_0$. We may now deduce from (5.26) that $\phi \equiv \phi_0$ as well. Therefore (5.25) implies that $(\phi, \psi) \equiv (\phi_0, \psi_0)$. Call this result (R).

Suppose, by way of contradiction of part (a), that either $\widetilde{F}$ or $\widetilde{G}$ does not converge to F or G, respectively, with probability 1. Then there is an infinite subsequence $\mathcal{S}$ of values of n, and there are subprobability distributions F_0 and G_0, such that (i) either $F_0 \neq F$ or $G_0 \neq G$, and (ii) as $n \to \infty$ through values in $\mathcal{S}$ we have $\widetilde{F} \to F_0$ and $\widetilde{G} \to G_0$ with probability 1, where the convergence is "in distribution" (i.e., weak convergence). It follows that $K_j(t, u \mid \tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ converges to $K_{0j}(t, u)$ for $j = 1, 2$, where the convergence is with probability 1, $K_{01}(t, u)$ and $K_{02}(t, u)$ denote the real and imaginary parts, respectively, of $\phi_0(t)\psi_0(tu) \times \psi_0\{t(1-u)\}$, and ϕ_0 and ψ_0 are the characteristic functions of F_0 and G_0, respectively.

Standard methods show that for each t and u the empirical characteristic function $\widehat{\chi}(t \mid u, 1-u)$ converges with probability 1 to $\chi(t \mid u, 1-u)$, defined at (2.4), and in particular that, for each t and u, $\widehat{\chi}_1(t \mid u, 1-u)$ and $\widehat{\chi}_2(t \mid u, 1-u)$ converge (with probability 1), respectively, to the real and imaginary parts of the limit, which we denote by $\chi_1(t, u)$ and $\chi_2(t, u)$, respectively. Since weak convergence of distributions implies convergence of characteristic functions, since the real and imaginary parts of characteristic functions are uniformly bounded and since the weight function w is integrable, $J(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$, defined at (2.12), converges with probability 1, as $n \to \infty$ through values in $\mathcal{S}$, to

$$
\begin{aligned}
J_0 = & \int_{-\infty<t<\infty} w(t)\, dt \\
& \times \int_0^1 \{|\chi_1(t, u) - K_{01}(t, u)|^2 + |\chi_2(t, u) - K_{02}(t, u)|^2\}\, du.
\end{aligned} \tag{5.28}
$$

By construction of $\mathcal{S}$ it cannot be true that both $\phi_0 \equiv \phi$ and $\psi \equiv \psi_0$. It follows from this property and result (R) that $J_0 \neq 0$. However, we may construct deterministic histogram approximations to f and g which involve a bin width h that converges to 0 as n increases and are such that the approximations converge to f and g, respectively, and satisfy the constraint (2.13). In consequence, the distributions derived from these histograms converge to F and G, respectively, and so their respective characteristic functions converge. Taking $\mathbf{f} = \mathbf{f}_n$, $\mathbf{g} = \mathbf{g}_n$ and $h = h_n$ to be the quantities associated with these particular histograms, we see that we may construct histograms such that $J(\mathbf{f}, \mathbf{g}, h)$ converges to 0 with probability 1 as $n \to \infty$ (through the full sequence of positive integers). Hence, for an infinite number of values of n in $\mathcal{S}$, the putative minimizer of $J(\mathbf{f}, \mathbf{g}, h)$, employed in the arguments in the two previous paragraphs, does not actually produce a minimum. This contradiction proves part (a) of the theorem.

Next we turn to part (b). We may construct deterministic histogram approximations to f and g which depend on n and are arbitrarily accurate and, in particular,

which are such that their respective characteristic functions ϕ_1 and ψ_1 have the property:

$$\int_{-\infty}^{\infty} w(t)\,dt \int_0^1 \bigl|\phi(t)\psi(tu)\psi\{t(1-u)\} - \phi_1(t)\psi_1(tu)\psi_1\{t(1-u)\}\bigr|^2\,du = O(n^{-1}). \tag{5.29}$$

Next we show that

$$\int_{-\infty}^{\infty} w(t)\,dt \int_0^1 \bigl|\widehat{\chi}(t \mid u, 1-u) - \phi(t)\psi(tu)\psi\{t(1-u)\}\bigr|^2\,du = O_p(n^{-1}). \tag{5.30}$$

Note that, for each real pair (a, b), $\widehat{\chi}(t \mid a, b) = \widehat{A}_2(t \mid a, b)\exp\{-(a+b)\,it\bar{U}\}$, where $\widehat{A}_2(t \mid a, b)$ is the characteristic function of the dataset $\{aU_{js_1} + bU_{js_2}: 1 \le j \le n, 1 \le s_1, s_2 \le r, s_1 \ne s_2\}$. Observe that

$$\tfrac{1}{2}\bigl|\widehat{\chi}(t \mid u, 1-u) - \chi(t \mid u, 1-u)\bigr|^2 \le |D(t,u)|^2 + |t\bar{U}|^2,$$

where $D(t,u) = \widehat{A}_2(t \mid u, 1-u) - \chi(t \mid u, 1-u)$. Now, $E\{D(t,u)\} = 0$. Using the property noted immediately below (5.8), we may show that $E\{|D(t,u)|^2\} = O(n^{-1})$ uniformly in t and u. And since the data have finite variance, $E(\bar{U}^2) = O(n^{-1})$. Therefore $E\{|D(t,u)|^2 + |t\bar{U}|^2\} = O\{(1+t^2)n^{-1}\}$ uniformly in t, u and n. Result (5.30) is an immediate consequence.

Combining (5.29) and (5.30), we deduce that the histogram estimators that minimize $J(\mathbf{f}, \mathbf{g}, h)$, when minimization over h is included, must satisfy

$$\int_{-\infty}^{\infty} w(t)\,dt \int_0^1 \bigl|\phi(t)\psi(tu)\psi\{t(1-u)\} - \tilde{\phi}(t)\tilde{\psi}(tu)\tilde{\psi}\{t(1-u)\}\bigr|^2\,du = O_p(n^{-1}).$$

Result (b) follows from this property.

Acknowledgments. We express our gratitude to David R. Cox, who raised the problem with us and encouraged its solution; to Anthony C. Atkinson, for drawing our attention to the calcium data analyzed in Section 3.4; and to two anonymous referees, whose helpful comments enabled us to improve our paper.

REFERENCES

AIRY, G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. Macmillan, London.

BERAN, R., FEUERVERGER, A. and HALL, P. (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. *Ann. Statist.* **24** 2569–2592.

BICKEL, P. and RITOV, Y. (1987). Efficient estimation in the errors in variables model. *Ann. Statist.* **15** 513–540.

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

BOWMAN, A. W., HALL, P. and PRVAN, T. (1998). Bandwidth selection for the smoothing of distribution functions. *Biometrika* **84** 799–808.

CARROLL, R. J. and HALL, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* **83** 1184–1186.

CORNFIELD, J. and TUKEY, J. W. (1956). Average values of mean squares in factorials. *Ann. Math. Statist.* **27** 907–949.

COX, D. R. and HALL, P. (2002). Estimation in a simple random effects model with nonnormal distributions. *Biometrika* **89** 831–840.

DANIELS, H. E. (1939). The estimation of components of variance. *J. Roy. Statist. Soc. Suppl.* **6** 186–197.

EISENHART, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* **3** 1–21.

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19** 1257–1272.

HECKMAN, M. (1960). Flame photometric determination of calcium in animal feeds. *J. Assoc. Official Analytical Chemists* **43** 337–340.

JOHNSON, N., KOTZ, S. and BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions* **1**, 2nd ed. Wiley, New York.

KEMPTHORNE, O. (1975). Fixed and mixed models in the analysis of variance. *Biometrics* **31** 473–486.

KHURI, A. I. and SAHAI, H. (1985). Variance components analysis: A selective literature survey. *Internat. Statist. Rev.* **53** 279–300.

MORAN, P. A. P. (1971). Estimating structural and functional relationships. *J. Multivariate Anal.* **1** 232–255.

NELDER, J. A. (1977). A reformulation of linear models (with discussion). *J. Roy. Statist. Soc. Ser. A* **140** 48–76.

NEYMAN, J. (1951). Existence of consistent estimates of the directional parameter in a linear structural relation between two variables. *Ann. Math. Statist.* **22** 497–512.

PLACKETT, R. L. (1960). Models in the analysis of variance (with discussion). *J. Roy. Statist. Soc. Ser. B* **22** 195–217.

REIERSØL, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica* **18** 375–389.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

SAHAI, H., KHURI, A. I. and KAPADIA, C. H. (1985). A second bibliography on variance components. *Comm. Statist. Theory Methods* **14** 63–115.

SARDA, P. (1993). Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inference* **35** 65–75.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York.

SPIEGELMAN, C. (1979). On estimating the slope of a straight line when both variables are subject to error. *Ann. Statist.* **7** 201–206.

TIPPETT, L. H. C. (1931). *The Methods of Statistics.* Williams and Norgate, London.

WOLFOWITZ, J. (1952). Consistent estimators of the parameters of a linear structural relation. *Skand. Aktuarietidskr* **35** 132–151.

YATES, F. (1966). A fresh look at the basic principles of the design and analysis of experiments. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **4** 777–790. Univ. California Press, Berkeley.

CENTRE FOR MATHEMATICS AND
ITS APPLICATIONS
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA ACT 0200
AUSTRALIA
E-MAIL: Peter.Hall@maths.anu.edu.au

DEPARTMENT OF STATISTICS
LONDON SCHOOL OF ECONOMICS
HOUGHTON STREET
LONDON WC2 2AE
UNITED KINGDOM
E-MAIL: q.yao@lse.ac.uk