Methodology for nonparametric deconvolution when the error distribution is unknown

Aurore Delaigle
Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia
A.Delaigle@ms.unimelb.edu.au

Peter Hall
Department of Mathematics and Statistics, University of Melbourne, VIC, 3010, Australia
halpstat@ms.unimelb.edu.au

Abstract: In the nonparametric deconvolution problem, in order to estimate consistently a density or distribution from a sample of data contaminated by additive random noise, it is often assumed that the noise distribution is completely known or that an additional sample of replicated or validation data is available. Methods also have been suggested for estimating the scale of the error distribution, but they require somewhat restrictive smoothness assumptions on the signal distribution, which can be hard to verify in practice. In the present paper we take a completely new approach to the problem, not requiring extra data of any type. We argue that data rarely come from a simple, regular distribution, and that this can be exploited to estimate the signal distributions using a simple procedure. Our method can be extended to other problems involving errors-in-variables, such as nonparametric regression estimation. Its performance in practice is remarkably good, often equalling (even unexpectedly) the performance of techniques that use additional data to estimate the unknown error distribution.

Some Key Words: errors-in-variables, kernel smoothing, measurement errors, nonparametric density estimation.

Short title: Nonparametric deconvolution with unknown error distribution.

1 Introduction

In the nonparametric deconvolution problem we seek to estimate the distribution F_X (or the density f_X, if it exists) of a random variable X, but only observe independent and identically distributed (i.i.d.) data $W_1, \ldots, W_n$ on $W = X + U$, where U denotes a measurement error independent of X. This classical error problem has received considerable attention since the late 80s. For several decades, research in the area was performed under the assumption that the distribution F_U of the error U was known.

The case of unknown F_U has also been considered in the literature, where a common approach is to assume the availability of additional data that make it feasible to estimate F_U, for example a sample of replicated contaminated data, or, in a less commonly encountered setting, a direct sample from the error distribution. There is also interest in estimating

F_X when F_U is unknown and no additional data are available, although work in this case consists largely of theoretical results in settings where a particular parametric model for F_U is available, not general methods that can be applied broadly and enjoy particularly good performance. References are given in the second-last paragraph of this section.

Taking a completely new viewpoint, in this paper we argue that F_X can in many cases be estimated consistently without knowing even the shape of F_U, and without extra data. All we require of F_U is that it have basic properties of symmetry. This level of generality is unusual in errors-in-variables problems. To achieve it we use mathematical models for F_X that are different from those that conventionally are assumed when techniques are devised, or theory is developed, or simulation studies are designed, in deconvolution problems. In particular, we suppose that F_X is drawn from a "random universe," and for example is not symmetric.

Of course, we are not going to sample distributions repeatedly. Our statistical analysis is conducted conditional on a particular but unknown distribution which has been drawn randomly from the universe, and we invoke the random universe principle, as we refer to it below, only to motivate and model the irregularity displayed by real data distributions. We submit that this approach is often appropriate in real, practical settings, although in terms of theoretical analysis and statistical simulation it is unconventional.

Indeed, from some viewpoints our approach to tackling deconvolution problems is highly unusual, because it is the sheer irregularity and unpleasantness of a real-world F_X distribution that allows us to do exciting, unexpected things in deconvolution. If F_X were nice, symmetric and conventional, for example if it were a normal distribution, then we could not recover it from data on W without knowing the distribution of U; but if it is reasonably irregular then we can estimate it consistently. Usually, irregularity is an unpleasant feature of statistical problems, but in deconvolution, we argue, it is an asset, and we exploit it.

Perhaps surprisingly, the role played by discrete distributions in this view of deconvolution is fundamental. In particular, a simple form of the random universe principle, applied to cases where F_X is discrete, implies that that distribution is indecomposable, under basic symmetry assumptions of F_U.

These considerations lead to a new approach to inference in deconvolution problems, having two novel, distinct components. First, we use a minimum variance method to pinpoint the basic distribution that has been sampled, with noise, from a random universe. Secondly, we use discrete rather than continuous distributions as the basis for our methodology. In particular, our initial estimator of F_X is discrete, and we suggest running a smoother through it to make it continuous. The performance of the resulting method is remarkably good, often equalling, or even exceeding, techniques that use additional data to estimate the distribution of measurement error.

Fourier transform methods are used predominantly for statistical inference in errors-in-variables problems (including deconvolution). We mention here particularly the work of Carroll and Hall (1988) and Stefanski and Carroll (1990), who constructed a consistent kernel deconvolution estimator of f_X, and Stefanski (1990) and Fan (1991, 1993), who derived its general asymptotic properties. Li and Vuong (1998), Lin and Carroll (2006), Hall and Ma (2007), Delaigle et al. (2008) and Stefanski and McIntyre (2011) addressed measurement error problems in the setting of replicated data, and Diggle and Hall (1993) and Neumann (1997) considered cases where samples of error data are available. Butucea and Matias (2005), Butucea et al. (2008) and Kneip et al. (2012) discussed problems where U has a supersmooth distribution known up to a scale parameter, and Meister (2006) addressed a similar context where U is known to be normally distributed. In these papers the distribution of X is assumed to be less smooth than that of U. The ideas in these articles provide an important step towards the possibility of estimating F_X without additional data or perfect knowledge of F_U, but they nevertheless require a parametric form for F_U. Moreover, the assumption that F_X is less smooth than F_U can be hard to justify in practice. Carroll et al. (2006) provided a particularly accessible, book-length account of general methodology.

A reviewer has suggested that we survey the literature on inference for symmetric distributions, but in truth it is particularly sparse. There is, of course, interest in inference for symmetric stable laws and related distributions (e.g. Dumouchel, 1973, 1975), and other particular symmetric classes (e.g. Augustyniak and Doray, 2012), but more generally there is greater interest in distributions that are constructed deliberately to be asymmetric; see, for example, Arellano-Vallea et al. (2005).

This paper is organised as follows. We introduce our model and assumptions in section 2, suggest our new methodology in section 3, and illustrate its numerical properties on real and simulated data in section 4. We motivate the assumptions imposed in section 2 by studying the decomposability of the distribution of X in section 5. Asymptotic theoretical properties of our estimators are established in section 6. Finally, we discuss the application of our ideas to other errors-in-variables problems in section 7. Proofs are given in the appendices.

2 Models and main assumptions

2.1 Model for data, and assumptions on the error distribution

Suppose we observe values of $W_j = X_j + U_j$ for $1 \leq j \leq n$, where the pairs (X_j, U_j) are i.i.d. as (X, U), and X and U are independent. We wish to estimate the distribution of X from data on W, without knowing the distribution of U. Write ϕ_X and ϕ_U for the characteristic functions, and F_X and F_U for the corresponding distributions (or distribution functions), of X and U, respectively. We assume that

> ϕ_U is real-valued and nonnegative, but otherwise unknown; in the discrete case it vanishes at at most a countable number of points, and in the continuous case it is strictly positive on the real line. (2.1)

The first part of (2.1) is equivalent to assuming that F_U is symmetric, a standard assumption in deconvolution problems. The other conditions in (2.1) are also standard, except that in deconvolution problems, F_U (or equivalently, ϕ_U) is usually assumed known. When F_U is not known perfectly, it is typically assumed that only its scale is unknown, or that samples drawn from the error distribution, or replicates of the Ws for the same X, are available; see section 1. Our methods do not require assumptions about the shape or scale of F_U, and neither do we rely on additional samples.

2.2 Assumption on the distribution of X

Our assumption on F_X will be given at (2.3) below, but it requires a little notation. Given distributions F and G, define $F \circ G$ by

$$(F \circ G)(x) = \int F(x-u)\, dG(u)\,.$$

Abusing terminology a little, we shall call $F \circ G$ the convolution of F and G. Since X and U are independent, we have $F_W = F_X \circ F_U$.

If F_X is symmetric, then since F_U is also symmetric, with the assumption at (2.1) alone, we can not distinguish F_X from F_U knowing only F_W, and thus we cannot estimate F_X from data on W. If F_X is not symmetric, and it is possible to decompose it as

$$F_X = F_Y \circ F_Z, \tag{2.2}$$

(equivalently, $\phi_X = \phi_Y \, \phi_Z$), where F_Y and F_Z are nondegenerate distributions (and ϕ_Y and ϕ_Z their respective characteristic functions) with F_Z symmetric, then we cannot distinguish F_X from F_Y knowing only F_W, since $F_Z \circ F_U$ is symmetric and could be confused with F_U.

However in real applications, F_X can rarely be expected to be "regular". For example, although simple, classical symmetric distributions are often convenient for inference, we rarely believe that our data come perfectly from such regular distributions. Rather, since data can come from very diverse populations, F_X can be viewed as a member of a very diverse universe, and perhaps can be considered to have been drawn randomly, in some sense, from a general class $\mathcal{D}$ of distributions, according to a particular but general sampling model. There, distributions like F_Y would not be candidates for F_X, because the factorisation of ϕ_X into $\phi_Y \, \phi_Z$ imposes constraints on the structure of the universe, which with probability 1 fail to hold for a member of that universe drawn at random (see section 5).

The fact that F_X does not have a symmetric component, i.e. cannot be decomposed as the convolution at (2.2), can be exploited to suggest an estimator of F_X from data on W. Specifically, in section 3 we suggest an estimation procedure based on the phase function, which, for a random variable V, is defined by $\rho_V = \phi_V / |\phi_V|$, with ϕ_V denoting the characteristic function of V. To understand the role of the phase function in our context, note that (2.1) implies that $\phi_U = |\phi_U|$. Since $\phi_W = \phi_X \, \phi_U$, we deduce that $\rho_X = \rho_W \equiv \phi_W / |\phi_W|$, except perhaps where ϕ_U vanishes. More generally, ρ_X is also equal to the phase function of distributions of the form $F_{W^{(1)}} = F_X \circ F_{U^{(1)}}$, where $F_{U^{(1)}}$ is a symmetric distribution. Importantly, the variance of $F_{W^{(1)}}$ is strictly larger than that of F_X. Similarly, if it were possible to write $F_X = F_Y \circ F_Z$, where F_Z symmetric, then we would have $\rho_X = \rho_Y$ and the variance of Y would be smaller than that of X, but we argue (and prove in section 5) that such a decomposition of F_X is often not possible. This motivates us to assume that:

$$F_X \text{ uniquely has least variance among all distributions sharing the phase function } \rho_X. \tag{2.3}$$

We argue that the distribution F_X that is sourced from $\mathcal{D}$ in the way described above will typically satisfy this assumption. If it does not then F_X cannot be estimated from data on W alone, and in that case we suggest taking the minimum variance version of F_X to be our target.

Remark 1. *We can write $\rho_X = \exp(ip_X)$, where the function p_X is real valued and $i = \sqrt{-1}$. In signal analysis, either of the functions $\pm p_X$ is referred to as the phase function, and sometimes that terminology is applied to ρ_X. We adopt the latter definition in this paper. Note that although the phase function is not well known to statisticians, it is connected to the cumulants of a distribution; see Appendix B.11 in the supplementary file.*

To motivate further the assumption at (2.3), in section 5 we examine the decomposability of F_X as at (2.2), where F_Z is symmetric. The extent of knowledge about decomposability of a probability distribution was described 35 years ago by Loève as "somewhat disturbing," because "the ingenuity and power of the methods and the great wealth of results still leave the basic problem unsolved: Find applicable general criteria so that, given a law, one can find all its components, and, in particular, find whether it is…indecomposable" (Loève, 1978). The same could be said today; developing useful theory in this area is remarkably challenging. The interested reader is referred to Linnik and Ostrovskii (1977) and Lukacs (1970, 1983) for details of the general mathematical theory of distribution decomposability.

3 Methodology

3.1 Estimating F_X

Motivated by the discussion in section 2.2, and in particular equation (2.3), we suggest estimating F_X by the distribution with minimum variance that has phase function equal to an estimator of ρ_X constructed from the data $W_1, \ldots, W_n$. In section 3.2 we show how this estimator can be used to construct an estimator of the density f_X in the continuous case.

First, we estimate $\rho_X = \rho_W$ by $\hat{\rho}_X = \hat{\phi}_W / |\hat{\psi}|^{1/2}$, where

$$\hat{\phi}_W(t) = n^{-1} \sum_{j=1}^{n} \exp(itW_j)\,, \quad \hat{\psi}(t) = \frac{1}{n(n-1)} \sum\sum_{1 \le j,k \le n \,:\, j \ne k} \exp\{it\,(W_j - W_k)\} \qquad (3.1)$$

are root-n consistent estimators of $\phi_W(t)$ and $\psi(t) = |\phi_W(t)|^2$, respectively.

If the quality of $\hat{\rho}_X(t)$ were the same for all t, we would search for the minimum variance distribution F with phase function $\rho = \hat{\rho}_X$, or equivalently, such that $\hat{\phi}_W(t) - \bigl|\hat{\psi}(t)\bigr|^{1/2} \rho(t) = 0$ for all t. Of course, the quality of $\hat{\rho}_X(t)$ degrades as $|t|$ increases, which indicates that we should put less emphasis on larger values of $|t|$. Motivated by this, let F denote a distribution function, let ρ be its phase function, and let

$$T(F) = \int_{-\infty}^{\infty} \Bigl|\hat{\phi}_W(t) - \bigl|\hat{\psi}(t)\bigr|^{1/2} \rho(t)\Bigr|^2 w(t)\, dt\,,$$

where w is a nonnegative weight function. Recalling that our target is the distribution that has least variance, subject to having the given phase function ρ_X, we suggest estimating F_X by choosing F to minimise $T(F)$, at the same time minimising variance.

To do this we use a sieve method, and in particular we approximate F_X by a distribution determined by a finite number of parameters, allowing that number potentially to diverge with sample size. Many types of approximation are possible, including those based on a discrete distribution or an orthogonal series; our development will focus on the former. That is, we approximate F_X by a distribution supported at atoms x_j with respective probability masses p_j for $1 \leq j \leq m$. Here, we have two options: (i) both the x_js and the p_js are subject to choice, or (ii) the x_js are predetermined (here the only fitted parameters are generally the probability masses). We argue that when approximating continuous distributions by discrete ones, a method based on (i) is unnecessarily complex. In part, this is because often we are interested in f_X, not F_X, and for this we need to smooth the estimator of F_X (we shall do this below). Therefore we take option (ii).

To choose the x_js, recall that the characteristic function of a lattice distribution is always periodic, which is not the case for continuous distributions. Therefore, instead of taking the x_js to be distributed on a regular grid of m points in an interval $\mathcal{I}$, say, we suggest distributing m points uniformly but randomly in $\mathcal{I}$, and taking these points, arranged in increasing order, to be x_js. In practice we suggest taking $\mathcal{I} = [\min_i W_i, \max_i W_i]$.

Let $p = (p_1, \ldots, p_m)$ and $x = (x_1, \ldots, x_m)$, write $F(\cdot\,|\,p)$ for the discrete distribution that puts mass p_j at x_j for $1 \leq j \leq m$. The characteristic function of $F(\cdot\,|\,p)$ is given by $\phi(t\,|\,p) = \sum_j p_j \exp(itx_j)$, its phase function has the formula $\rho(t\,|\,p) =$

$\sum_j p_j \exp(itx_j)/\left|\sum_j p_j \exp(itx_j)\right|$, and its variance is equal to

$$v(p) = \sum_{j=1}^{m} p_j\, x_j^2 - \left(\sum_{j=1}^{m} p_j\, x_j\right)^2 . \tag{3.2}$$

To compute our estimator of F_X, we search for $\hat{p}_1, \ldots, \hat{p}_m$ that minimises

$$T(p) = \int_{-\infty}^{\infty} \left| \hat{\phi}_W(t) - \left|\hat{\psi}(t)\right|^{1/2} \frac{\sum_j p_j \exp(itx_j)}{\left|\sum_j p_j \exp(itx_j)\right|} \right|^2 w(t)\, dt\,, \tag{3.3}$$

under the constraint of minimising the variance and that $\hat{\phi}_U = \hat{\phi}_W(t)/\phi(\,\cdot\,|\,\hat{p})$ is symmetric and less than or equal to 1. We solve this non convex optimisation problem using Matlab's constrained optimisation program. Let $\hat{p}$ denote the value of p obtained in this way. Then $F(\,\cdot\,|\,\hat{p})$ is our discrete approximation to F_X. The practical implementation of this procedure will be discussed in section 4.1.

3.2 Estimating f_X in the continuous case

In the continuous case we often wish to estimate the density f_X. There are several ways to construct a density estimator from a discrete approximation to its distribution. We suggest using $\hat{f}_X(x) = \sum_{j=1}^{m} \hat{p}_j K_h(x - x_j)$, where K is a kernel and $h > 0$ a bandwidth; see Hall and Presnell (1999) for related tilting methods. Define $\phi(t\,|\,\hat{p}) = \sum_j \hat{p}_j \exp(itx_j)$ and let ϕ_K denote the Fourier transform of K. In this notation we can write $\hat{f}_X$ as

$$\hat{f}_X(x) = \frac{1}{2\pi} \int e^{-itx}\, \phi(t\,|\,\hat{p})\, \phi_K(ht)\, dt. \tag{3.4}$$

While this estimator works well (see Appendix B.4 in the supplementary file), the fact that $\phi(t\,|\,\hat{p})$ was constructed from values of t such that $|t| \le t^*$ implies that $\phi(t\,|\,\hat{p})$ is less reliable for $|t| > t^*$, which can affect the variability of $\hat{f}_X$. The same phenomenon arises with the estimator of Delaigle et al. (2008), which is recalled in Appendix B.2 in the supplementary file. Like there, the variability of $\hat{f}_X$ can be reduced by using a standard ridging approach where $\phi(t\,|\,\hat{p})$ is replaced by

$$\tilde{\phi}(t) = \begin{cases} \phi(t\,|\,\hat{p}) & \text{if } |t| \le t^* \\ r(t) & \text{otherwise,} \end{cases}$$

with r denoting a ridge function. This leads to the modified estimator

$$\tilde{f}_X(x) = \frac{1}{2\pi} \int e^{-itx}\, \tilde{\phi}(t)\, \phi_K(ht)\, dt. \tag{3.5}$$

As in Delaigle et al. (2008), we take $r = \hat{\phi}_W / \hat{\phi}_{U,P}$, with $\hat{\phi}_{U,P}$ the characteristic function of a Laplace distribution having variance equal to an estimator $\hat{\sigma}_U^2$ of the error variance σ_U^2 (see Appendix B.1 in the supplementary file for how to construct $\hat{\sigma}_U^2$). Since Laplace distributions are very unsmooth, this is usually a good choice; see Meister (2006) and Delaigle (2008).

4 Numerical properties

4.1 Computing the estimators in practice

Here we show how to compute our estimator in the continuous case, which is the most challenging context and the one that interests us the most. The discrete case requires only computation of the estimated probabilities $\hat{p}_j$, which can be done along the same lines.

In practice, to avoid numerical problems arising when dividing by a small number, we let $w(t) = \omega(t) \, |\sum_j p_j \exp(itx_j)|$ at (3.3), where we take $\omega(t)$ to be the Epanechnikov kernel rescaled to an interval $[-t^*, t^*]$. To choose t^*, note that the role of this quantity is to avoid using $\hat{\phi}_W(t)$ when the latter is a poor estimator of $\phi_W(t)$. Since $\hat{\phi}_W$ is unbiased for ϕ_W, and has variance less than or equal to $1/n$, then if t is such that $\hat{\phi}_W(t)$ is an order of magnitude larger than $n^{-1/2}$, we can expect $\hat{\phi}_W(t)$ to be a reasonable estimator of $\phi_W(t)$. Motivated by this, we take t^* to be the smallest $t > 0$ for which $|\hat{\phi}_W(t)| \leq n^{-1/4}$. While our recommendation is sensible, determining the optimal value of t^* requires more detailed theoretical results about our procedure than those that we currently have.

We know from the theory in section 6.3 that, when X is continuous, m should diverge to infinity as $n \to \infty$, but determining the rates at which m should increase is particularly challenging and is beyond the scope of this paper. Therefore, we chose m from numerical investigations based on various examples, with sample sizes ranging from $n = 200$ to $n = 2000$. These indicated that $m = 5\sqrt{n}$ is a reasonable choice.

Once we have the $\hat{p}_j$s, it remains to compute the density estimator using $\tilde{f}_X$ at (3.5). For this we need to choose the bandwidth h. In Appendix B.3 in the supplementary file, we show that $\tilde{f}_X$ can be viewed as an estimator of the deconvolution estimator suggested by Carroll and Hall (1988) and Stefanski and Carroll (1990) when F_U is known, and defined by

$$\hat{f}_{\mathrm{DEC}}(x) = \frac{1}{2\pi} \int e^{-itx} \{\hat{\phi}_W(t)/\phi_U(t)\} \, \phi_K(ht) \, dt \, . \tag{4.1}$$

As in Delaigle et al. (2008), this suggests taking h to be an estimator of the 2-stage plug-in bandwidth of Delaigle and Gijbels (2002, 2004), which is usually employed when computing $\hat{f}_{\mathrm{DEC}}$; see Appendix B.3 for details. As is the case for $\hat{f}_{\mathrm{DEC}}$, our estimator $\tilde{f}_X$ is not guaranteed to be positive everywhere, and so we redefine it to equal zero at places where it was otherwise negative, and renormalise so that it integrates to 1, producing the revised density estimator which, for simplicity, we shall also denote by $\tilde{f}_X$.

4.2 Real data examples

We start by illustrating the effectiveness of our approach by applying it to two real datasets where replicated measurements W_{i1} and W_{i2} are available, with

$$W_{i1} = X_i + U_{i1}, i = 1, \dots, n \text{ and } W_{i2} = X_i + U_{i2}, i = 1, \dots, R, \text{ with } R \leq n, \tag{4.2}$$

where the U_{ij}s are independent, and independent of the X_is. This permits us to compare $\tilde{f}_X$, defined in the last paragraph of section 4.1, which does not use the replicates, with methods that can only be applied when replicates are available. As we shall see, our estimator worked extremely well in these two examples, illustrating the fact that real life distributions are often sufficiently irregular for the arguments supporting our approach to hold.

Below we compare our method with Delaigle et al.'s (2008) estimator ($\hat{f}_{\mathrm{DHM}}$), which estimates F_U completely nonparametrically from the differences of replicates (see Appendix B.2 in the supplementary file), and two variants often used in practice, $\hat{f}_{\mathrm{DEC,L}}$ and $\hat{f}_{\mathrm{DEC,N}}$, where F_U in $\hat{f}_{\mathrm{DEC}}$ at (4.1) is estimated parametrically by a Laplace ($\hat{f}_{\mathrm{DEC,L}}$) or a normal ($\hat{f}_{\mathrm{DEC,N}}$) distribution whose scale $\hat{\sigma}_U$ is computed from the replicates, as described in Appendix B.2. We apply these four methods using the 2-stage plug-in bandwidth of Delaigle and Gijbels (2002, 2004), where, for each method, we replace F_U by its related estimator. We also compute the naive estimator, $\hat{f}_{\mathrm{naive}}$, which is the standard kernel estimator of f_W based on the data $W_{11}, \dots, W_{n1}$, and using the plug-in bandwidth of Sheather and Jones (1991).

Our first example comes from the Framingham study carried our by the National Heart, Lung, and Blood Institute (Kannel et al., 1986 and Carroll et al., 2006). This dataset comprises measurements, for $n = 1615$ patients, of several variables related to coronary heart disease. Our interest is in the systolic blood pressure, which is measured with a great

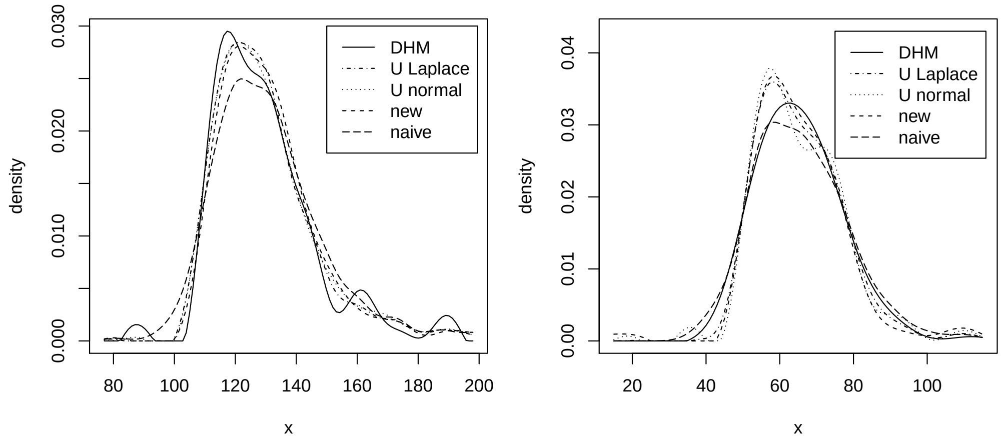

Figure 1: Estimators of the density f_X for the Framingham data (left) and for the diet and heart data (right): $\hat{f}_{\mathrm{DHM}}$ (DHM), $\hat{f}_{\mathrm{DEC,L}}$ (U Laplace), $\hat{f}_{\mathrm{DEC,N}}$ (U normal), our estimator $\tilde{f}_X$ (new) and $\hat{f}_{\mathrm{naive}}$ (naive).

deal of noise. In this study, for each i, two measurements of long term systolic blood pressure SBP_i were collected at each of two exams (exams 1 and 2). As in Carroll et al. (2006), for each i we let M_{ij} be the average of the two measurements at exam j, for $j = 1$ and 2, and take $W_{ij} = \log(50 - M_{ij})$ and $X_i = \log(50 - \mathrm{SBP}_i)$, where the W_{ij}s satisfy (4.2) with $R = n$.

The left panel of Figure 1 shows the five estimators of f_X on the same graph. We can see that our estimator, which is computed without the replicates, is very similar to $\hat{f}_{\mathrm{DHM}}$, $\hat{f}_{\mathrm{DEC,L}}$ and $\hat{f}_{\mathrm{DEC,N}}$, which cannot be computed without the replicates. In this example, our estimator even seems more attractive than $\hat{f}_{\mathrm{DHM}}$, which is perhaps too wiggly. As usual, the naive estimator, which estimates f_W rather then f_X, strongly oversmoothes the data.

Our second example comes from a pilot study on coronary heart disease; see Morris et al. (1977) and Clayton (1992). In this example we have error-prone measurements W_{i1}, $i = 1, \ldots, n$ of the ratio X_i of poly-unsaturated fat to saturated fat intake for $n = 336$ men in a one-week dietary survey. A second measurement of this ratio, W_{i2} for $i = 1, \ldots, R$, is available for $R = 60$ men, who completed a second survey several months after the first. Our goal is to estimate the density f_X of the X_is. Proceeding as in the previous example, we can compute the five estimators of f_X discussed above. The results are shown in the right panel of Figure 1. Compellingly, although our estimator does not use the replicates, it is close to $\hat{f}_{\mathrm{DEC,L}}$ and $\hat{f}_{\mathrm{DEC,N}}$, which both use the replicates. In this example, too, $\hat{f}_{\mathrm{DHM}}$ is less attractive than thoseestimators; it seems to oversmooth the data (although not as much as

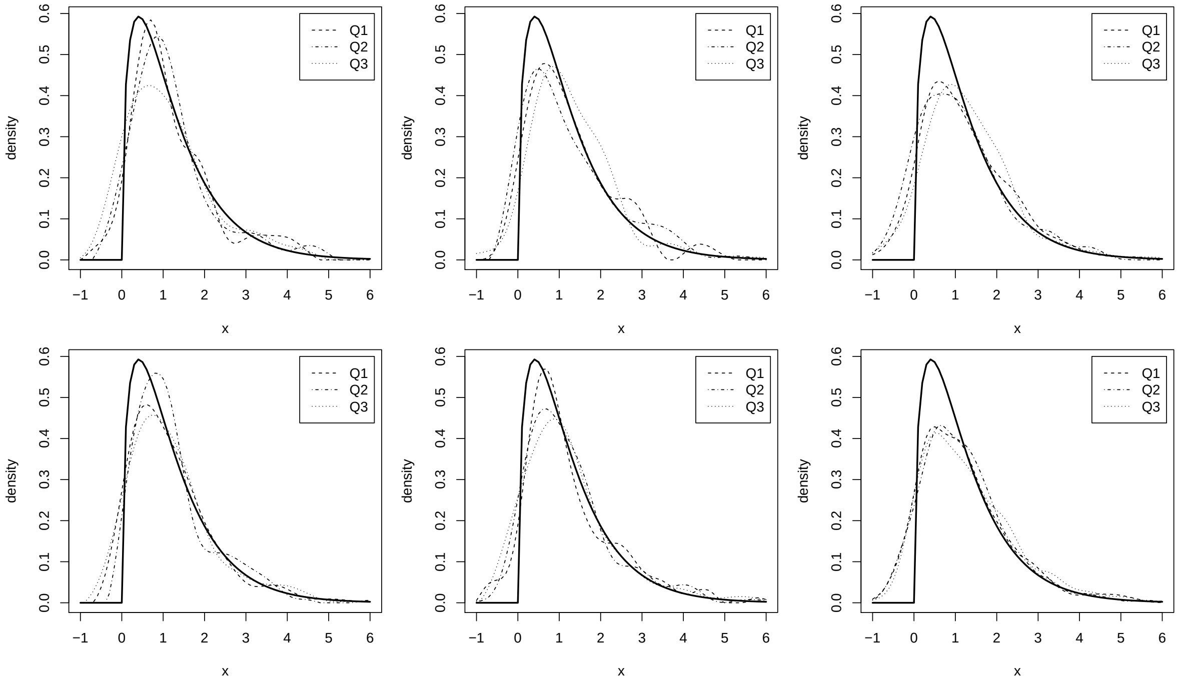

Figure 2: Curves Q1, Q2 and Q3 for model (i), when the errors are normal, NSR=20% and $n = 200$ (top) or $n = 500$ (bottom), for our estimator $\tilde{f}_X$ (left), $\hat{f}_{\text{DEC}}$ (middle) and $\hat{f}_{\text{naive}}$ (right). The true curve is depicted by the solid line.

$\hat{f}_{\text{naive}}$), probably because F_U is estimated nonparametrically from too few (60) observations.

4.3 Simulated examples

Next we consider simulated examples. Reflecting the issues discussed in section 1, we take F_X to be asymmetric and not straightforwardly decomposable as $F_X = F_Y \circ F_Z$, where F_Z is symmetric. Specifically, we generated data $W_j = X_j + U_j$, $j = 1, \ldots, n$, where U_j was normal or Laplace, and X_j came from one of the following distributions F_X:

(i) $X_j \sim \chi^2(3)/\sqrt{6}$;

(ii) $X_j \sim \{0.5\,\mathrm{N}(1,1) + 0.5\,\chi^2(5)\}/\sqrt{4.5}$.

(iii) $X_j \sim \{0.5\,\mathrm{N}(5,.6^2) + 0.5\,\chi^2(3)\}/\sqrt{4.08}$;

For each of these three distributions F_X, we generated 100 samples $W_1, \ldots, W_n$ of size $n = 200$ or $n = 500$, obtained by taking $W_j = X_j + U_j$, where $X_j \sim F_X$ and U_j was Laplace or normal, with noise to signal ratio, $\text{NSR} = \text{var}(U)/\text{var}(X)$, equal to 20% or 40%. For each of these eight configurations, and for each sample, we computed our estimator $\tilde{f}_X$ at (3.5) for unknown F_U, implemented as in section 4.1, the standard deconvolution estimator $\hat{f}_{\text{DEC}}$ at

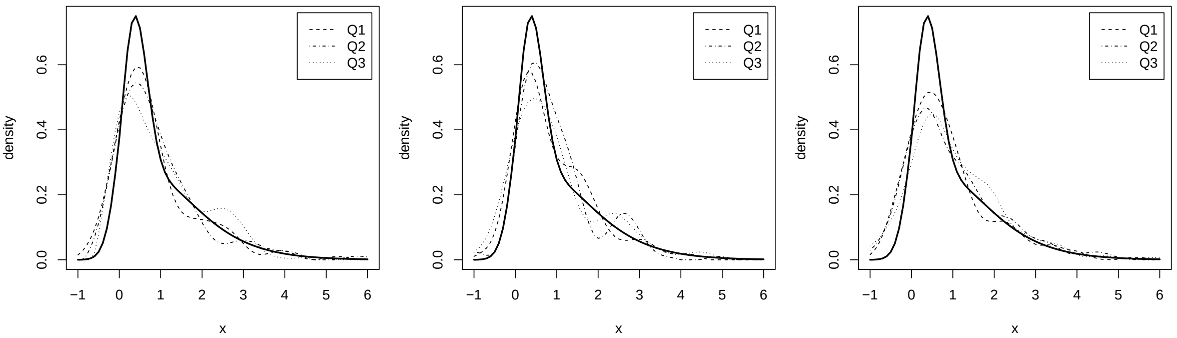

Figure 3: Curves Q1, Q2 and Q3 for model (ii), when the errors are Laplace, $n = 200$ and NSR=20%, for our estimator $\tilde{f}_X$ (left), $\hat{f}_{\text{DEC}}$ (middle) and $\hat{f}_{\text{naive}}$ (right). The true curve is depicted by the solid line.

Table 1: Simulation results for densities (i) to (iii). The numbers show 10× the median absolute deviation [1st quartile,3rd quartile] calculated from 100 simulated samples.

			U ∼Normal			U ∼Laplace		
F_X	NSR	n	$\tilde{f}_X$	$\hat{f}_{\text{DEC}}$	$\hat{f}_{\text{naive}}$	$\tilde{f}_X$	$\hat{f}_{\text{DEC}}$	$\hat{f}_{\text{naive}}$
(i)	0.4	200	3.67[3.36,4.13]	3.69[3.26,4.29]	4.17[3.84,4.58]	3.08[2.73,3.51]	3.06[2.68,3.35]	3.75[3.48,4.06]
		500	3.27[3.01,3.68]	3.31[3.00,3.68]	3.94[3.73,4.20]	2.66[2.35,2.97]	2.61[2.20,2.92]	3.46[3.20,3.75]
	0.2	200	2.88[2.51,3.15]	2.93[2.60,3.26]	3.26[2.93,3.53]	2.60[2.29,3.08]	2.62[2.29,3.09]	3.09[2.82,3.46]
		500	2.45[2.18,2.76]	2.53[2.38,2.89]	3.02[2.79,3.27]	2.13[1.92,2.40]	2.22[1.93,2.40]	2.75[2.55,2.92]
(ii)	0.4	200	3.90[3.50,4.23]	3.77[3.31,4.17]	4.16[3.80,4.43]	3.33[2.92,3.65]	3.22[2.70,3.51]	3.68[3.35,4.03]
		500	3.53[3.22,3.74]	3.49[3.08,3.70]	4.01[3.76,4.20]	2.57[2.21,2.87]	2.48[2.08,2.78]	3.36[3.14,3.54]
	0.2	200	2.92[2.61,3.36]	2.99[2.60,3.38]	3.34[3.05,3.68]	2.62[2.31,2.95]	2.53[2.29,2.96]	3.00[2.73,3.28]
		500	2.62[2.33,2.93]	2.61[2.40,2.88]	3.08[2.94,3.23]	1.98[1.65,2.39]	2.04[1.71,2.29]	2.66[2.42,2.87]
(iii)	0.4	200	5.14[4.68,5.48]	4.97[4.25,5.45]	5.47[5.19,5.77]	4.04[3.53,4.48]	3.84[3.35,4.35]	4.73[4.42,5.07]
		500	4.64[4.29,5.15]	4.40[3.92,5.02]	5.28[5.07,5.52]	3.21[2.76,3.59]	3.22[2.74,3.47]	4.59[4.31,4.72]
	0.2	200	3.93[3.52,4.38]	3.93[3.51,4.33]	4.51[4.16,4.77]	3.19[2.88,3.65]	3.28[2.95,3.70]	3.97[3.63,4.25]
		500	3.51[2.97,3.99]	3.44[3.11,3.85]	4.23[4.01,4.48]	2.50[2.28,2.91]	2.61[2.35,2.87]	3.62[3.40,3.77]

(4.1) for known F_U with the 2-stage plug-in bandwidth of Delaigle and Gijbels (2002, 2004), and the naive estimator $\hat{f}_{\text{naive}}$ (the standard kernel estimator of f_W computed from the W_is) with the Sheather and Jones (1991) plug-in bandwidth. Let $\hat{f}$ denote any of those estimators; for each estimator, we computed the integrated absolute error IAE $= \int |\hat{f} - f_X|$. In the figures we show three curves for each estimator: Q1, Q2 and Q3, which, for a given configuration, correspond to the estimated curves obtained for the samples that gave, respectively, the first, second and third quartile IAE values for that estimator in that configuration.

Figure 2 shows the curves Q1, Q2 and Q3 for the three estimators, for density (i) when the U_is are normal, NSR=20% and increasing n. We can see that our estimator performed very well compared to the standard estimator $\hat{f}_{\text{DEC}}$ that uses the correct error distribution, and much better than the naive estimator $\hat{f}_{\text{naive}}$. As expected, the results improve as n increases.

In Figure 3 we show Q1, Q2 and Q3 for density (ii), Laplace U_is, $n = 200$ and NSR=20% (see Figure B.1 in Appendix B.4 in the supplementary file for NSR=40%). Here too we see that our estimator fares particularly well compared to $\hat{f}_{\text{DEC}}$, whereas $\hat{f}_{\text{naive}}$ performs more poorly. We obtained similar results in other cases, as can be seen from Table 1, where we present the median, first and third quartiles of the IAE for all the cases we considered.

5 Indecomposability of F_X

5.1 Introduction

Owing to the challenges of decomposability, discussed briefly at the end of section 2, we do not have all the theory that we would prefer, and so we provide only partial results. To make the concept of a random universe more concrete, we start by constructing an explicit example of a discrete random universe; see section 5.2. Then, in sections 5.3 and 5.4, we show that, in a wide range of settings where F_X comes from a discrete random universe, F_X cannot be decomposed as at (2.2) where F_Y and F_Z are two nondegenerate discrete distributions, and F_Z is symmetric. We treat the continuous case in section 5.5.

Note that, if F_X is discrete, then cases where F_Z is continuous are not relevant, since the methodology we suggested in section 3 fits only discrete distributions to the phase function, and is not confounded by potential continuous components in a decomposition of the distribution of Z that will, on account of the assumed positivity of their characteristic functions, cancel from a formula for ρ_X. This is also true for the "generalised characteristic function" ϕ, introduced below; it too does not require consideration in cases where (a component of) the generalised distribution of Z has a continuous distribution. Therefore, to examine the existence of a decomposition

$$F_X = F_Y \circ F_Z$$

when F_X is assumed to be discrete, it is sufficient to consider instances where F_Z is discrete.

Our nondecomposability results imply that we cannot write $\phi_X = \phi_Y \, \phi_Z$, where ϕ_Y is the characteristic function of F_Y and ϕ_Z is the real-valued characteristic function of a symmetric distribution F_Z. Furthermore, our theorems can be extended, either directly or with mild extra assumptions, to the case where ϕ_Z is replaced by a "generalised characteristic function"

ϕ, i.e. a real-valued function, not necessarily a characteristic function but nevertheless a Fourier transform of a discrete, symmetric but potentially signed "probability" distribution. In such instances we can write $\phi(t) = \sum_{1 \le j \le m_Z} r_j \exp(it\, z_j)$, where the sequences of values z_j and r_j are each symmetric, but where it is not necessarily true that each r_j is positive. Further discussion is given below Theorems 1, 2 and 4.

5.2 An example of model for sampling F_X from a random universe

Recall from section 1 that we introduce random sampling from large sets of distributions only to motivate and model the irregularity of real-world populations. In our statistical analysis we condition on the randomly sampled distribution, and in particular, sampling from a random universe is not a feature of our methodology.

Let $\mathcal{D} = \mathcal{D}_{\text{disc}}$ denote the class of all discrete distributions having two or more distinct atoms. In the example considered here, we first we choose randomly the number, M, of atoms of F_X. We take $M \ge 2$ to be a random variable having a discrete distribution of which the support is the set $\{2, 3, \ldots\}$. Next, we choose randomly the M atoms of the distribution F_X. To do this, conditional on M, we let $V_1, V_2, \ldots$ denote independent continuous random variables (for example, normal random variables). Then, we choose randomly the M probability masses of the distribution F_X. For this, put $V = (V_1, \ldots, V_M)$, and, conditional on M and V, let $P = (P_1, \ldots, P_M)$ be chosen randomly from the simplex defined by $P_j \ge 0$ for $1 \le j \le M$ and $P_1 + \ldots + P_M = 1$. Finally, denote by $\mathcal{F}$ the sigma-field generated by M, V and P, and, conditional on $\mathcal{F}$, let X have the discrete distribution F_X that places mass P_j at V_j for $1 \le j \le M$.

This sampling model prescribes the joint distribution of (M, V, P). The model associates, with each choice of (M, V, P), a discrete distribution, $F(\,\cdot\,|\,M, V, P)$ say, in $\mathcal{D}_{\text{disc}}$, the distribution being defined by $P(X = x\,|\,M, V, P) = F(x\,|\,M, V, P)$, for all real numbers x, where X is as in the previous paragraph. Each distribution in $\mathcal{D}_{\text{disc}}$ is in the support of the class of all distributions $F(\,\cdot\,|\,m, v, p)$, in the sense that if $F_X \in \mathcal{D}_{\text{disc}}$ then, for each $\epsilon > 0$, $P\big[d\big\{F_X, F(\,\cdot\,|\,M, V, P)\big\} \le \epsilon\big] > 0$, where $d(F, G)$ can be any conventional measure of distance between discrete distributions F and G, for example a measure based on the Lévy metric for distribution functions.

The construction of F_X described above is just an example, and our theoretical analysis in sections 5.3 and 5.4 will show that the definition given above of the random universe $\mathcal{D}_{\text{disc}}$ is actually more elaborate than necessary. It is sufficient to take M to be fixed, and either fix P and choose V randomly in the continuum, or fix V and choose P randomly. Likewise, it is not essential that P be independent of M and V.

5.3 Discrete distributions with irregularly spaced atoms

In this section we prove that the decomposition at (2.2) is not possible for discrete distributions F_X with irregularly spaced atoms. More precisely, Theorem 1 below states that if F_X is drawn from a random universe of discrete distributions where the atoms of F_X are irregularly spaced, then with probability 1 with respect to the operation of sampling from that universe, the distribution of X cannot be expressed as that of $Y + Z$, where the random variables Y and Z are independent, nondegenerate and have discrete distributions. Although the case where Z is symmetric is particularly important when working with the phase function, we note that Theorem 1 is true regardless of whether or not the distribution of Z is symmetric.

Theorem 1. *Assume that the atoms of the discrete distributions, in the random universe from which we draw F_X, are irregularly spaced in the sense that the joint distribution of any finite number of the atoms is continuous. Then, with probability 1 with respect to drawing F_X from the random universe, that distribution cannot be written as a convolution of two discrete distributions.*

See Appendix A.1 for a proof. Theorem 1 is derived under the assumption that F_Z is a proper probability distribution. In particular, it satisfies $\sum_{j=1}^{m_Z} r_j = 1$, where $r_j = P(Z = z_j) \geq 0$ are the atoms of the discrete distribution of Z. However, the proof of the theorem also is correct when $\phi_X = \phi_Y \, \phi$ and ϕ is a generalised characteristic function as in section 5.1, provided we add a mild extra assumption; see Appendix B.6 in the supplementary file.

5.4 Discrete distributions with regularly spaced atoms

Next we consider cases where the randomness of the universe from which F_X is drawn comes from values of the probabilities $p_j = P(X = x_j)$, and not from arrangements of the atoms x_j. In particular, here drawing F_X from a random universe involves drawing the probability masses of the atoms of F_X randomly, and in the continuum, from the simplex $\mathcal{S}$ of values $p_1, \ldots, p_{m_X}$ such that each $p_j > 0$ and $p_1 + \ldots + p_{m_X} = 1$. (Here m_X denotes the number of atoms of F_X.) However, in that random universe the atoms need not be irregularly arranged; for example they could be constrained to lie on a regular grid.

The next theorem shows that, if the distribution F_X comes from this random universe, then it is not possible to decompose F_X as at (2.2). More precisely, Theorem 2 demonstrates that, with probability 1 with respect to sampling from the random universe, F_X cannot be expressed as the distribution of $Y + Z$, where Y and Z are independent, nondegenerate and have discrete distributions, and Z is symmetric.

Theorem 2. *Assume that F_X is chosen at random from a random universe of discrete distributions, each having $m_X \geq 3$ atoms constrained to lie at particular consecutive values (the same for each F_X) on a regular grid, and for which the respective probability masses are chosen at random and uniformly from $\mathcal{S}$. Then, with probability 1 with respect to the operation of sampling from the random universe, nondegenerate discrete distributions F_Y and F_Z cannot be found, supported on consecutive values on the regular grid and such that F_Z is symmetric about some point and $F_X = F_Y \circ F_Z$.*

As for Theorem 1, Theorem 2 is derived and proved under the assumption that F_Z is a proper probability distribution, but our proof is valid in cases where Z is interpreted as having a generalised discrete distribution. We conclude this section with a result showing that Theorem 2 does not generally hold when F_Z is not constrained to be symmetric. See Appendix A.2 for a proof and Appendix B.7 in the supplementary file for an example.

Theorem 3. *Assume that F_X is chosen at random from a random universe of discrete distributions, each having $m_X \geq 3$ atoms constrained to lie at particular consecutive values (the same for each F_X) on a regular grid, and for which the respective probability masses are chosen at random and uniformly from $\mathcal{S}$. Then there is strictly positive probability that nondegenerate discrete distributions F_Y and F_Z can be found, each supported on consecutive values on the regular grid, such that $F_X = F_Y \circ F_Z$.*

5.5 Continuous distributions

Next we study the decomposability of F_X in the continuous case, which is much more complex. The main difficulty is that, while there exist many results in the probability literature about decomposability of distributions, these do not permit checking whether or not a given distribution is decomposable, let alone decomposable into a convolution where one of the components is symmetric. As noted by Lukacs (1970), "There is no general method for finding the prime factors of a given characteristic function; our knowledge consists mostly of interesting special examples." One property that is known (see Parthasarathy et al., 1962; Loève, 1978) is that the set of indecomposable, absolutely continuous distributions is G_δ dense in the set of absolutely continuous distributions; those authors wrote that "This lends substance to the statement that, in general, a distribution is indecomposable."

Thus, the set of absolutely continuous distributions that cannot be decomposed into $F_X = F_Y \circ F_Z$, where F_Z is symmetric, is vast. Next we give an example in the context of finite-parameter models. Let $m \geq 3$ be an integer, let $x_1, \ldots, x_m$ be fixed, regularly spaced numbers, and denote by $P = (P_1, \ldots, P_m)$ a multivariate probability distribution drawn randomly and uniformly from the simplex $\mathcal{S}$ defined in section 5.4, with m_X there replaced by m. Let $U = (U_1, \ldots, U_m)$ be a vector of independent random variables, with each U_j distributed uniformly on $[c_1, c_2]$, where $0 < c_1 < c_2 < \infty$. It is assumed too that U is independent of P. Define $F_X(\,\cdot\,|\,h, P, U)$ to be the continuous distribution that has density

$$f_X(x\,|\,h, P, U) = \sum_{j=1}^{m} \frac{P_j}{hU_j}\, K\Big(\frac{x - x_j}{hU_j}\Big)\,, \tag{5.1}$$

where $h > 0$ and K is the density of a symmetric, compactly supported probability distribution G that cannot be decomposed as $G = G_1 \circ G_2$ for nondegenerate G_1 and G_2.

As h decreases to zero in the definition of $F_X(\,\cdot\,|\,h, P, U)$, that distribution converges to the discrete distribution $F_X(\,\cdot\,|\,P)$ that has an atom of mass P_j at x_j for $1 \leq j \leq m$. In view of Theorem 2, with probability 1 with respect to the operation of sampling P from the simplex, $F_X(\,\cdot\,|\,P)$ is not decomposable as $F_Y \circ F_Z$, where F_Z is symmetric.

The property of "indecomposability with probability 1" is also true for the distribution $F_X(\,\cdot\,|\,h, P, U)$, if h is sufficiently small and positive; see Theorem 4, below. Of course, in this setting the indecomposability property involves drawing (P, U), rather than simply P, from the joint distribution prescribed two paragraphs above. Theorem 4 exploits the fact that the U_js are chosen randomly and independently, and would not necessarily hold in more restrictive circumstances. For example, if the U_js were all equal to U^0, say, then we could write $F_X(\,\cdot\,|\,h, P, U) = F_Y \circ F_Z$ where $F_Y = F_X(\,\cdot\,|\,P)$ and F_Z is the continuous, symmetric distribution with density $f_Z(z) = (hU^0)^{-1} K(z/hU^0)$. See Appendix B.8 in the supplementary file for a proof of the theorem.

Theorem 4. *If $F_X(\,\cdot\,|\,h, P, U)$ is as defined three paragraphs above, if the continuous, symmetric, compactly supported distribution with density K is not decomposable, and if h is sufficiently small, then, with probability 1 with respect to the operation of sampling from the random universe, nondegenerate distributions F_Y and F_Z cannot be found such that F_Y is a distribution, F_Z is symmetric and $F_X = F_Y \circ F_Z$.*

As for Theorems 1 and 2, although this result is proved under the assumption that F_Z is a proper probability distribution, with a mild extra assumption it can be extended to the case where $\phi_X = \phi_Y\, \phi$ and ϕ is a generalised characteristic function as in section 5.1.

6 Large-sample properties

6.1 Asymptotic theory for the discrete case

Here we establish consistency and rates of convergence of our estimators of F_X when that distribution is discrete with a finite number of atoms. In this setting, let m^0, $p^0 = (p^0_1, \ldots, p^0_{m^0})$ and $x^0 = (x^0_1, \ldots, x^0_{m^0})$ denote the true values of m, of $p = (p_1, \ldots, p_m)$ and of $x = (x_1, \ldots, x_m)$, respectively. We make the following assumptions, where constants denoted by $C_\cdot$ or $c_\cdot$ are respectively large or small:

(a) $|\phi_W(t)| > c_1\,(1 + |t|)^{-C_1}$ for all t, where $c_1, C_1 > 0$; (b) the function w in (3.3) is continuous, and $0 < w(t) \le C_2 \exp(-c_2\,|t|)$ for all t, where $C_2, c_2 > 0$; (6.1)

(a) the true F_X is discrete and has exactly $m^0 \geq 2$ atoms, at the points $x_1^0, \ldots, x_{m^0}^0$ and with respective probability masses $p_1^0, \ldots, p_{m^0}^0 > 0$, and ϕ_X satisfies $|\phi_X(t)| > c_3\,(1 + |t|)^{-C_3}$ for all t, where $c_3 > 0$ and $C_3 \geq 0$; (b) any other distribution F_Y, say, with the same phase function as F_X, has strictly greater variance than F_X, and, if $m = m^0$ is held fixed, the minimum variance for F_X is achieved in the usual quadratic way as $p \to p^0$ and $x \to x^0$; (c) when estimating m^0, x^0 and p^0 we know the value of a constant $C_4 > 0$ such that $m^0 \leq C_4$ and $\max_j |x_j^0| \leq C_4$, and consequently we confine attention to m satisfying $m \leq C_4$ and x_j, satisfying $|x_j| \leq C_4$. (6.2)

It follows from (6.2)(c) that the variance of X is bounded above by C_4^2, and hence that the minimum variance criterion is well defined.

The assumption in (6.2)(a) that $|\phi_X(t)| > c_3\,(1 + |t|)^{-C_3}$ for all t often holds for discrete distributions, having either regularly or irregularly spaced atoms. Indeed, the assumption is often valid with $C_3 = 0$. To appreciate why, suppose F_X is the discrete distribution with mass π_j at ξ_j for $1 \leq j \leq m$, and write $\pi_k = \pi_{\max}$ for the largest π_j. Then

$$|\phi_X(t)| = \bigl|\phi_X(t)\,\exp(-it\,\xi_k)\bigr| = \left|\sum_{j=1}^{m} \pi_j\,\exp\{it\,(\xi_j - \xi_k)\}\right| \geq \pi_{\max} - \sum_{j\,:\,j\neq k} \pi_j = 2\,\pi_{\max} - 1\,.$$

Therefore, if $\pi_{\max} > \frac{1}{2}$, and we choose $c_3 \in (0, 2\,\pi_{\max} - 1)$, then $|\phi_X(t)| > c_3$ for all t.

Next we define estimators $\hat{m}$, $\hat{p}$ and $\hat{x}$ of m^0, p^0 and x^0, respectively. In a theoretical account of the discrete case it is feasible to minimise T, at (3.3), over m and x as well as p. Therefore it is appropriate to write $T(p)$ and $v(p)$, at (3.3) and (3.2), as $T(m,p,x)$ and $v(m,p,x)$, respectively. Noting (6.2)(c), for each integer $m \in [1, C_4]$ we choose $(m, \hat{p}_{[m]}, \hat{x}_{[m]})$ to minimise $T(m,p,x)$; we define $(m_j, \hat{p}_{[m_j]}, \hat{x}_{[m_j]})$, for $1 \leq j \leq J$, say, to be the local minima of the sequence $\{v(m, \hat{p}_{[m]}, x_{[m]}),\ 1 \leq m \leq C_4\}$; and we take $(\hat{m}, \hat{p}, \hat{x})$ to be the value of $(m_j, \hat{p}_{[m_j]}, \hat{x}_{[m_j]})$ that minimises $v(m_j, \hat{p}_{[m_j]}, \hat{x}_{[m_j]})$ over $1 \leq j \leq J$. Defining $\hat{p}_k$ and $\hat{x}_k$, for $1 \leq k \leq \hat{m}$, to be the components of $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_{\hat{m}})$ and $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_{\hat{m}})$, we take $\widehat{F}_X$ to be the distribution function that places mass $\hat{p}_j$ at $\hat{x}_j$ for $1 \leq j \leq \hat{m}$.

Theorem 5. *If* (6.1) *and* (6.2) *hold, and if* $(\hat{m}, \hat{p}, \hat{x})$ *is defined as suggested above, then* $P(\hat{m} = m^0) \to 1$ *and* $(\hat{p}, \hat{x}) = (p^0, x^0) + O_p(n^{-1/2})$.

A proof of the theorem is given in Appendix A.3. To better interpret the theorem, let L denote the Lévy metric, both here and in Theorem 7 below. Then Theorem 5 implies that, under the assumptions stated there, $L(\widehat{F}_X, F_X^0) = O_p(n^{-1/2})$. Theorem 7 will establish consistency of $\widehat{F}_X$ for F_X^0 in a general setting.

6.2 Minimising both $T(p)$ and $v(p)$ in a general setting

This section is a prelude to section 6.3, where we establish consistency of our estimator of F_X in a context that includes both continuous and discrete cases. Recall the definitions of $T(p)$ and $v(p)$ at (3.3) and (3.2), and put

$$T_0(p) = \int_{-\infty}^{\infty} \left| \phi_W(t) - |\phi_W(t)| \, \frac{\sum_j p_j \exp(itx_j)}{|\sum_j p_j \exp(itx_j)|} \right|^2 w(t) \, dt \, . \tag{6.3}$$

As in section 3.1 we have in mind cases where both m and the grid (either regular or irregular) on which the components x_j of x lie either do not depend on the data, or are determined from the data in a rudimentary way. In the continuous case, m would diverge with n, and the grid represented by x would become infinitesimally fine and expand to fill the support of the distribution of W.

Let $\hat{t}$ denote any sequence of positive random variables with the property that

$$1 - \pi \equiv P\Big\{ \sup_p |T(p) - T_0(p)| \leq \hat{t} \Big\} \to 1 \, , \quad \hat{t} \to 0 \tag{6.4}$$

as $n \to \infty$, where the second convergence is in probability, and, here and below, $\sup_p$ denotes the supremum over m-point probability distributions defined on the grid. For example, regardless of choice of m and x, (6.4) holds if we take $\hat{t} = c_n$ where c_n denotes any sequence of positive constants satisfying $c_n \to 0$ and $n^{1/2} \, c_n \to \infty$, and if the weight function w, in (3.3) and (6.3), is integrable. (This follows from the fact that $n^{1/2} \, (\hat{\phi}_W - \phi_W)$ and $n^{1/2} \, (\hat{\psi} - \psi)$ both converge weakly to Gaussian processes.) Alternatively we could take $\hat{t}$ to be a bootstrap approximation to the upper π-level quantile of the distribution of $\sup_p |T(p) - T_0(p)|$, and let π decrease slowly to zero as n increases.

Recall from section 3.1 that $F(\,\cdot\,|\,p)$ denotes the distribution that places mass p_j on x_j for $1 \leq j \leq m$. We shall view $T_0(p)$ as a measure of the distance of the distribution $F(\,\cdot\,|\,p)$ from the class $\mathcal{D}_X$ of all distributions with phase function ρ_X. (This interpretation is appropriate because $T_0 = 0$ if and only if the ratio $\sum_j p_j \exp(itx_j)/|\sum_j p_j \exp(itx_j)|$, in the definition at (6.3), is replaced by a function that equals ρ_X almost everywhere.) Given a positive random variable δ_n converging to 0, let $\mathcal{P}(\delta_n)$ denote the set of all p (of length m) such that $T_0(p) \leq \delta_n$. We interpret $\{F(\,\cdot\,|\,p) : p \in \mathcal{P}(\delta_n)\}$ as the class of distributions that are representable on our grid and are no further than δ_n from $\mathcal{D}_X$. In practice we can often obtain an appropriate value for δ_n by combining an assumption about the smoothness of F_X with knowledge of the fineness of the grid.

Put $\widehat{\mathcal{P}}(\delta_n) = \{p : T(p) \leq \hat{t} + \delta_n\}$. In section 6.3 we shall take $\hat{p}$, the value of p that we use to construct our estimator of F_X, to be the minimiser of $v(p)$ over $p \in \widehat{\mathcal{P}}(\delta_n)$. (With probability at least $1 - \pi$, $\widehat{\mathcal{P}}(\delta_n)$ contains any value of p that minimises $T_0(p)$ over $\mathcal{P}(\delta_n)$. Therefore, since $\mathcal{P}(\delta_n)$ is nonempty whenever $\delta_n > 0$, the probability that $\widehat{\mathcal{P}}(\delta_n)$ is nonempty converges to 1.) Theorem 6, below, shows that $T_0(p)$ converges to 0 uniformly in $p \in \widehat{\mathcal{P}}(\delta_n)$, and that with probability converging to 1, $\widehat{\mathcal{P}}(\delta_n)$ contains $\mathcal{P}(\delta_n)$.

Theorem 6. *If* (6.4) *holds, and $\delta_n \to 0$ as $n \to \infty$, then*

$$\sup_{p \in \widehat{\mathcal{P}}(\delta_n)} T_0(p) \xrightarrow{P} 0\,, \quad P\bigl\{\mathcal{P}(\delta_n) \subseteq \widehat{\mathcal{P}}(\delta_n)\bigr\} \to 1\,. \tag{6.5}$$

6.3 General consistency of $\widehat{F}_X$, and consistency of $\hat{f}_X$ in the continuous case

Here we establish consistency of our estimator $\hat{F}_X$ in a general context that encompasses discrete and continuous distributions. In the continuous case we also prove consistency of the estimator $\hat{f}_X$ of section 3.2. Our setting is that of section 6.2, and we adopt the notation there. Additionally we constrain $\widehat{F}_X$ to satisfy $\int x^4\, d\widehat{F}_X(x) \leq C\, n^{-1} \sum_{j=1}^n W_j^4$, for any fixed constant $C \geq 1$. This condition is justified by the fact that $E(W^4) = E(X^4) + 6\, E(X^2)\, E(U^2) + E(U^4)$, and requires only minor adjustments to m and x. The constraint ensures that our estimator of F_X does not have tails that are unduly heavy.

Define $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m) = \operatorname{argmin}\{v(p) : p \in \widehat{\mathcal{P}}(\delta_n)\}$, and let $\widehat{F}_X$ denote the discrete distribution with atoms of mass $\hat{p}_j$ at the respective components x_j of the vector $x = (x_1, \ldots, x_m)$. We take this version of $\widehat{F}_X$ to be our estimator of F_X, and again write L for the Lévy metric. Recall from section 6.2 that $\mathcal{P}(\delta_n) = \{p : T_0(p) \leq \delta_n\}$. It follows from the definition of T_0 at (6.3) that, if the grid is sufficiently extensive and sufficiently fine, then for each n there exists $p^0 = p^0(n)$ such that $F(\cdot \,|\, p^0) \to F_X^0$ as $n \to \infty$, where F_X^0 is the version of F_X that has minimum variance subject to having phase function ρ_X. We assume that the fineness of the grid decreases sufficiently quickly (i.e. m increases sufficiently fast), or δ_n decreases to zero sufficiently slowly, to ensure that $\mathcal{P}(\delta_n)$ contains, for sufficiently large n, such a probability vector p^0:

$$\text{there exists } p^0 = p^0(n) \text{ such that } T(p^0) \le \delta_n \text{ for all sufficiently large } n, \text{ and } F(\,\cdot\,|\,p^0) \to F_X^0. \tag{6.6}$$

The next theorem establishes consistency of our estimator $\hat{F}_X$. See Appendix B.9 in the supplementary file for a proof.

Theorem 7. *If $E(W^4) < \infty$ and there exists a unique distribution F_X^0 with minimum variance over all distributions F_X for which $\phi_X = |\phi_X|\,\rho_X$ almost everywhere on the real line, and if* (6.1), (6.4) *and* (6.6) *hold, then $L(\widehat{F}_X, F_X^0) \to 0$ in probability.*

Finally, we state a general theorem which, in the context of Theorem 7 and the additional assumption that F_X^0 is smooth, ensures that a simple kernel smooth of $\widehat{F}_X$, like the one in section 3.2, produces a uniformly consistent estimator of the density $f_X^0 = F_X^0{}'$. Of course, the quantities $\widehat{F}_X$ and F_X^0 mentioned in Theorem 8 are not required to be those introduced earlier in this paper. The proof is given in Appendix B.10 in the supplementary file.

Theorem 8. *Let $\widehat{F}_X$ be an estimator of a distribution function F_X^0, and assume that $L\big(\widehat{F}_X, F_X^0\big) \xrightarrow{P} 0$ as $n \to \infty$. Suppose too that F_X^0 is absolutely continuous with density f_X^0, which is bounded and uniformly continuous, and define $\hat{f}_X(x) = h^{-1} \int K\{(x-u)/h\}\, d\widehat{F}_X(u)$, where the nonnegative kernel K is of bounded variation and satisfies $\int K = 1$, and $h > 0$ represents a bandwidth. Then, if $h = h(n)$ decreases to 0 sufficiently slowly as n increases,*

$$\sup_x \big|\hat{f}_X(x) - f_X^0(x)\big| \xrightarrow{P} 0\,. \tag{6.7}$$

It follows from an integration by parts argument, and formula (B.13) in Appendix B.10 and the fact that K is of bounded variation, that the probability that $\hat{f}_X$ is well defined and uniformly bounded converges to 1 as $n \to \infty$.

7 Application to other problems

There are many problems involving measurement errors where knowing, or having an accurate estimator of, the error distribution is critical. Examples include nonparametric errors-in-variables regression, cure rate models (Ma and Yin, 2008) and variance estimation (Wang at al. 2009). Our method can be used in those cases too, since once we have an estimator of the distribution of U, we can use it to replace F_U. See Appendix B.5 in the supplementary file for details about the errors-in-variables regression problem.

A Appendix

A.1 Proof of Theorem 1

Recall that if F and G are distributions then $F \circ G$ denotes their convolution. We derive Theorem 1 by contradiction. If Theorem 1 does not hold, then

> F_X is a discrete probability distribution with mass $p_j > 0$ at respective distinct points x_j, for $1 \leq j \leq m_X$, and F_X can be written as the convolution $F_Y \circ F_Z$ of discrete probability distributions F_Y and F_Z, where F_Y has mass $q_j > 0$ at distinct points y_j for $1 \leq j \leq m_Y$ (with $m_Y \geq 2$), and F_Z has mass $r_j > 0$ at distinct points z_j for $1 \leq j \leq m_Z$ (with $m_Z \geq 2$). (A.1)

(We permit m_X, m_Y and m_Z to be either finite or infinite. Of course, if one of m_Y and m_Z is infinite then m_X must be too.) We claim that (A.1) entails:

> There exist atoms x_1, x_2, x_3 and x_4 of F_X, at least three of them distinct, such that $x_1 < x_2 < x_4$, $x_1 < x_3 < x_4$ and $x_2 - x_1 = x_4 - x_3$. (A.2)

If the m_X atoms of F_X are chosen randomly, in such a manner that each finite subset of them has a continuous joint distribution, then the probability of (A.2) holding equals 0. This proves Theorem 1.

It remains to show that (A.2) follows from (A.1). To this end, note that the atoms x of F_X all have the form $x = y + z$, where y and z are atoms of F_Y and F_Z, respectively. The assumptions in (A.1) imply the existence of atoms $y_1 < y_2$ and $z_1 < z_2$ of F_Y and F_Z, respectively. Then (A.2) holds with $x_1 = y_1 + z_1$, $x_2 = y_1 + z_2$, $x_3 = y_2 + z_1$ and $x_4 = y_2 + z_2$.

A.2 Proofs of Theorems 2 and 3

Let the atoms of the distributions F_X, F_Y and F_Z be x_j, y_j and z_j, respectively, for $1 \leq j \leq m_X$, $1 \leq j \leq m_Y$ and $1 \leq j \leq m_Z$, where each of m_X, m_Y and m_Z is a finite integer strictly greater than 1. The atoms are assumed to be defined on the same regular grid, which without loss of generality is the set of integers. Define $p_j = P(X = x_j)$, $q_j = P(Y = y_j)$ and $r_j = P(Z = z_j)$. Then the atoms of F_X are the points $y_j + z_k$, with respective probabilities $q_j r_k$, although typically not all the atoms listed in this way are distinct. Indeed, if $y_1, \ldots, y_{m_Y}$ and $z_1, \ldots, z_{m_Z}$ are each sets of consecutive integers, say $0, \ldots, m_Y - 1$ and $1, \ldots, m_Z$, respectively, then $m_X = m_Y + m_Z - 1$, the atoms of F_X are the points $j = 1, \ldots, m_Y + m_Z - 1$, and

$$p_j = \sum_{i=\max(0,\, j-m_Z)}^{\min(m_Y-1,\, j-1)} q_i\, r_{j-i}\,, \quad 1 \le j \le m_Y + m_Z - 2\,, \tag{A.3}$$

where we have omitted the last equation in (A.3), corresponding to $j = m_X = m_Y + m_Z - 1$, because the p_js are constrained to add to 1.

Assume that $p_1, \ldots, p_{m_X} - 1$ are known; we wish to determine q_js and r_js such that $F_X = F_Y \circ F_Z$. Since $q_1 + \ldots + q_{m_Y} = r_1 + \ldots + r_{m_Z} = 1$, then there are exactly $m_Y + m_Z - 2$ unknowns q_js and r_js to be determined, and these are given by the $m_Y + m_Z - 2$ equations (A.3). In particular, in the context of Theorem 3, the number of unknowns is the same as the number of equations. It follows that in some but not all instances, if the values $p_1, \ldots, p_{m_X}$ are chosen at random on the simplex of values for which each $p_j \ge 0$ and $p_1 + \ldots + p_{m_X} = 1$, then nondegenerate distributions of Y and Z can be selected such that $F_X = F_Y \circ F_Z$. Examples will be discussed in Appendix B.7 in the supplementary file.

On the other hand, in the context of Theorem 2 the distribution F_Z is symmetric, entailing $r_1 = r_{m_Z}$, $r_2 = r_{m_Z - 1}$, and so on. In particular, the number of unknown values of r_j is now reduced from m_Z to $\frac{1}{2} m_Z$ if m_Z is even, and to $\frac{1}{2}(m_Z + 1)$ if m_Z is odd. The number of unknowns q_j and r_j is now strictly less than the number of equations in (A.3), and so, since the p_js on the left-hand sides of those equations are chosen randomly in the continuum, then with probability 1 (with respect to the operation of drawing $(p_1, \ldots, p_{m_X})$ from the m_X-variate simplex), probability distributions $(q_1, \ldots, q_{m_Y})$ and $(r_1, \ldots, r_{m_Z})$ (the latter symmetric), satisfying equations (A.3), cannot be found.

In cases where the atoms of F_X are not at consecutive points of a grid, the number of equations still exceeds the number of unknowns, and so, since the p_is are sampled randomly and in the continuum, there is zero probability of finding F_Z such that $F_X = F_Y \circ F_Z$.

A.3 Proof of Theorem 5

Step 1. Simplified formula for $T(m, p, x)$. The simplification is given at (A.6) below. Define $\Delta_1(t) = n^{-1} \sum_j \{\exp(itW_j) - \phi_W(t)\}$, and note that in this notation, $\hat{\phi}_W = \phi_W + \Delta_1$.

Observe too that

$$\hat{\psi}(t) = \frac{1}{n(n-1)} \sum\sum_{1\le j,k\le n\,:\,j\ne k} \exp\{it\,(W_j - W_k)\} = |\phi_W(t)|^2 + \Delta_2(t) + O_p\big(n^{d_1-1}\big)\,,$$

uniformly in $|t| \le n^{D_1}$ for any $d_1, D_1 > 0$, where

$$\Delta_2(t) = \frac{1}{n} \sum_{j=1}^{n} \Big[\big\{\exp(itW_j) - \phi_W(t)\big\}\,\bar{\phi}_W(t) + \big\{\exp(-itW_j) - \bar{\phi}_W(t)\big\}\,\phi_W(t)\Big]\,.$$

Furthermore,

$$\max_{j=1,2} |\Delta_j(t)| = O_p\big(n^{d_2-(1/2)}\big)\,, \tag{A.4}$$

uniformly in $|t| \le n^{D_2}$, for any $d_2, D_2 > 0$. Therefore, in view of (6.1)(a), given $d_3 > 0$, if $D_2 = D_2(d_3) > 0$ is sufficiently small then

$$\begin{aligned}\big|\hat{\psi}(t)\big|^{1/2} &= |\phi_W(t)|\,\Big|1 + |\phi_W(t)|^{-2}\,\big\{\Delta_2(t) + O_p\big(n^{d_2-1}\big)\big\}\Big|^{1/2}\\ &= |\phi_W(t)|\,\Big\{1 + \tfrac{1}{2}\,|\phi_W(t)|^{-2}\,\Delta_2(t) + O_p\big(n^{d_3-1}\big)\Big\}\,,\end{aligned}$$

uniformly in t such that $|t| \le n^{D_2}$. Hence, recalling the definition of $T(m,p,x)$ at (3.3), we deduce that since $w(t)$ decreases at least as fast as $\exp(-c_2\,|t|)$ as $|t|$ diverges (see (6.1)(b)), then for any $D_3 > 0$,

$$\begin{aligned}T(m,p,x) = \int_{-n^{D_2}}^{n^{D_2}} \Bigg|\phi_W(t) &+ \Delta_1(t) - \Big\{|\phi_W(t)| + \tfrac{1}{2}\,|\phi_W(t)|^{-1}\,\Delta_2(t)\\ &+ O_p\big(n^{d_3-1}\big)\Big\}\,\frac{\sum_j p_j\,\exp(itx_j)}{\big|\sum_j p_j\,\exp(itx_j)\big|}\Bigg|^2\,w(t)\,dt + O_p\big(n^{-D_3}\big)\\ = \int_{-n^{D_2}}^{n^{D_2}} \Bigg|\phi_W(t) &+ \Delta_1(t) - \big\{|\phi_W(t)| + \tfrac{1}{2}\,\Delta_3(t)\big\}\,\rho(t\,|\,m,p,x)\\ &+ O_p\big(n^{d_3-1}\big)\Bigg|^2\,w(t)\,dt + O_p\big(n^{-D_3}\big)\,,\end{aligned} \tag{A.5}$$

where, defining $\alpha(t) = \phi_W(t)/|\phi_W(t)|$, we put

$$\Delta_3(t) = \frac{1}{n}\sum_{j=1}^{n} \Big[\bar{\alpha}(t)\,\big\{\exp(itW_j) - \phi_W(t)\big\} + \alpha(t)\,\big\{\exp(-itW_j) - \bar{\phi}_W(t)\big\}\Big]\,;$$

the ratio $\rho(t\,|\,m,p,x) = \sum_j p_j\,\exp(itx_j)\big/\big|\sum_j p_j\,\exp(itx_j)\big|$ is interpreted as 1 if the denominator equals 0; and the remainders are of the stated orders uniformly in all choices of m in the range $1 \le m \le C_4$ (see (6.2)(c)), and of the p_js and x_js, and, in the case of the remainder $O_p(n^{d_3-1})$, also uniformly in t such that $|t| \le n^{D_2}$.

Bearing in mind the exponential rate of decrease of the weight function w (see (6.1)(b)), we can deduce from (A.5) that if $d_3, D_3 > 0$ are given then $D_4 = D_4(d_3, D_3) > 0$ can be found, sufficiently large, such that

$$\begin{aligned} T(m,p,x) = \int_{-D_4 \log n}^{D_4 \log n} \Big| \phi_W(t) - |\phi_W(t)|\, \rho(t \,|\, m,p,x) + \Delta_1(t) \\ - \tfrac{1}{2}\, \Delta_3(t)\, \rho(t \,|\, m,p,x) + O_p\big(n^{d_3-1}\big) \Big|^2 w(t)\, dt + O_p\big(n^{-D_3}\big)\,, \end{aligned} \tag{A.6}$$

where the same interpretation of the remainders applies.

Step 2. Simple consistency. Under assumption (6.1), properties (A.4) and (A.5) imply that

$$T(m,p,x) = \int_{-\infty}^{\infty} \Big| \phi_W(t) - |\phi_W(t)|\, \rho(t \,|\, m,p,x) \Big|^2 w(t)\, dt + o_p(1)\,, \tag{A.7}$$

uniformly in (m,p,x). Define $(\hat m, \hat p, \hat x)$ as in the paragraph preceding Theorem 5. If $n(1), n(2), \ldots$ is any infinite sequence of values of n, then a further infinite subsequence $n_1, n_2, \ldots$ can be chosen such that (a) the random variable $\hat m$ has a proper limiting distribution, with atoms confined to the positive integers not exceeding the integer part of C_4; and, (b) conditional on $\hat m = m$ say, the random vector $(\hat p, \hat x)$ has a proper limiting distribution, depending on m. Using this result, an argument by contradiction, employing properties (6.2)(a), (6.2)(b) and (A.7), shows that:

the probability that $\hat m = m^0$ converges to 1 as $n \to \infty$, and, conditional on $\hat m = m^0$, $(\hat p, \hat x)$ converges in probability to (p^0, x^0). (A.8)

Step 3. Root-n consistency. In view of (A.8) there exists a positive sequence of real numbers ϵ_n, decreasing to 0, such that $n^{1/2}\, \epsilon_n \to \infty$ and $P\big\{ \big\| (\hat p, \hat x) - \big(p^0, x^0\big) \big\| \le \epsilon_n \big\} \to 1$. By (A.6), for (p,x) satisfying $\|(p,x) - (p^0, x^0)\| \le \epsilon_n$, we have

$$\begin{aligned} T\big(m^0,p,x\big) = \int_{-D_4 \log n}^{D_4 \log n} \Big| \phi_W(t) - |\phi_W(t)|\, \rho\big(t \mid m^0,p,x\big) + \Delta_1(t) \\ - \tfrac{1}{2}\, \Delta_3(t)\, \rho\big(t \mid m^0,p^0,x^0\big) + n^{-1/2}\, \Delta_4(t \,|\, p,x) \Big|^2 w(t)\, dt + O_p\big(n^{-D_3}\big)\,, \end{aligned} \tag{A.9}$$

where the stochastic process Δ_4 satisfies $\sup_{(p,x)\,:\, \|(p,x)-(p^0,x^0)\| \le \epsilon_n} |\Delta_4(t \,|\, p,x)| \xrightarrow{P} 0$. From this point standard arguments, which involve minimising $T(m^0,p,x)$ in (A.9) over (p,x) satisfying

$\|(p,x)-(p^0,x^0)\| \le \epsilon_n$, show that the minimum is attained at a value of (p,x) that is within $O_p(n^{-1/2})$ of (p^0,x^0). Hence, by (A.8), $(\hat{p},\hat{x}) = (p^0,x^0)+O_p(n^{-1/2})$, conditionally on $\hat{m}=m^0$. Equivalently, since $P(\hat{m}=m^0) \to 1$, we have $(\hat{p},\hat{x}) = (p^0,x^0)+O_p(n^{-1/2})$ in the standard, unconditional sense.

A.4 Proof of Theorem 6

Observe that $T_0(p) \le T(p) + \hat{t} \le 2\,\hat{t} + \delta_n$, where, in view of (6.4), the first inequality holds simultaneously for all p, with probability $1-\pi$, and the second inequality holds simultaneously for all $p \in \widehat{\mathcal{P}}(\delta_n)$, with probability 1. The first part of (6.5) is therefore a consequence of the second part of (6.4) and the assumption that $\delta_n \to 0$.

Note that $p \in \mathcal{P}(\delta_n)$ is equivalent to $T_0(p) \le \delta_n$. The assertions at (6.4) imply that with probability at least $1-\pi$, whenever $p \in \mathcal{P}(\delta_n)$ it holds too that $T(p) \le \hat{t} + \delta_n$; that is, $p \in \widehat{\mathcal{P}}(\delta_n)$. Therefore $P\{\mathcal{P}(\delta_n) \subseteq \widehat{\mathcal{P}}(\delta_n)\} \ge 1-\pi$, and so the second part of (6.5) follows from the first part of (6.4).

Acknowledgement

We thank Ethan Anderes and Jiashun Jin for helpful discussion, and referees and an associate editor for their useful suggestions. Delaigle and Hall were supported by the Australian Research Council. Hall also acknowledges support from the National Science Foundation.

References

Arellano-Vallea, R.B., Gómezb, H.W. and Quintanaa, F.A. (2005). Statistical inference for a general class of asymmetric distributions. *J. Statist. Plann. Inf.* **128**, 427–443.

Augustyniak. M. and Doray, L.G. (2012). Inference for a leptokurtic symmetric family of distributions represented by the difference of two gamma variates. *J. Statistic. Comput. Simul.* **82**, 1621–1634.

Butucea, C. and Matias, C. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, **11**, 309-340.

Butucea, C., Matias, C. and Pouet, P. (2008). Adaptivity in convolution models with partially known noise distribution. *Electron. J. Statist.*, **2**, 897–915.

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

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models*, 2nd Edn. Chapman and Hall CRC Press, Boca Raton.

Clayton, D.G. (1992). Models for the analysis of cohort and case control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health* (edss Dwyer, Feinleib, Lippert and Hoffmeister), 301–331, Oxford University Press, New York.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statist. Sinica*, **18**, 1025–1045

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *J. Roy. Statist. Soc., Ser. B*, **64**, 869–886.

Delaigle, A. and Gijbels, I. (2004). Comparison of data-driven bandwidth selection procedures in deconvolution kernel density estimation. *Comp. Statist. Data Anal.*, **45**, 249–267.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.*, **103**, 280–287.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, **36**, 665–685

Diggle, P. and Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *J. Roy. Statist. Soc.* Series B, **55**, 523–531.

Dumouchel, W.H. (1973). Stable distributions in statistical inference: 1. Symmetric stable distributions compared to other symmetric long-tailed distributions. *J. Amer. Statist. Assoc.* **68**, 469–477.

Dumouchel, W.H. (1975). Stable distributions in statistical inference: 2. Inference from stably distributed samples. *J. Amer. Statist. Assoc.* **70**, 386–393.

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

Fan, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.*, **21**, 600-610.

Hall, P. and Presnell, B. (1999). Density estimation under constraints. *J. Comput. Graph. Statist.*, **8**, 259–277.

Hall, P. and Ma, Y. (2007). Semiparametric estimators of functional measurement error models with unknown error. *J. Roy. Statist. Soc.* Series B, **69**, 429-446.

Kneip, A., Simar, L. and Van Keilegom, I. (2012). Boundary estimation in the presence of measurement error with unknown variance. *J. Econometrics* (conditionally accepted).

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivariate Anal.*, **65**, 139–165.

Lin, X. and Carroll, R.J. (2006). Semiparametric estimation in general repeated measures problems. *J. Roy. Statist. Soc.* Series B, **68**, 69-88.

Linnik, J.V. and Ostrovskii, I.V. (1977). *Decomposition of Random Variables and Vectors.* Translations of Mathematical Monographs **48**, American Mathematical Society, Providence, R.I.

Loève, M. (1978). Review of Linnik and Ostrovskii (1977). *Bull. Amer. Math. Soc.*, **84**, 638–642.

Lukacs, E. (1970). *Characteristic Functions.* Charles Griffin, London; Hafner Publishing Company, New York.

Lukacs, E. (1983). *Developments in Characteristic Function Theory.* Macmillan, New York.

Ma, Y. and Yin, G. (2008). Cure rate model with mismeasured covariates under transformation. *J. Amer. Statist. Assoc.*, **103**, 743-756.

Meister, A. (2006). Density estimation with normal measurement error with unknown variance. *Statist. Sinica*, **16**, 195–211.

Morris, J.N., Marr, J.W. and Clayton, D.G. (1977). Diet and heart: a postscript. *British Med. J.*, **2**, 1307–1314.

Neumann, M.H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametric Statist.*, **7**, 307–330.

Parthasarathy, K.R., Rao, R.R. and Varadhan, S.R.S. (1962). On the category of indecomposable distributions on topological groups. *Trans. Amer. Soc.*, **102**, 200-217.

Sheather, S.J. and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.* Series B, **53**, 683-690.

Stefanski, L.A. (1990). Rates of convergence of some estimators in a class of deconvolution problems. *Statist. Probab. Lett.* **9**, 229-235.

Stefanski, L.A and Carroll, R.J., (1990). Deconvoluting kernel density estimators. *Statistics*, **21**, 169–184.

Stefanski, L.A. and McIntyre, J. (2011). Density estimation with replicate heteroscedastic measurements. *Ann Inst Stat Math.*, **63**, 81-99.

Wang, Y., Ma, Y. and Carroll, R.J. (2009). Variance estimation in the analysis of microarray data. *J. Roy. Statist. Soc.* Series B, **71**, 425-445.