Aurore Delaigle, Ingrid Van Keilegom

Deconvolution with unknown error distribution

1

Deconvolution with unknown error distribution

CONTENTS

"nobreak

1.1 Introduction

In the nonparametric literature on the classical measurement error problem, where the goal is to do inference on the distribution of a variable X observed with independent additive error U, it is often assumed that the error density f_U is perfectly known. In applications this is often too restrictive, since in many cases one does not know much about the complexity and imperfect nature of the underlying data. In this chapter we present approaches for non- and semiparametric inference on the distribution of X developed in papers that do not assume the density f_U to be (fully) known. Unless otherwise specified, we assume throughout that all the variables are continuous and the data follow the classical additive measurement error model (see (1.1) below) or a heteroscedastic variant of it (see model (1.15)). Apart from some qualitative or parametric assumptions on f_U, we do not use other models. In particular,

we do not consider the case where the variable X is a covariate in a regression model, that one wishes to estimate by taking the measurement error on X into account.

There are basically two ways to avoid assuming that f_U is known: either we observe additional data (in the form of repeated contaminated measurements, longitudinal data, validation data, auxiliary variables, instrumental variables, etc.) that can be used to estimate f_U, or we make assumptions on the laws of X and U that allow to identify and estimate f_U without the need to collect additional data. In Section 1.2 we show how to estimate f_X nonparametrically when additional data are available; the case without additional observations is the topic of Section 1.3. In Section 1.4 we discuss the related challenging problem of estimating the boundary of the support of X in the case without additional data. Finally, in Section 1.5, we consider the case where one has neither enough data nor enough information about the underlying densities to be able to estimate f_U consistently; we discuss how nonparametric estimators of f_X are impacted by misspecification of f_U.

1.2 Estimation of f_X when additional data are available

In this section we discuss techniques for estimating the density f_X nonparametrically from data coming from the classical measurement error model, in the case where the error density is unknown but it can be estimated from additional observations. For example, we may have access to a sample from the error distribution or may be able to measure the data repeatedly. There are other situations where some type of additional data are available, which can be used for identifying and constructing an estimator of f_U and f_X. They include the case when validation data are available (i.e. when some of the data points are measured without error), when panel or longitudinal data are observed, or when auxiliary variables or instrumental variables are available. In each of these cases it is possible to identify the distribution of the signal X under certain assumptions.

In this section we consider three different scenarios. In Section 1.2.1, we discuss the simplest case where we observe a direct sample from the errors, which are assumed to be homoscedastic. Section 1.2.2 introduces techniques that can be applied in the more common situation where the contaminated observations are observed repeatedly and the errors are homoscedastic. We extend those technique to the more general heteroscedastic case in Section 1.2.3.

1.2.1 Sample from f_U in the classical error model

Suppose we observe an i.i.d. sample of contaminated data $X_1^*, \ldots, X_n^*$ distributed like a random variable X^* coming from the classical error model

$$X^* = X + U, \; X \text{ and } U \text{ independent}, \; X \sim f_X \text{ and } U \sim f_U \,. \tag{1.1}$$

As usual in the deconvolution literature, we assume that

$$f_U \text{ is symmetric and } \varphi_U(t) \neq 0 \text{ for all } t \in \mathbb{R} \,, \tag{1.2}$$

where here and throughout, $\varphi_A(t)$ denotes the characteristic function of A if A is a random variable, or the Fourier transform of A if A is a function. (Note that since f_U is symmetric, φ_U is real and thus $\varphi_U(t) \neq 0$ is equivalent to $\varphi_U(t) > 0$.)

In the basic setting where f_U is known, a popular nonparametric estimator of f_X is the deconvolution kernel estimator (Carroll and Hall, 1988; Stefanski and Carroll, 1990). See also Apanasovich and Liang (2021) for more details regarding kernel deconvolution methods for regression with classical errors, and Kang and Qiu (2021) for other deconvolution approaches. For all $x \in \mathbb{R}$, it is defined by

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} K_U\Big(\frac{x - X_j^*}{h}\Big) \,, \tag{1.3}$$

where $h > 0$ is a bandwidth and, letting K be a kernel function,

$$K_U(x) = \frac{1}{2\pi} \int e^{-itx} \frac{\varphi_K(t)}{\varphi_U(t/h)} \, dt \,, \tag{1.4}$$

where i denotes the complex number such that $i^2 = -1$; see Delaigle (2021).

When f_U is unknown, we cannot compute the estimator at (1.3). In Diggle and Hall (1993), the authors considered the case where, in addition to the X_i^*'s, an i.i.d. sample $\widetilde{U}_1, \ldots, \widetilde{U}_m$ from the error density f_U is observed. There, they suggested to replace the unknown $\varphi_U(t)$ at (1.4) by the empirical characteristic function $\hat{\varphi}_U(t) = m^{-1} \sum_{j=1}^{m} e^{it\widetilde{U}_j}$ to obtain

$$\hat{K}_U(x) = \frac{1}{2\pi} \int e^{-itx} \frac{\varphi_K(t)}{\hat{\varphi}_U(t/h)} \, dt \,, \tag{1.5}$$

and they proposed to estimate $f_X(x)$ by

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} \hat{K}_U\Big(\frac{x - X_j^*}{h}\Big) \,. \tag{1.6}$$

Under certain regularity conditions, they showed that as long as $n = O(m)$, estimating φ_U has no first order asymptotic impact on the integrated squared error of the estimator at (1.3). However, in practice $\hat{\varphi}_U(t)$ is a poor estimator

of $\varphi_U(t)$ for large $|t|$ (i.e. where $\varphi_U(t)$ takes small values). Indeed, since $\hat{\varphi}_U$ is unbiased and has a variance of order m^{-1}, then values of $|\hat{\varphi}_U|$ much smaller than $m^{-1/2}$ are essentially noise and can create numerical instability of the integral at (1.6). These difficulties can be avoided by regularising $\hat{\varphi}_U$ in its tails, using for example the estimator $\tilde{f}_X(x) = (nh)^{-1} \sum_{j=1}^{n} \tilde{K}_U\{(x - X_j^*)/h\}$ suggested by Neumann (1997), where

$$\tilde{K}_U(x) = \frac{1}{2\pi} \int e^{-itx} \frac{\varphi_K(t)}{\hat{\varphi}_U(t/h)} 1\{|\hat{\varphi}_U(t/h)| \geq m^{-1/2}\} \, dt \,. \tag{1.7}$$

In practice, h needs to be chosen from the data; Neumann (1997) briefly mentions the possibility of using cross-validation of the type as in Stefanski and Carroll (1990), but without giving any detail. Perhaps this could be done by replacing, in each term of the cross-validation criterion used by Stefanski and Carroll (1990), φ_U by $\hat{\varphi}_U$, using as well $1\{|\hat{\varphi}_U(t/h)| \geq m^{-1/2}\}$ as above, but this approach is studied neither in theory nor in practice.

1.2.2 Replicated measurements with homoscedatisc errors

Observing a sample from f_U, like in Section 1.2.1, is only possible in some particular applications. Instead, what is often more realistic is to assume that we can observe replicated contaminated measurements

$$X_{jk}^* = X_j + U_{jk}, \text{ for } k = 1, \ldots, m_j \text{ and } j = 1, \ldots, n \,, \tag{1.8}$$

where $m_j \geq 1$ is the number of replicates available for the jth individual, at least a fraction of the m_j's are greater or equal to 2, $X_j \sim f_X$, $U_{jk} \sim f_U$ and the X_j's and the U_{jk}'s are totally independent.

For example, in nutrition studies, the long term nutrient intake X of a patient is often measured through several 24 hour recalls. At each 24 hour recall, the jth patient reports their food intake over the last 24 hours, which is only a rough approximation to their long term intake X_j. After a transformation (typically the log transform), the data from the 24 hour recalls are often assumed to behave like X_{jk}^* at (1.8).

To understand why replicates can be used to make progress when f_U is unknown, note that if we subtract two replicates, say X_{j1}^* and X_{j2}^*, for a given individual, then we have $X_{j1}^* - X_{j2}^* = U_{j1} - U_{j2}$. Since the U_{jk}'s are i.i.d., this can be exploited to estimate quantities related to f_U. In the simplest version of this problem, a parametric model is assumed for f_U. Often, the errors are assumed to be of zero mean, and the only unknown in f_U is the variance σ_U^2, which can be estimated by

$$\hat{\sigma}_U^2 = \sum_{j=1}^{n} \sum_{k=1}^{m_j} (X_{jk}^* - \bar{X}_j^*)^2 / \sum_{j=1}^{n} (m_j - 1). \tag{1.9}$$

This approach is commonly used in the parametric literature on measurement

errors. See for example Madansky (1959), Carroll and Stefanski (1990), Gleser (1990), Carroll et al. (1993) and Carroll et al. (2006), page 71. It can be combined with either a parametric estimator of f_X or with the deconvolution estimator at (1.3), replacing there all instances of φ_U by the corresponding parametric estimator.

However, if we assume that (1.2) holds, the distribution of U can be estimated consistently without such parametric assumptions, as we show now. Under (1.2), since the U_{jk}'s are i.i.d., we have

$$X^*_{j1} - X^*_{j2} = U_{j1} - U_{j2} \sim f_U * f_U \,,$$

where, for two functions f and g, $f * g(x) = \int f(x-u)g(u)\,du$ denotes the convolution of f and g. Therefore, the characteristic function of $X^*_{j1} - X^*_{j2}$ is equal to φ_U^2 and we can estimate φ_U^2 by the empirical characteristic function of all the possible differences $X^*_{jk} - X^*_{j\ell}$, $k \neq \ell$,

$$\hat{\varphi}_U^2(t) = \frac{1}{N} \sum_j \sum_{k<\ell} e^{it(X^*_{jk} - X^*_{j\ell})} = \frac{1}{N} \sum_j \sum_{k<\ell} e^{it(U_{jk} - U_{j\ell})} \,, \tag{1.10}$$

where $\sum_j$ refers to the sum over the individuals for which $m_j \geq 2$, and N denotes the number of terms in the triple sum $\sum_j \sum_{k<\ell}$. See Delaigle et al. (2007) and Delaigle et al. (2008).

Under (1.2), for all $t \in \mathbb{R}$ we can write $\varphi_U(t) = \{\varphi_U^2(t)\}^{1/2}$. Since $\hat{\varphi}_U^2$ is an empirical characteristic function, it is not guaranteed to be real nor positive, unlike $\varphi_U^2(t)$. However, these two properties can be enforced by replacing $\hat{\varphi}_U^2(t)$ by $\big|\Re\{\hat{\varphi}_U^2(t)\}\big|$, where $\Re$ denotes the real part (the imaginary part is of no interest in practice, and ignoring it makes no asymptotic difference since it vanishes asymptotically). Motivated by these arguments and noting that, for all $x \in \mathbb{R}$ we have $\Re(e^{ix}) = \cos x$, Delaigle et al. (2008) proposed to estimate $\varphi_U(t)$ by

$$\hat{\varphi}_U(t) = \big|\Re\{\hat{\varphi}_U^2(t)\}\big|^{1/2} = \left| \frac{1}{N} \sum_j \sum_{k<\ell} \cos\{t(X^*_{jk} - X^*_{j\ell})\} \right|^{1/2} , \tag{1.11}$$

where N and the sums are as at (1.10). Then, they estimate $f_X(x)$ by

$$\hat{f}_X(x) = \frac{1}{nh} \sum_{j=1}^{n} \hat{K}_U\Big(\frac{x - X^*_j}{h}\Big), \tag{1.12}$$

with

$$\hat{K}_U(x) = \frac{1}{2\pi} \int e^{-itx} \frac{\varphi_K(t)}{\hat{\varphi}_U(t/h) + \rho}\, dt, \tag{1.13}$$

where $\rho > 0$ is a ridge parameter (usually a small number) introduced to avoid problems with the denominator getting too close to zero. Under sufficient regularity conditions, Delaigle et al. (2008) show that, if $n = O(N)$ and the

distribution of X is sufficiently smooth compared to that of U, $\hat{f}_X$ at (1.12) has the same first order asymptotic properties as $\hat{f}_X$ at (1.3).

In practice it is difficult to choose an appropriate ρ from the data, and other techniques can be employed to reduce the impact of the unreliability of $\hat{\varphi}_U$ in its tails. For example, one could use Neumann's (2007) truncation approach introduced in Section 1.2.1. Delaigle et al.'s (2008) alternative suggestion is to replace $\hat{\varphi}_U(t) + \rho$ by

$$\tilde{\varphi}_U(t) = \hat{\varphi}_U(t)\, 1\{t \in A\} + \hat{\varphi}_{U,P}(t)\, 1\{t \notin A\}, \tag{1.14}$$

where A denotes an interval on which $\hat{\varphi}_U$ is not too unreliable, and $\hat{\varphi}_{U,P}$ denotes a parametric estimator of φ_U. For example, A could be taken to be the truncation interval used by Neumann (2007), or the largest interval around 0 on which the estimator $\hat{\varphi}_U(t)$ does not oscillate. For the parametric model $\varphi_{U,P}$, arguments from Meister (2004) and Delaigle (2008) suggest that a good choice is to use a Laplace distribution. Delaigle et al. (2008) follow this approach, and take $\hat{\varphi}_{U,P}(t) = (1 + \hat{\sigma}_P^2 t^2)^{-1}$, where $\hat{\sigma}_P^2$ is half the empirical variance of U computed as at (1.9). Delaigle (2008) suggest using the plug-in method bandwidth of Delaigle and Gijbels (2002, 2004), replacing there each occurrence of φ_U by $\tilde{\varphi}_U$.

Li and Vuong (1998) suggested a more complex approach which, in principle, could be used regardless of symmetry properties of the error density f_U. Their technique is based on properties of the characteristic function of pairs of replicates, but it suffers from at least two drawbacks. First, it is not clear how it could be implemented in practice. Second, the authors established consistency of their estimator under the assumption that the density and the characteristic function of U and X are both compactly supported. As noted by Delaigle et al. (2008), it seems difficult to satisfy those conditions simultaneously.

1.2.3 Replicated measurements with heteroscedastic errors

The estimator at (1.12) introduced in Section 1.2.2 can be extended to the setting where the U_{jk}'s at (1.8) are not identically distributed. In practice, heteroscedastic errors arise when individuals are not observed under similar conditions. For example, the observations may have been collected in several laboratories or from different studies, and groups of individuals (e.g. smoker/non smoker, male/female) may be contaminated by different types of errors. There are several ways to model heteroscedasticity of the errors. One possibility is to assume that instead of coming from the model at (1.8), the replicated contaminated observations are generated from the model

$$X_{jk}^* = X_j + U_{jk},\ X_j \sim f_X,\ U_{jk} \sim f_{U_j}, \text{ for } k = 1, \ldots, m_j \text{ and } j = 1, \ldots, n, \tag{1.15}$$

where $m_j > 1$ is the number of replicates available for the jth individual, the X_j's and the U_{jk}'s are totally independent and, letting $n_j = m_j(m_j-1)/2$,

$$f_{U_j} \text{ is symmetric for } j = 1, \ldots, n \text{ and } \sum_{j=1}^{n} n_j |\varphi_{U_j}(t)|^2 > 0 \text{ for all } t \in \mathbb{R}\,, \tag{1.16}$$

where φ_{U_j} denote the characteristic function of U_{jk}, for $k = 1, \ldots, m$. Note that unlike the homoscedastic case from Section 1.2.2, the measurements need to be replicated at least once ($m_j > 1$) for all individuals.

Under these conditions, Delaigle and Meister (2008) proposed an alternative to (1.12) that can be computed from the data at (1.15). To understand their technique, note that using (1.16), since $(X^*_{jk} + X^*_{j\ell})/2 = X_j + (U_{jk} + U_{j\ell})/2$, we have

$$E\Big(\sum_{j=1}^{n}\sum_{k<\ell} e^{it\frac{X^*_{jk}+X^*_{j\ell}}{2}}\Big) = \varphi_X(t)\sum_{j=1}^{n} n_j|\varphi_{U_j}(t/2)|^2\,.$$

Likewise, since $(X^*_{jk} - X^*_{j\ell})/2 = (U_{jk} - U_{j\ell})/2$, we have

$$E\Big(\sum_{j=1}^{n}\sum_{k<\ell} e^{it\frac{X^*_{jk}-X^*_{j\ell}}{2}}\Big) = \sum_{j=1}^{n} n_j|\varphi_{U_j}(t/2)|^2\,.$$

These calculations suggest estimating $\varphi_X(t)$ by

$$\hat\varphi_X(t) = \sum_{j=1}^{n}\sum_{k<\ell} e^{it(X^*_{jk}+X^*_{j\ell})/2} \Big/ \sum_{j=1}^{n}\sum_{k<\ell} e^{it(X^*_{jk}-X^*_{j\ell})/2}\,.$$

However, as in Section 1.2.2, since the denominator is an empirical characteristic function, it can get close to zero, especially for large $|t|$, even if (1.16) holds. To overcome this difficulty, Delaigle and Meister (2008) use a ridge parameter $\rho > 0$, and propose to estimate $f_X(x)$ by

$$\hat f_X(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\sum_{j=1}^{n}\sum_{k<\ell} e^{it(X^*_{jk}+X^*_{j\ell})/2}}{\sum_{j=1}^{n}\sum_{k<\ell} e^{it(X^*_{jk}-X^*_{j\ell})/2}+\rho}\,\varphi_K(ht)\,dt\,.$$

To choose (ρ, h) in practice, we can use the SIMEX bandwidth of Delaigle and Hall (2008). Alternatives such as the one discussed at (1.14) could also be used to avoid having to use ρ. In this case, the only parameter to choose is h. While for this we could also use the SIMEX bandwidth of Delaigle and Hall (2008), inspired by Delaigle et al. (2008), another approach that seems reasonable and which we prefer is to use a plug-in procedure of the type in Delaigle and Gijbels (2002, 2004). This approach consists in choosing h by minimising an estimator of the asymptotic mean squared error of version of the estimator $\hat f_X$ that assumes that the φ_{U_j}'s are known, i.e. of

$$\tilde f_X(x) = \frac{1}{2\pi}\int e^{-itx}\sum_{j=1}^{n}\sum_{k<\ell} e^{it(X^*_{jk}+X^*_{j\ell})/2}\,\varphi_K(ht)\Big/\sum_{j=1}^{n} n_j|\varphi_{U_j}(t/2)|^2\,dt\,.$$

1.2.4 Related literature

In the case where replicated data are available and the errors U_i are normally distributed, McIntyre and Stefanski (2011) suggested an alternative kernel density estimator in the heteroscedastic context. In the regression context, Delaigle and Meister (2007) also suggested an alternative SIMEX procedure, which can be easily adapted to the density estimation setting. Meister and Neumann (2010) considered the replication model in cases where φ_U has zeros, and Horowitz and Markatou (1996) and Neumann (2007) considered a related problem using panel data. See also Schennach (2004).

1.3 Estimation of f_U and f_X without additional data

In the previous section we assumed that we had additional data that could be used to estimate the error density f_U. However, additional data are not always available, a situation that has received increasing attention in the recent literature. Although this problem is much harder than the previous ones, fortunately one can still correct for the measurement error in certain cases. In this section we present some settings where we can identify and estimate the densities f_U and f_X without the additional data from Section 1.2. This will be done first for the case where f_U is partially known (e.g. normal distribution with unknown variance), and then for the case where only some qualitative information about f_U is available (like symmetry). The former model is referred to as the semiparametric model (as f_X is nonparametric and f_U is parametric), whereas the latter is nonparametric.

We will assume throughout this section that the classical additive error model is valid, i.e. we have a sample $X_1^*, \ldots, X_n^*$ of i.i.d. copies of X^*, where X^* comes from the model at (1.1).

1.3.1 Semiparametric estimation without additional data

In the literature on this estimation problem, two quite different main streams of ideas are developed. The first one makes assumptions on the behavior of the characteristic function of X; in the second approach, the identifiability of the model is obtained by making assumptions on the support of X. We refer also to Wang (2021), which goes deeper into the important issue of identifiability of measurement error models.

The first line of research is initiated by Matias (2002), and later extended to more general densities f_X and f_U by Butucea and Matias (2005), Meister (2006) and Butucea et al. (2008). It focuses on the supersmooth class of error distributions introduced in Delaigle (2021). Below we explain in detail the

method developed by Butucea and Matias (2005). Then we briefly explain the different contributions of the other three papers.

Suppose that the error U is supersmooth of a known order, i.e. the error is such that its characteristic function φ_U satisfies

$$d_0 \exp(-|t\sigma|^\beta) \leq |\varphi_U(t)| \leq d_1 \exp(-|t\sigma|^\beta) \quad \text{for all } |t| > M \tag{1.17}$$

for some constants $0 < d_0, d_1, M < \infty$, where $\beta > 0$ is a known constant and $\sigma > 0$ is an unknown scale parameter. This includes the case of normal errors with unknown variance, in which case $\beta = 2$ and $\sigma^2 = \sigma_U^2/2$, where σ_U^2 is the variance of U. It also includes the case where U has a stable distribution with unknown scale parameter σ, and with all other parameters known.

The density of U then only depends on the unknown parameter σ, and we will therefore focus in what follows on the identifiability and estimation of σ. Once the density of U is estimated, classical deconvolution methods, like the one at (1.6), can be used to estimate the density f_X.

Suppose that the characteristic function φ_X of X satisfies either

$$|\varphi_X(t)| \geq c_0 |t|^{-\alpha} \quad \text{for } |t| \text{ large enough,} \tag{1.18}$$

for some unknown constant $c_0 > 0$ and a known constant $\alpha > 1$ (which is e.g. the case for the exponential density) or

$$|\varphi_X(t)| \geq c_0 \exp(-c_1 |t|^\alpha) \quad \text{for } |t| \text{ large enough,} \tag{1.19}$$

for some known constants $0 < \alpha < \beta$ and $c_1 > 0$, and an unknown constant $c_0 > 0$, which is satisfied for the normal density, among many others. Under (1.17), since $\varphi_{X^*} = \varphi_X \varphi_U$ by (1.1), we have, for $|t|$ large enough,

$$\log\{|\varphi_X(t)|\}/|t|^\beta + \log(d_0)/|t|^\beta - \sigma^\beta \leq \log\{|\varphi_{X^*}(t)|\}/|t|^\beta \leq 0\,,$$

where we used the fact that $\varphi_{X^*}(t) \leq 1$ for all t. Under either of (1.18) or (1.19), $\log\{|\varphi_X(t)|\}$ tends to infinity more slowly than $|t|^\beta$. Therefore, we have

$$\lim_{|t|\to\infty} \log\{|\varphi_{X^*}(t)|\}/|t|^\beta = -\sigma^\beta.$$

This shows that σ can be identified from the distribution of X^* alone.

To derive an estimator of σ, define $\Psi(s,t) = \varphi_{X^*}(t) \exp(s^\beta t^\beta)$, and note that for $s > \sigma$,

$$\lim_{t\to+\infty} |\Psi(s,t)| = \lim_{t\to\infty} |\varphi_X(t)|\, |\varphi_U(t)| \exp(s^\beta t^\beta) = \infty,$$

whereas for $s \leq \sigma$, $\lim_{t\to+\infty} |\Psi(s,t)| = 0$, since $|\varphi_U(t)| \exp(s^\beta t^\beta)$ tends to a positive constant when $s = \sigma$, to 0 when $s < \sigma$ and to infinity faster than $|\varphi_X(t)|$ tends to zero when $s > \sigma$. Motivated by this, Butucea and Matias (2005) proposed to estimate σ by

$$\hat{\sigma} = \inf\big\{s : s > 0 \text{ and } |\hat{\Psi}(s, t_n)| \geq 1\},$$

where $\hat{\Psi}(s,t) = \hat{\varphi}_{X^*}(t)\exp(s^\beta t^\beta)$, with $\hat{\varphi}_{X^*} = n^{-1}\sum_{j=1}^n e^{itX_j^*}$ the empirical characteristic function of X^*, and where t_n is a smoothing parameter satisfying $t_n \to +\infty$ as $n \to \infty$.

Butucea and Matias' (2005) work shows that it is possible to estimate the unknown parameter σ of a parametric supersmooth error distribution, but their method requires strong assumptions on the target density f_X, since we need to know that the characteristic function φ_X decreases to zero in the tails more slowly than the characteristic function of the errors, φ_U, and this information is often not available in practice. Moreover, their estimator depends in a sensitive way on the smoothing parameter t_n, as they explain at the end of their Section 2.

In Butucea et al. (2008) the above method was adapted to the case where U has a stable symmetric distribution, that is has characteristic function

$$\varphi_U(t) = \exp(-|\sigma t|^\beta)$$

for all t, where the scale parameter $\sigma > 0$ is supposed to be known and the self-similarity index $\beta > 0$ is unknown. So in this case instead of assuming that σ is unknown and β is known (which is a special case of the model in Butucea and Matias, 2005), they make the reverse assumption. They show that the parameter β is identified, and they propose an estimator which is (as in Butucea and Matias, 2005) based on the idea of selecting the smallest value of β that satisfies a certain property. As before, the estimator depends on a tuning parameter t_n that needs to go to infinity as $n \to \infty$ and that depends on parameters that are unknown in practice.

The work of Butucea and Matias (2005) is also related to the papers by Matias (2002) and Meister (2006), who considered the special case of a normal error $U \sim N(0, \sigma_U^2)$. We focus here on the method proposed by Meister (2006); the one by Matias (2002) is rather similar. Meister (2006) assumed that for a constant $M > 0$,

$$c_0|t|^{-\alpha} \le |\varphi_X(t)| \le c_1|t|^{-\alpha} \quad \text{for all } |t| > M,$$

where c_0 and c_1 are positive finite constants and $\alpha > 1$ (not necessarily known). Under these conditions, since $\varphi_{X^*} = \varphi_X\varphi_U$, by (1.1), for all $\gamma > 1$ and $c_2 > 0$, we have

$$\lim_{|t|\to\infty} t^{-2}\log\left\{\frac{|\varphi_{X^*}(t)|}{c_2|t|^{-\gamma}}\right\} \le \lim_{|t|\to\infty}\frac{\log(c_1/c_2) + \log\left(|t|^{\gamma-\alpha}\right)}{t^2} - \frac{\sigma_U^2}{2} = -\frac{\sigma_U^2}{2},$$

since $\varphi_U(t) = \exp(-\sigma_U^2 t^2/2)$.

Using these ideas, Meister (2006) proposed to estimate σ_U^2 by

$$\hat{\sigma}_U^2 = \min\left\{\max\left(0, -2\log\left\{|\hat{\varphi}_{X^*}(t_n)|/(c_2 t_n^{-\gamma})\right\}/t_n^2\right), \sigma_n^2\right\}, \tag{1.20}$$

with σ_n^2 and t_n two sequences tending to infinity as $n \to \infty$ (see Meister, 2006),

and where $c_2 = c_0$ and $\gamma = \alpha$ if c_0 and α are known. If they are unknown, c_2 and γ are chosen arbitrarily.

Let us now turn to the second stream of techniques mentioned at the start of this subsection, namely the idea of identifying the measurement error model by making assumptions on the support of X. Schwarz and Van Bellegem (2010) assumed that $U \sim N(0, \sigma_U^2)$ with $\sigma_U^2 > 0$ unknown, and that the density f_X vanishes on a set of positive measure. They show that under this model, σ_U^2 and f_X are identified. The proof of this result relies on the fact that for any probability distributions P_1 and P_2 and for any $0 < \sigma_1 < \sigma_2$, we have that $P_1 = P_2 * N(0, \sigma_2^2 - \sigma_1^2)$ whenever $P_1 * N(0, \sigma_1^2) = P_2 * N(0, \sigma_2^2)$.

Based on this identification result, Schwarz and Van Bellegem (2010) proposed to estimate σ_U^2 and φ_X by minimizing a weighted L_1 distance between the characteristic function $\hat{\varphi}_{X^*}$ of the data, and the characteristic function $\varphi_{X^*} = \varphi_X \varphi_{\sigma_U}$ under the model, with φ_{σ_U} the characteristic function of a normal with variance σ_U^2. More precisely, they defined their estimators by any pair $(\hat{\sigma}_U^2, \hat{\varphi}_X)$ that satisfies

$$\rho(\hat{\sigma}_U^2, \hat{\varphi}_X) \leq \inf_{\tilde{\sigma}_U^2 > 0, \tilde{\varphi}_X \in \Phi_{\mathcal{C}}} \rho(\tilde{\sigma}_U^2, \tilde{\varphi}_X) + \delta_n,$$

where δ_n is some sequence tending to zero, $\Phi_{\mathcal{C}}$ is the set of all characteristic functions corresponding to some class of distributions $\mathcal{C}$, and

$$\rho(\tilde{\sigma}_U^2, \tilde{\varphi}_X) = \int_{\mathbb{R}} |\tilde{\varphi}_X(t) \varphi_{\tilde{\sigma}_U}(t) - \hat{\varphi}_{X^*}(t)| w(t)\, dt,$$

for some strictly positive and continuous weight function w. While they established consistency of their estimator under some conditions, they did not propose a way to compute their estimator in practice (e.g. how to choose δ_n, $\mathcal{C}$ and w nor how to solved the optimisation problem).

The approaches in the preceding paragraphs are important in that they come up with theoretical ways to identify the model without the need to collect additional data. However, those estimators all suffer from one or more of the following problems: they depend on one or several tuning parameters for which no clear guidelines are given, no (or limited) practical implementation is discussed, and they might lead to practical identifiability issues.

To overcome these problems, Bertrand et al. (2019) proposed a new method for the case where the error U is normal with unknown variance σ_U^2, and the support of X is compact. The identifiability of this model is guaranteed by Schwarz and Van Bellegem (2010). Assume that the support of X is equal to $[b, a + b]$, with $a > 0$ and $b \in \mathbb{R}$ two unknown constants, and write

$$X = aS + b\,,$$

where $S = (X - b)/a$ is a continuous random variable taking values on $[0, 1]$.

The advantage of formulating the problem through S is that Bernstein

(1912) showed that any continuous function $f(s)$ defined on $[0,1]$ can be approximated uniformly by the limit, as m tends to infinity, of a Bernstein polynomial of degree m, defined as follows:

$$B_m(s) = \sum_{k=0}^{m} \alpha_{k,m} b_{k,m}(s), \qquad s \in [0,1]\,,$$

where $b_{k,m}(s) = \binom{m}{k} s^k (1-s)^{m-k}$, for $k = 0, \ldots, m$ and $\alpha_{k,m} = f(k/m)$. Thus, for m large enough, the density $f_S(s)$ of S can be approximated by

$$\tilde{f}_{S,m}(s; \boldsymbol{\theta}_m) = \sum_{k=0}^{m} f_S(k/m) b_{k,m}(s) = \sum_{k=0}^{m} \theta_{k,m} f_{B_{k+1,m-k+1}}(s), \quad s \in [0,1],$$

where $\boldsymbol{\theta}_m = (\theta_{0,m}, \ldots, \theta_{m,m})$, with $\theta_{k,m} = (m+1)^{-1} f_S(k/m)$ some unknown positive parameters, and where $f_{B_{k+1,m-k+1}}$ denotes the Beta density with parameters $k+1$ and $m-k+1$.

We deduce, using (1.1), that $f_{X^*}(t) = f_X * f_U(t) = a^{-1} \int f_S\big\{(x-b)/a\big\} \phi_{\sigma_U}(t-x)\, dx$, where ϕ_{σ_U} is the density of a $N(0, \sigma_U^2)$. Thus, we can approximate $f_{X^*}(t)$ by

$$\tilde{f}_{X^*,m}(t; \sigma_U, a, b, \boldsymbol{\theta}_m) = \sum_{k=0}^{m} \theta_{k,m} \int f_{a,B_{k+1,m-k+1}}(x-b) \phi_{\sigma_U}(t-x)\, dx\,,$$

where $f_{a,B_{k+1,m-k+1}}(x) = a^{-1} f_{B_{k+1,m-k+1}}(x/a)$. It can be shown that if f_S is continuous, we have $\lim_{m\to\infty} \sup_t \big|\tilde{f}_{X^*,m}(t; \sigma_U, a, b, \boldsymbol{\theta}_m) - f_{X^*}(t)\big| = 0$.

The unknown parameters σ_U, a, b and $\boldsymbol{\theta}_m$ of this approximation are estimated by maximum likelihood, under the constraint that the $\theta_{k,m}$'s are non negative and sum to 1, which ensures that the estimator of $\tilde{f}_{S,m}(\cdot\,; \boldsymbol{\theta}_m)$ is a density. Letting $\hat{\sigma}_U, \hat{a}, \hat{b}$ and $\hat{\boldsymbol{\theta}}_m$ denote the resulting maximum likelihood estimators, the estimator of $f_X(x)$ is then given by

$$\hat{f}_X(x) = \sum_{k=0}^{m} \hat{\theta}_{k,m} f_{\hat{a},B_{k+1,m-k+1}}(x - \hat{b})\,.$$

Bertrand et al. (2019) suggested choosing m using the Bayesian Information Criterion (BIC).

1.3.2 Nonparametric estimation without additional data

In the previous section we made some parametric assumptions on the error density f_U, assuming for example that U is normal with unknown variance or that U has a stable distribution with either scale or self-similarity index unknown and estimated from the data. In this section we go one step further and make only qualitative assumptions on the density f_U, from which we can still identify and estimate f_X.

A first paper on this topic is by Meister (2007), who assumed that the characteristic function φ_U is known on a compact interval around zero. Based on this assumption, he shows that the density f_X is identifiable. To estimate f_X in practice, he decomposes its characteristic function φ_X in a Legendre polynomial basis, estimates the unknown coefficients of the decomposition, and then deduces an estimator of f_X by an inverse Fourier transform smoothed by some kernel. While this work constitutes a first step in this difficult estimation problem, assuming that φ_U is known on a compact interval around zero is only slightly less strong than assuming that f_U is completely known; for example it implies that all moments of U are known (if they exist).

Delaigle and Hall (2016) relaxed the conditions on f_U, assuming only that f_U satisfies the standard deconvolution assumption at (1.2). They introduced their technique for both continuous and discrete distributions but here we discuss only the continuous case. Under (1.2), if we do not impose any constraint on f_X then it is impossible to distinguish f_X from f_U based on data on X^*. For example, if X has a symmetric density f_X and we observe $X^* = X + U$ whereas all we know about X and U is that they are independent and f_U is symmetric, then f_{X^*}, f_X and f_U could each play the role of f_U since these three densities are symmetric (here and below we let f_T denote the density of any variable T).

Thus, under (1.2), f_X cannot be identified from data on X^* if it is not asymmetric, but asymmetry does not suffice to ensure identifiability. Indeed, suppose that X can be expressed as

$$X = Y + Z\,, \tag{1.21}$$

where Y and Z are independent and have a non degenerate distribution, f_Z is symmetric and f_Y is asymmetric. Then f_X is asymmetric, but since $X^* = X + U = Y + Z + U$, and each of f_U, f_Z and f_{Z+U} are symmetric, then each of these three densities can play the role of the symmetric error density f_U, which shows that we cannot identify f_U from data on X^*. Motivated by these considerations, Delaigle and Hall (2016) assumed that

$$f_X \text{ is asymmetric and } X \text{ cannot be decomposed as at (1.21).} \tag{1.22}$$

The condition at (1.22) may appear very restrictive as it excludes many distributions commonly used in statistics. However, Delaigle and Hall (2016) argued that in real life, data can come from such a diverse, infinite set of distributions, that we can rarely expect that f_X will be so nice and regular that it exactly satisfies (1.21). By contrast, the errors tend to be more structured (for example they are often averages of several random variables) and can often reasonably be assumed to have a nice and symmetric distribution.

The distributions for which (1.21) does not hold are known in the probability literature as non-decomposable distributions, and are notoriously difficult to characterise. However, some results suggest that "in general, a distribution is indecomposable"; see Parthasarathy et al. (1962). See also Delaigle and Hall (2016), who proved that (1.21) often does not hold in the discrete case.

Under (1.2) and (1.22), Delaigle and Hall (2016) developed a methodology based on the phase function, which, for a random variable T, is defined by $\rho_T(t) = \varphi_T(t)/|\varphi_T(t)|$, for all $t \in \mathbb{R}$. In this notation, since (1.2) implies that $\varphi_U = |\varphi_U|$ and recalling that $X^* = X + U$, we have

$$\rho_{X^*} = \varphi_{X^*}/|\varphi_{X^*}| = \varphi_X \varphi_U/|\varphi_X \varphi_U| = \rho_X \,. \tag{1.23}$$

Now, we can easily estimate ρ_X from data on X^* (it suffices to replace φ_{X^*} in (1.26) by its empirical estimator). Thus, if we could uniquely identify a distribution from its phase function, we could deduce an estimator of f_X.

However, a phase function does not uniquely identify a distribution. To see why, let

$$T = X + V \,, \tag{1.24}$$

where V and X are independent and V has a non degenerate distribution. If V has a symmetric density f_V, then $\rho_T = \rho_X$ (this is shown in the same way as the proof of (1.26)). Likewise, if we could express X as at (1.21), then we would have $\rho_X = \rho_Y$. Note too that, in these notations, we have $\text{var}(Y) < \text{var}(X) < \text{var}(T)$. Now, as argued above, in many cases, (1.21) will not hold. Since, for T at (1.24), we have $\text{var}(X) < \text{var}(T)$, this motivated Delaigle and Hall (2016) to assume that, in addition to (1.22), the distribution F_X of X is such that

$$\begin{aligned} &F_X \text{ is the unique distribution with smallest variance among all} \\ &\qquad \text{distributions with phase function } \rho_X. \end{aligned} \tag{1.25}$$

Note that (1.25) does not follow from (1.22). Indeed, even though, under (1.22), (1.21) cannot hold, where $\rho_Y = \rho_X$ and $\text{var}(Y) < \text{var}(X)$, this does not ensure that there is no random variable V which has neither the form at (1.24) nor that of Y at (1.21), but which is such that $\rho_V = \rho_X$ and $\text{var}(V) < \text{var}(X)$. As mentioned above, too little is known about properties of non-decomposable distributions for it to be possible to characterise the distributions that satisfy (1.22) and (1.25). However, Delaigle and Hall (2016) argued that, in many real applications, it is reasonable to assume that these conditions hold.

Thus, in theory, their goal is to find the distribution F_X with phase function ρ_X such that $\rho_X = \rho_{X^*}$, or equivalently, such that

$$\varphi_{X^*}(t) - \rho_X(t)|\varphi_{X^*}(t)| = 0 \quad \text{for all } t \,, \tag{1.26}$$

which has the smallest possible variance. In practice, to estimate F_X under assumptions (1.2), (1.22) and (1.25), they start by estimating the phase function of X by $\hat{\rho}_X = \hat{\rho}_{X_*} = \hat{\varphi}_{X^*}/|\hat{\varphi}_{X^*}|$, where $\hat{\varphi}_{X^*}(t) = n^{-1}\sum_{j=1}^n e^{itX_j^*}$ is the empirical characteristic function of X^*. Then, to make the optimisation problem tractable in practice, they approximate $F_X(x)$ by a discrete distribution $F(x|p) = \sum_{j=1}^m p_j I(x_j \le x)$ which puts mass $0 \le p_j \le 1$ at x_j, for $j = 1, \ldots, m$, where $\sum_{j=1}^m p_j = 1$; they take $m = 5\sqrt{n}$ and choose the x_j's at random, uniformly over $[\min_i X_i^*, \max_i X_i^*]$. Let $\rho(t|p) =$

$\varphi(t|p)/|\varphi(t|p)|$ denote the phase function of the distribution $F(x|p)$, where $\varphi(t|p) = \sum_{j=1}^{m} p_j \exp(itx_j)$. Motivated by (1.26), and noting that the quality of $\hat{\varphi}_{X^*}(t)$ degrades as $|t|$ increases, Delaigle and Hall (2016) propose to search for $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_m)$ that minimises

$$T(p) = \int \left| \varphi_{X^*}(t) - \rho(t|p)|\hat{\varphi}_{X^*}(t)| \right|^2 w(t)\, dt\,,$$

where w is a weight function, and minimises, at the same time, the variance of the distribution $F(x|p)$. See Delaigle and Hall (2016) for details of implementation. Then they estimate $\varphi_U(t)$ by $\hat{\varphi}_U(t) = \hat{\varphi}_{X^*}(t)/\hat{\varphi}(t|\hat{p})$. In the continuous case, they deduce a density estimator of $f_X(x)$ by taking $\hat{f}_X(x) = (2\pi)^{-1} \int e^{-itx} \hat{\varphi}(t|\hat{p}) \varphi_K(ht)/\tilde{\varphi}_U(t)\, dt$, where $\tilde{\varphi}_U$ is obtained from $\hat{\varphi}_U$ as at (1.14), K is a kernel function, and h is the plug-in bandwidth of Delaigle and Gijbels (2002, 2004), replacing there each occurrence of φ_U by $\tilde{\varphi}_U$.

It is worth mentioning that typically, in practice, when minimising $T(p)$ most of the $\hat{p}_j$'s are equal to zero and a more recent suggestion is to take m small and minimise over both the x_j's and the p_j's; see the R package `deconvolve` by Delaigle et al. (2021).

1.3.3 Related literature

So far, we focused on the case where the only available model equation is the measurement error model. In some case, we also know that X is the covariate of a regression model of the form $Y = g(X) + \epsilon$, where X, ϵ and U are mutually independent, $E(\epsilon) = 0$, and we have an i.i.d. sample (X_i^*, Y_i), $i = 1, \ldots, n$ with the same distribution as (X^*, Y). In such cases, this regression model can be used in an attempt to identify and estimate f_U (in addition to the other model components). This approach has been followed by Reiersøl (1950) in a seminal paper that assumes that g is linear, and by Schennach and Hu (2013), who showed conditions under which this model is identified in a fully nonparametric setting. As it turns out, there are three mutually exclusive cases, and for each case Schennach and Hu (2013) established conditions under which the model components are identified. The result establishes when the knowledge of the joint distribution of the observable variables Y and X^* uniquely determines the unobservable quantities of interest, namely the function g and the distributions of X, U and ϵ. We also refer to Schennach and Hu (2013) for related papers regarding identification in this type of models, and to the review paper by Schennach (2016) (Section 5) for papers that deal with partial identification in these models.

1.4 Boundary estimation with measurement error

We now turn to the estimation of the endpoints of the support of X, when we observe X^* at (1.1), where the distribution of U is (partially) unknown and no additional data are available. We suppose that X is not supported on the whole real line and our goal is to estimate the endpoints of its support. Without loss of generality, we suppose that it is the right endpoint τ of the support of X that is finite. Our goal in this section is to estimate τ. The case of a left endpoint is dealt with similarly. Note that most of the methods presented in the previous sections implicitly assumed that f_X was either supported on the whole real line or vanished at the endpoints of its support. For example, it is well known that kernel density estimators are not consistent near the boundary of their domain unless they vanish at the boundary.

The problem of estimating the frontier or boundary of a variable has received a lot of attention in the literature, in particular in economics where the efficiency of a firm is often measured by the distance between the output (production) of the firm and the maximal possible output. This output variable is often measured with some noise, and often not much information is available about the distribution of this noise. In that example, τ above corresponds to the largest possible output.

If X is assumed to be supported on a bounded interval, and if $U \sim N(0, \sigma_U^2)$ with unknown σ_U^2, then the method proposed by Bertrand et al. (2019) (see Section 1.3.1) can be applied since it provides estimators of each of σ_U^2,f_X and the support $[b, a+b]$ of X.

If X is not assumed to be supported on a compact interval, a number of other approaches exist in the literature. When the error U has an unknown compact support $[-c, c]$ and is unimodal with unique mode at zero (and is otherwise unknown), Hall and Simar (2002) proposed to estimate τ by:

$$\hat{\tau} = \text{argmax}_{t \in \mathcal{J}} |\hat{f}'_{X^*}(t)|, \tag{1.27}$$

where $\mathcal{J}$ is an interval in the right tail of X^* including τ, and $\hat{f}'_{X^*}$ denotes the derivative of a standard kernel density estimator based on the data $X_1^*, \ldots, X_n^*$.

Note that in the error-free case, that is if $X^* = X$, this is a standard procedure for estimating an endpoint of the support of a distribution. In the error case, a natural approach that leads to a consistent estimator of τ consists in replacing $\hat{f}'_{X^*}$ at (1.27) by a deconvolution kernel density estimator of f'_X (see Delaigle and Gijbels, 2006). However, Hall and Simar (2002) show that, under some conditions, the naive approach at (1.27) provides a consistent estimator of τ too (recall that in the error setting, a naive estimator is an estimator that applies to contaminated data, a technique that is consistent in the error-free case).

To understand why, they considered the toy example where the density

f_X is constant (say equal to $K > 0$) on some (small) interval $[a, \tau]$. Then, if $\tau > a + 2c$ (which is the case for small c), it can be easily seen that

$$f'_{X^*}(t) = \begin{cases} 0 & \text{if } a + c < t < \tau - c \\ -K f_U(t - \tau) & \text{if } \tau - c < t < \tau + c \\ 0 & \text{if } t > \tau + c. \end{cases}$$

Hence, the maximizer of $|f'_{X^*}(t)|$ for t in the right tail equals the maximizer of $f_U(t - \tau)$, which is τ thanks to the unimodality of f_U. They showed that when the density f_X is not flat to the left of τ but satisfies certain regularity conditions, and σ_U (and hence c) is small, then $\text{argmax}_t |f'_{X^*}(t)| = \tau + O(\sigma_U^2)$. Exploiting the fact that $\sigma_U^2 = o(1)$, they proposed a refined estimator, obtained by adding a bias correction term to $\hat{\tau}$. The improved estimator has a bias term of order $O(\sigma_U^3)$ if f_U has a symmetric density around zero.

Another approach was considered by Schwarz et al. (2012), under the assumption that the error $U \sim N(0, \sigma_U^2)$ with σ_U^2unknown. Their work considered a general problem of frontier estimation, but here we discuss only the part of their technique that can be used to estimate the right endpoint τ of the distribution of X. To ensure that the model is identifiable, the authors followed Schwarz and Van Bellegem (2010) (see Section 1.3.1) and assumed that the probability distribution of X vanishes on a set of positive measure. Without loss of generality, suppose that $\tau > 0$ (can be satisfied by pre-transforming the data to the positive real line).

Their approach for estimating τ is based on the observation that for any $A > \tau$, we have $\int_0^A \lim_{m\to\infty} \{F_X(x)\}^m \, dx = \int_0^A 1\{x \geq \tau\} \, dx = A - \tau$, which suggests that, for m large enough, we can estimate τ by $A - \int_0^A \lim_{m\to\infty} \{\hat{F}_X(x)\}^m \, dx$, where $\hat{F}_X$ denotes a consistent estimator of F_X. To construct $\hat{F}_X$, the authors use a sieve estimator. Specifically, they approximate the distribution of X by a discrete distribution with atoms $0 \leq \delta_1 \leq \ldots \leq \delta_{K_n}$ where $K_n \to \infty$ as $n \to \infty$; that is, they take $F_{X,\delta}(x) = K_n^{-1} \sum_{k=1}^{K_n} I(\delta_k \leq x)$. Let $\hat{F}_{X^*}(t) = n^{-1} \sum_{j=1}^n I(X_j^* \leq t)$ denote the empirical distribution function and $F_{\delta,\sigma_U} = F_{X,\delta} * \phi_{\sigma_U}$. To estimate the δ_k's and σ_U^2, they propose to minimise the following weighted L_1-distance:

$$\int_{\mathbb{R}} \left| F_{\delta,\sigma_U}(t) - \hat{F}_{X^*}(t) \right| w(t) \, dt \,,$$

under the constraint that the δ_k's and σ_U are less than a constant D_n, which is such that $D_n \to \infty$ as $n \to \infty$. Finally they estimate τ by the order-m_n frontier estimator, defined by

$$\hat{\tau} = A - \int_0^A \left\{ F_{X,\hat{\delta}}(x) \right\}^{m_n} dx, \tag{1.28}$$

where $m_n \to \infty$ as $n \to \infty$, but they didn't propose a practical choice of m_n.

Next, we consider the method proposed by Kneip et al. (2015), which

also assumes that $U \sim N(0, \sigma_U^2)$ with σ_U^2 unknown. Their approach relies on transforming the error model at (1.1) to the exponential scale by taking $\tilde{X}^* = \tilde{X} \cdot \tilde{U}$, where for a random variable T, $\tilde{T} = \exp(T)$. This has the advantage that the support of $\tilde{X}$ is compact since it is included in $[0, \tilde{\tau}]$, where $\tilde{\tau} = \exp(\tau)$ (by contrast, all we know of X is that is its support is included in $(-\infty, \tau]$).

Then they estimate $\tilde{\tau}$ by maximum likelihood under this multiplicative model. To do this, for any $\check{\tau} > 0$, $\sigma > 0$ and density h with support included in $[0, 1]$, let

$$f_{\check{\tau},\sigma,h}(t) = t^{-1} \int_0^1 \phi_\sigma \big[\log\{t/(z\check{\tau})\} \big] h(z)\, dz\,. \tag{1.29}$$

We have $f_{\tilde{X}^*}(t) = \int_0^{\tilde{\tau}} f_{\tilde{U}}(t/x) f_{\tilde{X}}(x)/x\, dx$, where $f_{\tilde{U}}(u) = \phi_{\sigma_U}(\log u)/u$ with ϕ_{σ_U} the density of a $N(0, \sigma_U^2)$. By a change of variable, we deduce that $f_{\tilde{X}^*}(t) = f_{\tilde{\tau},\sigma_U,h^*}(t)$, with $h^*(z) = \tilde{\tau} f_{\tilde{X}}(z\tilde{\tau})$.

Since h^*, $\tilde{\tau}$ and σ_U^2 are unknown, to estimate τ by maximum likelihood, we need to maximise $\sum_{i=1}^n \log f_{\check{\tau},\sigma,h}(\tilde{X}_i^*)$ with respect to the parameters $\check{\tau} > 0$ and $\sigma > 0$, and the density h whose support is included in $[0, 1]$. Since minimising over the set of such densities h is not possible in practice, Kneip et al. (2015) approximate $h(z)$ by a histogram of the type

$$h_\gamma(z) = \gamma_1 I(z = 0) + \sum_{k=1}^{M_n} \gamma_k I(q_{k-1} < z \le q_k)\,,$$

for $0 \le z \le 1$, where $q_k = k/M_n$, $M_n \to \infty$ and $n \to \infty$ and $\gamma = (\gamma_1, \ldots, \gamma_{M_n}) \in \Gamma_n$, defined by

$$\Gamma_n = \Big\{ \gamma = (\gamma_1, \ldots, \gamma_{M_n}) : \gamma_k > 0 \text{ for all } k \text{ and } \sum_{k=1}^{M_n} \gamma_k = M_n \Big\}\,.$$

Finally, estimators $\hat{\tilde{\tau}}$, $\hat{\sigma}_U$ and $\hat{h} = h_{\hat{\gamma}}$ are obtained by maximizing the following penalized likelihood :

$$(\hat{\tilde{\tau}}, \hat{\sigma}_U, \hat{\gamma}) = \mathrm{argmax}_{\check{\tau}>0,\sigma_U>0,\gamma\in\Gamma_n} \Big\{ n^{-1} \sum_{i=1}^n \log f_{\check{\tau},\sigma_U,h_\gamma}(\tilde{X}_i^*) - \lambda\, \mathrm{pen}(f_{\check{\tau},\sigma_U,h_\gamma}) \Big\},$$

where $\lambda \ge 0$ is a fixed constant independent of n, and $\mathrm{pen}(f_{\check{\tau},\sigma_U,h_\gamma}) = \max_{3\le j\le M_n} |\gamma_j - 2\gamma_{j-1} + \gamma_{j-2}|$ is a smoothness penalty.

Under regularity conditions, Kneip et al. (2015) showed that, as long as the density $f_{\tilde{X}}$ has a jump at the boundary $\tilde{\tau}$, the estimators $\hat{\tilde{\tau}}$ and $\hat{\sigma}_U$ are consistent and converge at a logarithmic rate. In practice they recommended using $M_n = \max\{3, 2\,[n^{1/5}]\}$, with $[a]$ the nearest integer to a; for λ they selected the value of λ by minimizing, over the grid $\log_{10}\lambda = -2, -1, 0, 1, 2$, a bootstrap estimator of the relative root mean squared error of $\hat{\tilde{\tau}}$.

Recently, Florens et al. (2020) proposed an estimator of τ that makes

minimal assumptions on the error density f_U. For X^* at (1.1), write $X^* = \tau - Z + U$, where $Z = \tau - X$ is an unobservable random variable independent of U, supported on $[0, \infty)$ and with density f_Z. Suppose that f_U is symmetric around zero.

Their approach is based on the identifiability of the cumulants of Z. Denote the ℓ-th cumulant of Z and X^* by $\kappa_Z(\ell)$ and $\kappa_{X^*}(\ell)$, respectively. Using standard properties of cumulants, it can be proved that

$$\kappa_{X^*}(1) = E(X^*) = \tau - \kappa_Z(1), \qquad \kappa_{X^*}(2p+1) = -\kappa_Z(2p+1), \quad p \geq 1\,,$$
$$\kappa_{X^*}(2p) = \kappa_Z(2p) + \kappa_U(2p) \quad \text{for } p \geq 1\,.$$

Thus, the odd cumulants of Z of order 3 and higher are identifiable since they can be expressed in terms of the observable X^*. On the other hand, since the even cumulants of f_U do not vanish in general, the even cumulants of Z are not identifiable from the distribution of X^*. This shows that to make the distribution of Z identifiable, we have to restrict f_Z to a class of densities that are determined by the odd cumulants of order 3 and higher.

The above formulation can also be expressed through characteristic functions, as follows. For any random variable V, let $\eta_V(t) = \frac{\partial}{\partial t}\{\mathrm{Im} \log \varphi_V(t)\}$, where Im denotes the imaginary part, and let

$$T_V(t) = \eta_V(t) - \eta_V(0) = \sum_{p=1}^{\infty} \frac{(-1)^p t^{2p}}{(2p)!} \kappa_V(2p+1)\,. \tag{1.30}$$

Then the class of densities f_V that are determined by their odd cumulants of order 3 and higher is equivalent to the class of densities that are determined by $T_V(t)$ for all t. Thus, if f_Z belongs to a class of densities f_V that are determined by $T_V(t)$ for all t, it is identifiable from the distribution of X^*.

In practice Florens et al. (2020) suppose that f_Z belongs to a flexible parametric family $\{f_Z(\cdot|\lambda) : \lambda \in \Lambda\}$, where Λ is a compact subset of $\mathbb{R}^m$; thus, $f_Z(\cdot) = f_Z(\cdot|\lambda_0)$ for some $\lambda_0 \in \Lambda$. Let $T_\lambda(t) = \eta_V(t) - \eta_V(0)$, where $V \sim f_Z(\cdot|\lambda)$. The results above show that λ (and hence f_Z) is identified if the parametric class is such that

$$T_\lambda \equiv T_{\lambda_0} \Longrightarrow \lambda = \lambda_0\,. \tag{1.31}$$

Note in particular that we have $T_{\lambda_0} = -T_{X^*}$. Once f_Z is identified, we can identify τ and $\sigma_U^2 = \mathrm{var}(U)$ from the distribution of X^* since

$$\tau = E(X^*) + E(Z) \quad \text{and} \quad \sigma_U^2 = \mathrm{var}(X^*) - \mathrm{var}(Z)\,. \tag{1.32}$$

For a parametric family satisfying (1.31), to estimate λ, Florens et al. (2020) minimise a penalised weighted L_2-distance

$$\hat{\lambda}_\alpha = \mathrm{argmin}_{\lambda \in \Lambda} \Big\{ \int_{\mathbb{R}} \{\hat{T}_{X^*}(t) - T_\lambda(t)\}^2 w(t)\, dt + \alpha \|\lambda\|^2 \Big\},$$

where $\alpha > 0$ is a regularization parameter, w is a weight function, $\|\lambda\|$ is the Euclidean norm of λ and $\hat{T}_{X^*}(t) = \hat{\eta}_{X^*}(t) - \hat{\eta}_{X^*}(0)$, with $\hat{\eta}_{X^*}(t) = \frac{\partial}{\partial t}\{\mathrm{Im}\log\hat{\varphi}_{X^*}(t)\}$ and $\hat{\varphi}_{X^*}(t) = n^{-1}\sum_{j=1}^{n}\exp(itX_j^*)$. Then, using (1.32), they estimate τ and σ_U^2 by $\hat{\tau}_\alpha = \overline{X^*} + E_{\hat{\lambda}_\alpha}(Z)$ and $\hat{\sigma}^2_{U,\alpha} = \hat{\sigma}^2_{X^*} - \mathrm{var}_{\hat{\lambda}_\alpha}(Z)$, where $\overline{X^*}$ and $\hat{\sigma}^2_{X^*}$ are the sample mean and variance of the X_i^*'s, and $E_{\hat{\lambda}_\alpha}(Z)$ and $\mathrm{var}_{\hat{\lambda}_\alpha}(Z)$ are the mean and variance corresponding to the density $f_Z(z|\hat{\lambda}_\alpha)$.

In practice, Florens et al. (2020) suggest the following flexible parametric family $\{f_Z(z|\lambda) : \lambda \in \Lambda\}$:

$$f_Z(z|\lambda) = \frac{e^{-z}}{\|\lambda\|^2}\Big\{\sum_{k=0}^{m}\lambda_k v_k(z)\Big\}^2,$$

where $\lambda_0 = 1$, $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$, and v_k are Laguerre polynomials, i.e. polynomials that are orthonormal with respect to the exponential density e^{-z}. It can be shown that as the number m of parameters tends to infinity, the Hellinger distance (restricted to an interval A) between any continuous density f_Z defined on $\mathbb{R}^+$ and its Laguerre approximation $f_Z(\cdot|\lambda)$ tends to zero. In practice Florens et al. (2020) recommended to choose the regularization parameters α and the number of polynomials m by minimizing the bootstrap estimate of the root mean squared error of the estimator $\hat{\tau}_\alpha$.

1.5 Effect of misspecifying the error distribution

Consistent estimation of the error density is not always possible and it is important to understand the effect that misspecifying the error distribution can have on the nonparametric deconvolution estimators of f_X such as those introduced in the previous sections. Clearly, in standard asymptotic terms, where properties of an estimator are analyzed for $n \to \infty$, misspecifying the error distribution implies non consistency of deconvolution estimators.

Indeed, if we assume that the error density is, say, f_ξ instead of the true f_U, then estimators such as the one at (1.3) will consistently estimate

$$(2\pi)^{-1}\int e^{-itx}\varphi_{X^*}(t)/\varphi_\xi(t)\,dt = (2\pi)^{-1}\int e^{-itx}\varphi_X(t)\varphi_U(t)/\varphi_\xi(t)\,dt,$$

which in general is different from f_X if $\varphi_U \neq \varphi_\xi$. Further this integral is not even guaranteed to exist, nor to be equal to a density. In particular, if $\varphi_\xi(t)$ tends to zero faster than $\varphi_U(t)$ as $|t| \to \infty$, i.e. if we misspecify the error density in such a way that we assume it is smoother than it really is, then not only is the resulting estimator of f_X not consistent, but it also does not converge to a well defined quantity. (Here, as usual in the deconvolution

literature, we qualify the smoothness of a distribution through the speed of decay of its characteristic function in the tails, where a faster decay corresponds to a smoother distribution). A thorough theoretical study of this error misspecification problem has been studied in Meister (2004).

These considerations suggest that the consequences of assuming that the error distribution is smoother than it really is are so severe that, when in doubt, we should rather assume that the error density is not very smooth. In particular, recalling the distinction between ordinary smooth and supersmooth errors introduced in Delaigle (2021), we should preferably assume that the error is ordinary smooth with a low level of smoothness, rather than supersmooth. For example, in many cases we could assume that the errors have a Laplace distribution, since this distribution is ordinary smooth with $\beta = 2$ (and thus not very smooth).

Using a completely different approach, arguments in favor of a Laplace distribution were also provided by Delaigle (2008). There, the author suggested that, in measurement error problems, if it reasonable to derive asymptotic properties of estimators (in particular of the deconvolution kernel density estimator) taking $n \to \infty$ and $\sigma_U^2 = \text{var}(U) \to 0$. The rationale behind this approach is that, in the conventional error-free case, asymptotics where $n \to \infty$ usually describe properties of an estimator in situations where the sample becomes ideal (i.e. of increasingly better quality). In the context with errors, the quality of a sample does not depend only on its size, but also on the value of σ_U^2, and an ideal sample is a sample that has both a large sample size and a small error variance. This approach was also used by Fan (1992) in the supersmooth error case, by Hall and Simar (2002) in the context of boundary estimation, and by Staudenmayer and Ruppert (2004) to justify their SIMEX procedure. Delaigle (2008) shows that the deconvolution kernel estimator is relatively robust to error misspecification, as long as the error variance σ_U^2 is reasonably well estimated. Further, as in the standard asymptotic approach, the effect of error misspecification is more serious if we assume that the error distribution is smoother than it is, compared to assuming that the error distribution is less smooth than it is. In particular, in this context too, it is preferable to assume that the error is ordinary smooth, and assuming that the error is Laplace seems to guarantee relative robustness of the estimator.

This double asymptotic approach is useful to understand several issues that are encountered in practice, but which are not accessible via classical theory. For example, this work explains why, in practice, assuming Laplace errors often results in estimators that perform well, even if the error distribution is not Laplace. In particular it provides some theoretical underpinning to the low-order estimator of Carroll and Hall (2004), which can work really well in practice when σ_U^2 is not too large (the advantage of the low-order approach is that it only requires to estimate low-order moments of the error, whereas the deconvolution estimator requires knowledge of the entire error distribution). These double asymptotics also justify theoretically the fact that, when the error variance is not too large, the rate of convergence of the deconvolution

kernel estimator with Gaussian errors can be much faster than logarithmic (it can be as fast as algebraic). Intuitively this is clear: if there is little noise in the data, then even if this noise is Gaussian, deconvolving should not be too difficult. However, the classical theory does not take into account the magnitude of the error variance. For example, even though, in finite samples, deconvolving Laplace errors with $\text{var}(U) = 2\text{var}(X)$ is much more difficult than deconvolving normal errors with $\text{var}(U) = 0.2\text{var}(X)$, the classical asymptotic theory only states that rates in the latter case are logarithmic, whereas rates in the former case are algebraic, thus considerably faster.

Acknowledgments

Delaigle acknowledges support from the Australian Research Council (DP170102434); Van Keilegom acknowledges support from the European Research Council (2016-2021, Horizon 2020/ ERC grant agreement No. 694409).

Bibliography

Apanasovich, T. and H. Liang (2021). Nonparametric measurement errors models for regression. In G. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter 14, pp. ? CRC.

Bernstein, S. (1912). Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Communications de la Société mathématique de Kharkow 13*, 1–2.

Bertrand, A., I. Van Keilegom, and C. Legrand (2019). Flexible parametric approach to classical measurement error variance estimation without auxiliary data. *Biometrics 75*(1), 297–307.

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

Butucea, C., C. Matias, and C. Pouet (2008). Adaptivity in convolution models with partially known noise distribution. *Electronic Journal of Statistics 2*, 897–915.

Carroll, R., J. Eltinge, and D. Ruppert (1993). Robust linear regression in replicated measurement error models. *Statistics & probability letters 16*(3), 169–175.

Carroll, R. and P. Hall (1988). Optimal rates of convergence for deconvolving a density. *Journal of the American Statistical Association 83*(404), 1184–1186.

Carroll, R. and P. Hall (2004). Low order approximations in deconvolution and regression with errors in variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 31–46.

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Carroll, R. and L. Stefanski (1990). Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association 85*(411), 652–663.

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

Delaigle, A. (2021). Deconvolution kernel density estimator. In G. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter 10, pp. ? CRC.

Delaigle, A. and I. Gijbels (2002). Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 869–886.

Delaigle, A. and I. Gijbels (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Computational statistics & data analysis 45*(2), 249–267.

Delaigle, A. and I. Gijbels (2006). Estimation of boundary and discontinuity points in deconvolution problems. *Statistica Sinica*, 773–788.

Delaigle, A. and P. Hall (2008). Using simex for smoothing-parameter choice in errors-in-variables problems. *Journal of the American Statistical Association 103*(481), 280–287.

Delaigle, A. and P. Hall (2016). Methodology for non-parametric deconvolution when the error distribution is unknown. *Journal of the Royal Statistical Society, Series B 78*, 231–252.

Delaigle, A., P. Hall, and A. Meister (2008). On deconvolution with repeated measurements. *The Annals of Statistics 36*(2), 665–685.

Delaigle, A., P. Hall, and H. Müller (2007). Accelerated convergence for nonparametric regression with coarsened predictors. *The Annals of Statistics 35*(6), 2639–2653.

Delaigle, A., T. Hyndman, and T. Wang (2021). deconvolve: Deconvolution tools for measurement error problems. R package version 0.1.0.

Delaigle, A. and A. Meister (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *Journal of the American Statistical Association 102*, 1416–1426.

Delaigle, A. and A. Meister (2008). Density estimation with heteroscedastic error. *Bernoulli 14*, 562–579.

Diggle, P. and P. Hall (1993). A fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal statistical society: series B (Methodological) 55*(2), 523–531.

Fan, J. (1992). Deconvolution with supersmooth distributions. *Canadian Journal of Statistics 20*(2), 155–169.

Florens, J.-P., L. Simar, and I. Van Keilegom (2020). Estimation of the boundary of a variable observed with symmetric error. *Journal of the American Statistical Association 115*(529), 425–441.

Gleser, L. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. *Contemp. Math 112*, 99–114.

Hall, P. and L. Simar (2002). Estimating a changepoint, boundary, or frontier in the presence of observation error. *Journal of the American Statistical Association 97*, 523–534.

Horowitz, J. and M. Markatou (1996). Semiparametric estimation of regression models for panel data. *The Review of Economic Studies 63*(1), 145–168.

Kang, Y. and P. Qiu (2021). Nonparametric deconvolution by fourier transformation and other related approaches. In G. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter 11, pp. ? CRC.

Kneip, L., L. Simar, and I. Van Keilegom (2015). Frontier estimation in the presence of measurement error with unknown variance. *Journal of Econometrics 184*, 379–393.

Li, T. and Q. Vuong (1998). Semiparametric deconvolution with unknown noise variance. *Journal of Multivariate Analysis 65*, 139–165.

Madansky, A. (1959). The fitting of straight lines when both variables are subject to error. *Journal of the American Statistical Association 54*, 173–205.

Matias, C. (2002). Semiparametric deconvolution with unknown noise variance. *ESAIM, Probability and Statistics 6*, 271–292.

McIntyre, J. and L. Stefanski (2011). Density estimation with replicate heteroscedastic measurements. *Annals of the Institute of Statistical Mathematics 63*(1), 81–99.

Meister, A. (2004). On the effect of misspecifying the error density in a deconvolution problem. *Canadian Journal of Statistics 32*(4), 439–449.

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

Meister, A. (2007). Deconvolving compactly supported densities. *Mathematical Methods of Statistics 16*, 63–76.

Meister, A. and M. H. Neumann (2010). Deconvolution from non-standard error densities under replicated measurements. *Statistica Sinica*, 1609–1636.

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

Neumann, M. H. (2007). Deconvolution from panel data with unknown error distribution. *Journal of Multivariate Analysis 98*(10), 1955–1968.

Parthasarathy, K., R. Rao, and S. Varadhan (1962). On the category of indecomposable distributions on topological groups. *Transactions of the American Mathematical Society 102*, 200–217.

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

Schennach, S. M. (2004). Nonparametric regression in the presence of measurement error. *Econometric Theory 20*(6), 1046–1093.

Schennach, S. M. (2016). Recent advances in the measurement error literature. *Annual Review of Economics 8*, 341–377.

Schennach, S. M. and Y. Hu (2013). Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association 108*, 177–186.

Schwarz, M. and S. Van Bellegem (2010). Consistent density deconvolution under partially known error distribution. *Statistics and Probability Letters 80*, 236–241.

Schwarz, M., S. Van Bellegem, and J.-P. Florens (2012). Nonparametric frontier estimation from noisy data. In *Exploring Research Frontiers in Contemporary Statistics and Econometrics - Festschrift in Honor of L. Simar, Van Keilegom, I. and Wilson, P.W. (eds.)*, pp. 45–64. Springer.

Staudenmayer, J. and D. Ruppert (2004). Local polynomial regression and simulation-extrapolation. *Journal of the Royal Statistical Society, Series B 66*, 17–30.

Stefanski, L. and R. Carroll (1990). Deconvolving kernel density estimators. *Statistics 21*(2), 169–184.

Wang, L. (2021). Identifiability in measurement error models. In G. Yi, A. Delaigle, and P. Gustafson (Eds.), *Handbook of Measurement Error Models*, Chapter 5, pp. ? CRC.