Published in final edited form as:
Comput Stat Data Anal. 2010 January 1; 54(1): 25–36. doi:10.1016/j.csda.2009.08.012.

Estimating smooth distribution function in the presence of heteroscedastic measurement errors

Xiao-Feng Wang[a,*], **Zhaozhi Fan**[b], and **Bin Wang**[c]

[a] Department of Quantitative Health Sciences/Biostatistics, Cleveland Clinic, Cleveland, OH 44195, USA

[b] Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada

[c] Department of Mathematics and Statistics, University of South Alabama, Mobile, AL, USA

Abstract

Measurement error occurs in many biomedical fields. The challenges arise when errors are heteroscedastic since we literally have only one observation for each error distribution. This paper concerns the estimation of smooth distribution function when data are contaminated with heteroscedastic errors. We study two types of methods to recover the unknown distribution function: a Fourier-type deconvolution method and a simulation extrapolation (SIMEX) method. The asymptotics of the two estimators are explored and the asymptotic pointwise confidence bands of the SIMEX estimator are obtained. The finite sample performances of the two estimators are evaluated through a simulation study. Finally, we illustrate the methods with medical rehabilitation data from a neuro-muscular electrical stimulation experiment.

Keywords

Smooth distribution function; Measurement errors; Kernels; Deconvolution; Fourier method; SIMEX; Heteroscedasticity; Confidence bands

1. Introduction

Many practical problems involve estimation of distribution functions or density functions from indirect observations. For example, in low level microarray data from either the complementary DNA (cDNA) microarray or the Affymetrix GeneChip system, each observation is an original signal coupled with a background noise. To obtain an expression measure, the goals often include developing better statistical tools or enhancing algorithms for background correction so that the disease genes can be detected accurately and efficiently. In medical image analysis, observable outputs are often blurred images. In astronomy, due to great astronomical distances and atmospheric noise, most data are subject to measurement errors. Statistical analyses that ignore measurement errors could be misleading. Measurement error model is an active, rich research field in statistics. There is an enormous literature on this topic in linear regression, as

*Corresponding author: Xiao-Feng Wang, Ph.D., Department of Quantitative Health Sciences/Biostatistics, Cleveland Clinic, 9500 Euclid Avenue/JJN3, Cleveland, OH 44195, wangx6@ccf.org (Xiao-Feng Wang).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

summarized by Fuller (1987) and in nonlinear models, as summarized by Carroll et al. (2006). In this paper, we investigate estimation methods of smooth distribution function from data contaminated with heteroscedastic measurement errors.

Nonparametric kernel type methods have been widely used in estimating density functions and their derivatives or in regression. Kernel smoothing is also an important tool for distribution function estimation. The kernel estimate of a distribution function, first introduced by Nadaraya (1964), has been investigated by many authors (Azzalini, 1981; Reiss, 1981; Sarda, 1993; Bowman et al., 1998). The kernel smooth estimator, which has good statistical properties, can be expressed as

$$\widehat{F}_X(x)=\frac{1}{n}\sum_{j=1}^{n}L\left(\frac{x-X_j}{h_n}\right), \tag{1}$$

where $X_1, X_2, \ldots, X_n$ are independent random variables with the common distribution function F_X and the density function f_X. The function L is defined from a kernel K as $L(x)=\int_{-\infty}^{x}K(t)dt$, where $K(\cdot)$ is some bounded density function with $K(t) = K(-t)$ for all $t \in \mathbb{R}$ and $h_n > 0$ is the smoothing parameter with $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. The conventional *empirical distribution function* (EDF) can be obtained by letting $h_n \to 0$, when $L(\cdot)$ is replaced by an indictor function $I(X_i \le x)$. It is noted that $\hat{F}$ in (1) can be written as

$$\widehat{F}_X(x)=\int_{-\infty}^{x}\widehat{f}_X(t)dt,$$

where $\hat{f}$ is the well-known kernel density estimator.

In real data applications, there are many examples where observable data are contaminated with measurement errors and it is not realistic to assume that the errors are homoscedastic. The measurement process might be subjective and differs among all individuals. Fuller (1987) had an early consideration of this problem. Cheng and Riu (2006) discussed the point estimation of the parameters in a linear measurement error (heteroscedastic errors in variables) model. Kulathinal et al. (2002) considered estimation problem of an errors-in-variables regression model when the variances of the measurement errors vary between observations in the analysis of aggregate data in epidemiology. Sun et al. (2002) studied a measurement error model with application to astronomical data that came with information on their heteroscedastic errors. In the research of density estimation with measurement errors, the literature is vast; see for example, Fan (1991); Zhang (1990); Stefanski and Carroll (1990); Carroll and Hall (1989); Delaigle and Gijbels (2004). They focused on the study of the *deconvoluting kernel density estimation* through an inverse Fourier transform with the case of homoscedastic errors. Until very recently, Delaigle and Meister (2008) studied the deconvoluting kernel estimator under the heteroscedastic setting. Staudenmayer et al. (2008) addressed a different type of model, where the observable data with heteroscedastic measurement errors were assumed normally distributed. A Monte Carlo Markov chain and a random-walk Metropolis-Hastings algorithm were proposed to estimate the unknown density.

Estimating cumulative distribution functions with measurement errors was also of interest. The sample cumulative distribution function is a nonlinear function of data and is biased when it is estimated ignoring measurement errors. Stefanski and Bay (1996) studied the estimation of a discrete population cumulative distribution function when data were contaminated with measurement errors. Using the method of simulation extrapolation (SIMEX) (Cook and

Stefanski, 1994; Stefanski and Cook, 1995), they proposed a bias-adjusted estimator that reduced much of the bias. Later, Cordy and Thomas (1996) considered an expectation-maximization algorithm for estimating a distribution function when data were from a mixture of a finite number of known distribution. Nusser et al. (1996) presented a method for estimating distributions with additive normal errors with application to a study of daily dietary intakes. Their approach differed from the kernel estimators in that they assumed that a transformation existed such that both the original observations and the measurement errors were normally distributed. Most recently, Hall and Lahiri (2008) studied Fourier-type estimation of distributions, moments, and quantiles with homoscedastic errors.

Challenges of estimation problems arise when errors are heteroscedastic where we literally have only one observation for each error distribution. In this paper, we study two types of methods to recover the unknown smooth distribution function: a Fourier-type deconvolution method and a SIMEX method, when data are contaminated with heteroscedastic errors. In section 2, we first extend Hall and Lahiri's (2008) Fourier-type estimator to the case of heteroscedastic errors and then generalize the work of Stefanski and Bay (1996) to estimate the smooth distribution function with heteroscedastic errors using SIMEX. The asymptotics of the two estimators are studied and the asymptotic pointwise confidence bands of the SIMEX estimator are obtained. In section 3, we conduct a simulation study to compare the finite sample performances of the two estimators. In section 4, we apply our proposed method to real data in a medical rehabilitation study. In section 5, we close this paper with a discussion.

2. Estimation methods

To investigate the estimation of the smooth distribution function for data contaminated with heteroscedastic errors, we first consider a general *heteroscedastic measurement error model*. Let $Y_1, \cdots, Y_n$ be an observed random sample such that

$$Y_j = X_j + U_j, \tag{2}$$

with the measurement error U_j independent of X_j. Each U_j has its own density f_{U_j}, $j = 1, \cdots, n$, where $f_{U_1}, \cdots, f_{U_n}$ are from a same distributional family, but the measurement error distribution's parameters vary with the observation index. If $f_{U_1} = \cdots = f_{U_n} = f_U$, the errors are said to be *homoscedastic*; otherwise, the errors are *heteroscedastic*. One is to recover the unknown distribution function $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$ of the unobserved continuous random variable X, where f_X is the density function of X.

2.1. Fourier-type Deconvolution

Denote the characteristic functions of X and U_j by φ_X and φ_{U_j} ($j = 1, \cdots, n$), respectively. Through an inverse Fourier transform, Delaigle and Meister's (2008) deconvolution estimator for the density with heteroscedastic errors can be written as a form of a kernel-type density estimator,

$$\widehat{f}_{X,Fourier}(x) = \frac{1}{nh_n} \sum_{j=1}^{n} \tilde{K}_j \left(\frac{x - Y_j}{h_n} \right), \tag{3}$$

where

$$\tilde{K}_j(z)=\frac{1}{2\pi}\int e^{-itz}\frac{\varphi_K(t)}{\psi_{U_j}(t/h_n)}dt, \quad \psi_{U_j}(t)=\frac{\frac{1}{n}\sum_{k=1}^{n}|\varphi_{U_k}(t)|^2}{\varphi_{U_j}(-t)},$$

and φ_K is the characteristic function of a symmetric probability kernel, $K(\cdot)$, with a finite variance. It is noted that $\hat{f}_{X,Fourier}$ in (3) becomes the conventional deconvoluting kernel estimator (Stefanski and Carroll, 1990) for homoscedastic errors when $f_{U_1} = \cdots = f_{U_n} = f_U$.

Similarly to Hall and Lahiri (2008), our distribution estimator $\hat{F}_{X,Fourier}$ in the case of heteroscedastic errors is defined as simply the integral of $\hat{f}_{X,Fourier}$ over $(-\infty, x]$. Let

$$\tilde{L}_j(z)=\frac{1}{2}+\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{\sin(tz)\cdot\varphi_K(h_n t)}{t\cdot\psi_{U_j}(t)}dt \quad , \quad j=1,\cdots,n.$$

By integrating $\hat{f}_{X,Fourier}$ in (3), we have

$$\begin{aligned}\widehat{F}_{X,Fourier}(x) &= \int_{-\infty}^{x}\widehat{f}_{X,Fourier}(t)dt \\ &= \frac{1}{n}\sum_{j=1}^{n}\tilde{L}_j(x-Y_j).\end{aligned} \tag{4}$$

We now study asymptotic properties of our estimator. Due to the nature of the deconvolution kernel estimation, similarly as defined in Fan (1991), we consider a slightly different definition of the distribution function estimator as in the following form to derive the asymptotics. For a sequence of positive numbers $M_n \to \infty$ as $n \to \infty$,

$$\widehat{F}_{n,Fourier}(x)=\int_{-M_n}^{x}\widehat{f}_{X,Fourier}(t)dt.$$

Here we included the terms about M_n in the uniform bound. Conventionally, M_n can be selected to be proportional to h_n^{-1}, similar to that discussed in Fan (1991). Note that the bandwidth h_n depends on n when studying the asymptotics. We assume the following conditions:

1. For an integer $m \geq 0$, $0 < \alpha \leq 1$,

$$|f^{(m)}(x) - f^{(m)}(y)| \leq C \cdot |x-y|^{\alpha};$$

2. $K(x) < D \cdot |x|^{-m-2}$ for all x;
3. $\varphi_K(t)$ has support $[-1, 1]$;
4. Condition C of Delaigle and Meister (2008).

The asymptotic behavior of the estimator is described in the next theorem.

Theorem 1—Assume Conditions (1) – (4). Then, for $x_0 \in \mathbb{R}$ with specifically chosen sequence of numbers $M_n \to \infty$ and bandwidth $h_n \to 0$, we have

$$\sup_{-\infty<x_0<\infty} E[\widehat{F}_{n,Fourier}(x_0) - F(x_0)]^2 \le O(h_n^{2(m+\alpha+1)}) + O(F^2(-M_n(1-h_n))) + O(M_n^{-2m-2}) + O\left(M_n\left[h_n\sum_{k=1}^{n}|\varphi_{U_k}(1/h_n)|^2\right]^{-1}\right).$$

Remark 1—The last term of the upper bound reflects the effect of the heteroscedastic measurement errors to the estimation of the unknown distribution function. The smoothness of the measurement error distributions influence the convergence rate of the distribution function estimation, as discussed in Fan (1991). In the above theorem, we present a general asymptotic result of our estimator. The specific convergence rate could be obtained similar as in Hall and Lahiri (2008), where they classified the distribution functions of the unobservable random variable and the measurement errors into eight classes.

2.2. SIMEX

SIMEX is a “jackknife”-type bias-adjusted method that has been widely applied in regression problems with measurement errors. Stefanski and Cook (1995) applied the SIMEX algorithm to parametric regression problems. Staudenmayer and Ruppert (2004), Carroll et al. (1999) discussed the nonparametric regression in the presence of measurement errors using SIMEX method. Stefanski and Bay (1996) studied SIMEX estimation of a finite population cumulative distribution function when sample units are measured with errors. Now we generalize Stefanski and Bay’s (1996) method to estimate the smooth distribution function with heteroscedastic Gaussian errors.

Under the model setting (2), we further assume $U_j \sim N(0, \sigma_j^2)$, for $j = 1, \cdots, n$. Typically, σ_j can be obtained from auxiliary data. By the general SIMEX algorithm, estimators are re-computed from a large number B of measurement error-inflated pseudo data sets, $\{Y_{jb}(\lambda)\}_{j=1}^n$, $b = 1, \cdots, B$, with

$$Y_{jb}(\lambda) = Y_j + \lambda^{1/2}U_{jb} = Y_j + \sigma_j\lambda^{1/2}Z_{jb}, \quad j=1,\cdots,n, \quad b=1,\cdots,B,$$

where Z_{jb} are independent, standard normal pseudo-random variables, and $\lambda \ge 0$ is a constant controlling the amount of added errors.

The smooth distribution function estimator from the bth variance-inflated data $\{Y_{jb}(\lambda)\}_{j=1}^n$ is

$$\widehat{G}_b(x) = \frac{1}{n}\sum_{j=1}^{n} L\left(\frac{x - Y_{jb}}{h}\right),$$

where $L(\cdot)$ is defined from a kernel $K(\cdot)$ as $L(x)=\int_{-\infty}^{x}K(t)dt$ and $K(\cdot)$ is a symmetric kernel with a finite variance such that $\int K(t)dt = 1$, $\int tK(t)dt = 0$, and $\int t^2K(t)dt < \infty$.

In the general SIMEX algorithm, the simulation and estimation steps are repeated a large number of times, and the average value of the estimators for each level of contamination is calculated by,

$$\widehat{G}(x)=\frac{1}{B}\sum_{b=1}^{B}\widehat{G}_b(x)$$
$$=\frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{n}\sum_{j=1}^{n}L\left(\frac{x-Y_{jb}}{h}\right)\right). \tag{5}$$

However, it is noted that $U_{jb} = \sigma_j Z_{jb}$, where Z_{jb} are independent, standard normal random variables. With a fixed j ($j = 1, \cdots, n$), let $\tilde{h}_j = h/(\sigma_j\lambda^{1/2})$, we have

$$\frac{1}{B}\sum_{b=1}^{B}L\left(\frac{x-Y_{jb}}{h}\right)$$
$$=\frac{1}{B}\sum_{b=1}^{B}L\left(\frac{x-Y_j-\lambda^{1/2}U_{jb}}{h}\right)$$
$$=\frac{1}{B}\sum_{b=1}^{B}L\left(\frac{\frac{x-Y_j}{\sigma_j\lambda^{1/2}}-Z_{jb}}{\tilde{h}_j}\right).$$

Notice that the smooth distribution estimator is asymptotically unbiased and has the same variance as the EDF. It uniformly converges to the true distribution function with probability one (Nadaraya, 1964). Hence, with a fixed j, conditional on Y_j,

$$\frac{1}{B}\sum_{b=1}^{B}L\left(\frac{x-Y_{jb}}{h}\right)\xrightarrow{\mathcal{D}}\Phi\left(\frac{x-Y_j}{\sigma_j\lambda^{1/2}}\right),$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution.

Therefore, the simulation step can be bypassed in the SIMEX algorithm for the distribution estimation. We use $\hat{G}^*$ in (6) to replace $\hat{G}$ in (5) for our estimation,

$$\widehat{G}^*(x,\lambda)\triangleq\frac{1}{n}\sum_{j=1}^{n}\Phi\left(\frac{x-Y_j}{\sigma_j\lambda^{1/2}}\right). \tag{6}$$

The above equation has some similarity with equation (3) in Stefanski and Bay (1996). We further calculate the quantity in (6) for a pre-determined sequence of λ, i.e. $0 \le \lambda_1 < \lambda_2 < \cdots < \lambda_l$. The success of SIMEX technique depends on the fact that the expectation of $\hat{G}^*$ is well-approximated by a nonlinear function of λ (Carroll et al., 2006). Here we consider the conventional quadratic function of λ, *i.e.*

$$\mathrm{E}(\widehat{G}^*(x,\lambda)) \approx \beta_0+\beta_1\lambda+\beta_2\lambda^2.$$

It is indeed a good approximation when $\max\{\sigma_j^2\}$ is not very large. The SIMEX estimator for the unknown distribution function F_X without measurement error can be obtained by the extrapolation step, *i.e.* letting $\lambda \to -1$,

$$\widehat{F}_{X,SIMEX}(x)=\widehat{\beta}_0-\widehat{\beta}_1+\widehat{\beta}_2. \tag{7}$$

The following proposition shows that, in the case of heteroscedastic Gaussian errors, our estimator (7) is an asymptotically unbiased estimator for the unknown distribution function F_X, if the extrapolant is exact.

Proposition 1—Assume (a) the polynomial extrapolant is exact, (b) the distribution function $\mathrm{F_X(x)}$ of unobserved $\mathrm{X_1, \cdots, X_n}$ has continuous fourth derivative, and (c) $\sigma_j^2<C_0$ for all j and $\mathrm{C_0} < \infty$. Then, the estimator (7) is an asymptotically unbiased estimator of $\mathrm{F_X(x)}$. The corresponding asymptotic variance is $\mathrm{F_X(x)(1 - F_X(x))/n}$, which can be consistently estimated by $\mathrm{\hat{F}_{X,SIMEX}(x)(1 - \hat{F}_{X,SIMEX}(x))/n}$.

Remark 2—Under realistic applications, the assumption (a) in proposition 1 will only be approximately true. See for instance Carroll et al. (1999); Staudenmayer and Ruppert (2004). As we show in the equation (8) in Appendix, the expectation of $\hat{\mathrm{G}}^*(\mathrm{x}, \lambda)$ is a quadratic function of λ at a given x plus the approximation error term, $o\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j^4\right)$. The success of the SIMEX method in practice depends on the fact that $\mathrm{E}[\hat{\mathrm{G}}^*(\mathrm{x}, \lambda)]$ can well-approximated by a quadratic function of λ.

From proposition 1, the estimated asymptotic pointwise confidence bands when data are contaminated with heteroscedastic Gaussian errors have the form

$$I(x)=\left(\widehat{F}_{X,SIMEX}(x)-c\sqrt{\widehat{F}_{X,SIMEX}(x)(1-\widehat{F}_{X,SIMEX}(x))/n},\right.$$
$$\left.\widehat{F}_{X,SIMEX}(x)+c\sqrt{\widehat{F}_{X,SIMEX}(x)(1-\widehat{F}_{X,SIMEX}(x))/n}\right),$$

where c is chosen as the $(1-\alpha/2)$ quantile of the standard normal distribution.

Remark 3—Due to the heterogeneity of measurement errors, the variance estimation method in Stefanski and Bay (1996) can not be applied to our proposed estimator $\hat{\mathrm{F}}_{\mathrm{X,SIMEX}}(\mathrm{x})$. However, the variance of $\hat{\mathrm{F}}_{\mathrm{X,SIMEX}}(\mathrm{x})$ can be well approximated by the variance of $\hat{\mathrm{G}}^*(\mathrm{x}, \lambda)$ as $\lambda \to -1$. The latter can then be approximately consistently estimated. The naive confidence bands therefore can be constructed. The confidence bands based on the aforementioned variance estimation are simulated. It almost coincides the nonparametric bootstrap confidence band. This variance estimation of the distribution function can also be applied to the case of homoscedastic measurement errors.

Remark 4—In the SIMEX distribution estimation, it is possible that the estimated values are out of [0, 1] in tail regions. This situation is similar to many classes of kernel methods, such as wavelet density estimators, sinc kernel estimators, and spline estimators. The disadvantage will not affect the global performance of our estimation. A simple correction version of our SIMEX estimator in practice is

$$\widehat{F}^*_{X,SIMEX}(x) = \begin{cases} 0 & : \widehat{F}_{X,SIMEX}(x)<0 \\ \widehat{F}_{X,SIMEX}(x) & : 0 \le \widehat{F}_{X,SIMEX}(x) \le 1. \\ 1 & : 1<\widehat{F}_{X,SIMEX}(t) \end{cases}$$

3. Numerical Study

3.1. Empirical choice of smoothing parameters

In the Fourier-type method, we consider a modified “plug-in” approach to select the bandwidth h_n. Fan (1992) proposed a plug-in asymptotical bandwidth for the density estimation with homoscedastic errors. Delaigle and Gijbels (2004) suggested a “normal reference” approach and Hall and Lahiri (2008) discussed it for distribution estimation with homoscedastic errors. Here we adopt Hall and Lahiri’s (2008) “normal reference” bandwidth approach for the case of heteroscedastic errors. Specifically, we shall temporarily take f_X be a normal $N(0, \sigma_X^2)$ density; and calculate an estimator $\widehat{\sigma}_X^2$ of σ_X^2 as the variance of the data Y minus $\overline{\sigma}_U^2$, where $\overline{\sigma}_U^2 = \sum_{j=1}^n \sigma_j^2/n$.

The SIMEX estimator requires specification of $\lambda_1, \cdots, \lambda_l$. They are not the smoothing parameters in the sense of the conventional kernel estimation. They act as “design points” of our estimator, which is similar to the knots in a spline (Kooperberg and Stone, 1992). So, the choice of $\lambda_1, \cdots, \lambda_l$ is not as sensitive as the bandwidth in density estimation is. Based on our extensive simulations, our experience suggests that the number of values, l, is not critical and neither is that of λ_l if λ_1 is determined. Note that $\sigma_j\lambda^{1/2}$ in (6) has some similarity to a bandwidth. Our proposed rule-of-thumb choice of λ_1 can be obtained by solving the equation

$$\overline{\sigma}_U \lambda_1^{1/2} = c_1 \widehat{h}_{rot,Y},$$

where $\hat{h}_{rot,Y}$ is the Silverman’s rule-of-thumb bandwidth (Silverman, 1986) based on the observed data Y and $c_1 = \sqrt{\widehat{\sigma}_Y^2 - \overline{\sigma}_U^2}/\widehat{\sigma}_Y$ is a coefficient to adjust the effect of measurement errors. The sequence $\lambda_1, \cdots, \lambda_l$ is then taking equally-spaced values over the interval $[\lambda_1, 3 + \lambda_1]$ with $l = 50$.

3.2. Finite-sample performance

We now investigate the finite sample performances of the Fourier-type estimator and the SIMEX estimator via a simulation study. Our study involves three types of target distributions: (1) $X \sim N(0, 1)$, (2) $X \sim 0.5\,N(-3, 1) + 0.5\,N(3, 1)$, and (3) $X \sim \Gamma(2, 1)$. From each of these distributions, 500 samples of size $n = 50$, 100, and 500 are generated, each of which is then contaminated by heteroscedastic errors. The measurement errors are generated from

$N(0, \sigma_j^2)$, where σ_j ($j = 1, \cdots, n$) are generated from $U(a, b)$ and (a, b) are chosen to be (0.4, 0.6), (0.8, 1), respectively.

To assess the quality of the smooth distribution function estimators, we use the *integrated squared error* (ISE) criterion:

$$\mathrm{ISE}(\widehat{F}(x)) = \int \left\{ \widehat{F}(x) - F(x) \right\}^2 dx.$$

We compare four estimators in the study: the Fourier-type estimator; the SIMEX estimator; the naive estimator, *i.e.* the smooth distribution estimator of Y where measurement errors are ignored; and the smooth distribution estimator from the uncontaminated sample X.

Table 1 summarizes the results of the average of the 500 ISEs for different estimators under different simulation conditions. The simulation results show that both the Fourier-type estimator and the SIMEX estimator perform much better than the naive estimator in terms of the ISE criterion. The ISEs for the Fourier-type method and the ISEs for the SIMEX method become smaller and closer to the ISEs from the uncontaminated sample when sample sizes become larger. We also note that the ISEs are larger at the case of $\sigma_U \sim U(0.8, 1)$ than those at the case of $\sigma_U \sim U(0.4, 0.6)$ under the same simulation conditions. This is due to the level of difficulty of deconvolution. Comparing the Fourier method and the SIMEX method, we find that SIMEX estimator performs better than the Fourier method when the sample sizes are small and error variances are large. They become very close as sample size is sufficiently large. This is not surprising because the converge rate is slow for the Fourier method while the SIMEX method is fast when the polynomial extrapolant is accurate.

Figure 1 allows us to display and compare estimated curves visually. In the upper plot, the true distribution is standard normal, $N(0, 1)$. The measurement errors are generated from $N(0, \sigma_j^2)$ and $\sigma_j \sim U(0.4, 0.6)$, $j = 1, \cdots, n$ with sample size $n = 500$. The solid line is the smooth distribution estimator from X; the dashed line is the SIMEX estimator; the dotted line is the Fourier estimator; and the dot-dashed line is the naive estimator. Both the Fourier estimator and the SIMEX estimator give very close results. They recover the true distribution accurately from the sample Y, while the naive estimator is far from the true distribution functions. In the lower plot, the true distribution is a normal mixture 0.5 $N(-3, 1)$ + 0.5 $N(3, 1)$. The measurement errors are generated from $N(0, \sigma_j^2)$ and $\theta_i \sim U(0.8, 1)$, $j = 1, \cdots, n$ with sample size $n = 500$. Despite of the complexity of the true distribution and measurement errors, we see that both the Fourier method and the SIMEX method perform well in recovering the true distribution function.

To investigate the performance of the estimated asymptotic confidence bands with the SIMEX method derived in the last section, we compare them with the nonparametric bootstrap confidence bands and the estimated confidence bands from the uncontaminated sample X. Figure 2 shows a simulated example of the SIMEX estimate and three different confidence bands of the distribution function $F(x)$. The true distribution is a normal mixture $0.5N(-3, 1) + 0.5N(3, 1)$ and the measurement errors are generated from $N(0, \sigma_j^2)$ and $\sigma_j \sim U(0.8, 0.9)$, $j = 1, \cdots, n$ with sample size $n = 500$. The solid lines are the SIMEX estimate and its 95% associated asymptotic confidence bands from contaminated sample Y. The dashed lines are the nonparametric bootstrap confidence bands from Y. The number of bootstrap replicates is 1000. The dotted lines are the estimated confidence bands from the uncontaminated sample X. The estimated asymptotic confidence bands are nearly identical with the nonparametric bootstrap confidence bands and are very close to the the estimated confidence bands from uncontaminated sample except in the tail areas. The asymptotic estimates of confidence bands, hence, are recommended due to their simplicity and computational easiness.

4. A real data application

Spinal cord injury (SCI) is damage to the spinal cord often due to traumatic accident, resulting in upper and lower motor neuron lesions. It typically leads to paralysis and loss of sensation in parts of the body controlled through the spinal cord below the level where the injury occurred. All individuals with SCI, and particularly those with complete lesions, are considered to be at high risk of *pressure ulcer* development throughout their lifetime. Pressure ulcers are areas of damaged skin and tissue that develop when sustained pressure occurs. Traditionally, techniques to reduce pressure ulcer incidence have focused on reducing extrinsic risk factors. These techniques include providing cushions to improve pressure distribution. Another approach is educating individuals on the importance of regular pressure relief procedures. *Neuro-muscular electrical stimulation* (NMES) is the application of electrical stimuli to a group of muscles, which is a new clinic tool for pressure ulcers to produce beneficial changes at the user/support system interface by altering the intrinsic characteristics of the user’s paralyzed tissue itself (Bogie et al., 2006, 2008).

The primary goal of the NMES study at Cleveland FES center was to investigate the distribution of pressure intensities for each patient under different clinical conditions such as before and after NMES treatment. Pressure intensity data at the seating interface for each patient were recorded by using the Tekscan advanced clinseat pressure mapping system (Tekscan, Inc.). The seating interface was divided as a 48 × 42 matrix. The pressure intensity of each element was measured simultaneously. Figure 3 displays an example of pressure intensity data in the NMES study. Pressure intensities at the seating area for one subject correspond to color-scale rectangular segments in the image. The color bar indicates the mapping from data values to colors. However, the outcomes of pressure intensities are subject to measurement errors. Measurements were taken at several different times for a patient, resulting in replicate measurements of pressure intensity maps. It is reasonable to assume that the observable intensities are contaminated with heteroscedastic Gaussian errors at the seating area. From the replicate measurement data, we are able to estimate the heteroscedastic variances of measurement errors for each active location. Hence, the fundamental statistical question is how to estimate the unobserved distribution of pressure intensities from the data contaminated with errors. In this example, our pressure data contain total 1518 activated observations. The pressure intensities with measurement errors have mean 42.99 and standard deviation 16.73. The range of the observed intensities is from 11.0 to 120.0. The heteroscedastic standard deviations of measurement errors vary from 6.32 to 8.01.

We conduct the analysis using both the Fourier-type and the SIMEX methods for the data from the NMES study. Figure 4 displays the smooth distribution estimation of pressure intensities in the single case study. The solid line is the estimated distribution function by the SIMEX method and the dotted line is the estimate by the Fourier-type method, while the dashed line is the estimated distribution function by the naive method where measurement errors are ignored. The recovered function shows asymmetric features. Both the Fourier-type estimator and the SIMEX estimator obtain coincident results and there is only a slight difference at the left tails. The curve for the naive estimator does not fall within the confidence region of the SIMEX estimator. The example demonstrates that correcting the measurement errors is critical in the statistical analysis. After recovering the distribution function, we will be able to make further statistical inferences, such as comparing two recovered distribution functions under different clinical treatments.

5. Discussion

We studied two bias-correction methods to estimate smooth distribution function for the data contaminated with heteroscedastic errors. The Fourier-type method was generalized from the conventional deconvolution kernel method (Hall and Lahiri, 2008), which can be applied to any arbitrary error distributions. The SIMEX method, extending the Stefanski and Bay's (1996) work, was easy to be implemented and computationally fast due to the fact that the simulation step was bypassed. As shown in the simulation results, both methods allowed us to recover accurately the true distribution function from a contaminated sample when the sample size was large. However, the SIMEX method worked better than the Fourier-type method when sample size was small and the variances of measurement errors were large. The asymptotic variance of the SIMEX estimator was obtained and the simulation showed that the native estimate of the asymptotic confidence bands performed very well. Thus, applying the SIMEX method in real applications would be attractive.

We addressed in section 2 that the success of the SIMEX method depended on the fact that the expectation of $\hat{G}^*(x, \lambda)$ was well-approximated by a quadratic function of λ for a small $\frac{1}{n}\sum_{j=1}^{n}\sigma_j^4$. How large should the error level be for the SIMEX method to be feasible? With a moderate sample size, our simulations suggested that $\max(\sigma_j) < 1$ could work adequately well for the three special cases considered here. For other cases, more simulations and study would be needed in order to answer the above question.

The SIMEX method we discussed only dealt with the case of the Gaussian errors, while the Fourier-type method could work with a large classes of errors including both super-smoothed errors and ordinary-smoothed errors as defined by Fan (1991). Indeed, the SIMEX method could also be applied to the case of exponential errors. Sometimes, measurement errors only could be positive in medical applications. An exponential distribution, as a non-zero mean, skewed distribution, was a common assumption in those studies (Ballico, 2001; Savin, 2000). We noticed that, however, the SIMEX estimator in the case of exponential errors was not asymptotically unbiased even if the polynomial extrapolant was exact. A natural idea was to apply an extra smoothing step (*i.e.* smoothing splines) on $\hat{F}_{X,SIMEX}(x)$ to reduce the bias in the estimation. Our extended simulation study showed that the extra smoothing step made the SIMEX estimator for exponential errors surprisingly well to recover the true distribution.

As one of reviewers pointed out, it would be interesting to compare the performances of our estimators with (homoscedastic) estimators that heteroscedastic measurement error variances were replaced by a constant measurement error variance (the average of heteroscedastic variances). Our preliminary simulation showed that the degree of variation of heteroscedastic measurement error variances affected the estimators using the constant variance. A more comprehensive study of the model misspecification problem for both density and distribution estimation could be done in our future research.

The research of the confidence bands for the Fourier-type method remained as an open problem. Bissantz et al.'s (2007) nonparametric confidence intervals in deconvolution density estimation was relevant where they only focused on the homoscedastic and ordinary-smoothed errors.

Acknowledgments

We are grateful to the reviewers and the associate editor for their valuable comments. We thank Dr. Kath Bogie from the Cleveland FES center and Case Western Reserve University for making NMES clinical data available for this analysis. The research of Xiao-Feng Wang is supported in part by the NIH Clinical and Translational Science Awards. The research of Zhaozhi Fan is supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

Azzalini A. A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika 1981;68:326–328.

Ballico M. Calculation of key comparison reference values in the presence of non-zero-mean uncertainty distributions, using the maximum-likelihood technique. Metrologia 2001;38:155–159.

Bogie KM, Wang X, Fei B, Sun J. New technique for real-time interface pressure analysis: Getting more out of large image data sets. Journal of Rehabilitation Research and Development 2008;45 (4):523–536. [PubMed: 18712638]

Bogie KM, Wang X, Triolo RJ. Long-term prevention of pressure ulcers in high-risk patients: a single case study of the use of gluteal neuromuscular electric stimulation. Arch Phys Med Rehabil 2006;87 (4):585–591. [PubMed: 16571402]

Bowman A, Hall P, Prvan T. Bandwidth selection for the smoothing of distribution functions. Biometrika 1998;85:799–808.

Carroll RJ, Hall P. Optimal rates of convergence for deconvolving a density. Journal of the American Statistical Associations 1989;83:1184–1186.

Carroll RJ, Maca J, Ruppert D. Nonparametric regression in the presence of measurement error. Biometrika 1999;86:541–554.

Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceanu, C. Measurement Error in Nonlinear Models: A Modern Perspective. 2. Chapman Hall; New York: 2006.

Cheng CL, Riu J. On estimating linear relationships when both variables are subject to heteroscedastic measurement errors. Technometrics 2006;48:511–519.

Cook JR, Stefanski LA. Simulation extrapolation estimation in parametric measurement error models. Journal of the American Statistical Association 1994;89:1314–1328.

Cordy C, Thomas D. Deconvolution of a distribution function. Journal of the American Statistical Association 1996;92:1459–1465.

Delaigle A, Gijbels I. Practical bandwidth selection in deconvolution kernel density estimation. Computational Statistics and Data Analysis 2004;45:249–267.

Delaigle A, Meister A. Density estimation with heteroscedastic error. Bernoulli 2008;14:562–579.

Fan J. On the optimal rates of convergence for nonparametric deconvolution problems. The Annals of Statistics 1991;19:1257–1272.

Fan J. Deconvolution with supersmooth distributions. The Canadian Journal of Statistics 1992;20:155–169.

Fuller, WA. Measurement Error Models. John Wiley & Sons; New York: 1987.

Hall P, Lahiri SN. Estimation of distributions, moments and quantiles in deconvolution problems. Annals of Statistics 2008;36 (5):2110–2134.

Kooperberg C, Stone C. Logspline density estimation for censored data. Journal of Computational and Graphical Statistics 1992;1:301–328.

Kulathinal SB, Kuulasmaa K, Gasbarra D. Estimation of an errors-in-variables regression model when the variances of the measurement errors vary between the observations. Statistics in Medicine 2002;21:1089–1101. [PubMed: 11933035]

Nadaraya EA. Some new estimates for distribution functions. Theory of Probability and its Applications 1964;9:497–500.

Nusser SM, Carriquiry AL, Dodd KW, Fuller WA. A semi-parametric transformation approach to estimating usual daily intake distributions. Journal of the American Statistical Association 1996;91 (436):1440–1449.

Reiss RD. Nonparametric estimation of smooth distribution functions. Scandinavian Journal of Statistics 1981;8:116–119.

Sarda P. Smoothing parameter selection for smooth distribution functions. Journal of Statistical Planning and Inference 1993;35:65–75.

Savin SK. Reliability of parameter control with exponential error. Measurement Techniques 2000;43:329–334.

Silverman, BW. Density Estimation. Chapman and Hall; London: 1986.

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

Staudenmayer J, Ruppert D, Buonaccorsi J. Density estimation in the presence of heteroskedastic measurement error. Journal of the American Statistical Association 2008;103:726–736.

Stefanski LA, Bay JM. Simulation extrapolation deconvolution of finite population cumulative distribution function estimators. Biometrika 1996;83:407–417.

Stefanski LA, Carroll RJ. Deconvoluting kernel density estimators. Statistics 1990;21:169–184.

Stefanski LA, Cook JR. Simulation extrapolation: the measurement error jackknife. Journal of the American Statistical Association 1995;90:1247–1256.

Sun, J.; Morrison, H.; Harding, P.; Woodroofe, M. Density and mixture estimation from data with measurement errors. 2002. Technical Report, http://sun.cwru.edu/~jiayang/india.ps

Zhang CH. Fourier methods for estimating mixing densities and distributions. The Annals of Statistics 1990;18:806–830.

Appendix

Proof of Theorem 1

Under the regularity conditions (1) ~ (4), for $x_0 \in (-\infty,\infty)$ with specifically chosen sequence of numbers $M_n \to \infty$ and bandwidth $h_n \to 0$, the bias of $\hat{F}_{n,Fourier}(x_0)$ is

$$\begin{aligned}
Bias[\widehat{F}_{n,Fourier}(x_0)] &= E[\widehat{F}_{n,Fourier}(x_0)] - F(x_0) \\
&= \int_{-M_n}^{x_0} E[\widehat{f}(x)]dx - f(x_0) \\
&= \int_{-M_n}^{x_0} \int_{-\infty}^{\infty} f(x-y)\frac{1}{h_n}K(\frac{y}{h_n})dydx - F(x_0) \\
&= \int_{-\infty}^{\infty} \frac{1}{h_n}[F(x_0-y) - F(-M_n-y)]K(\frac{y}{h_n})dy - F(x_0).
\end{aligned}$$

The norm of the bias is then

$$\begin{aligned}
\left|Bias[\widehat{F}_{n,Fourier}(x_0)]\right| \le & \left|\int_{-\infty}^{\infty} \frac{1}{h_n}F(x_0-y)K(\frac{y}{h_n})dy - F(x_0)\right| \\
& + \int_{-\infty}^{\infty} F(-M_n - h_n y)K(y)dy.
\end{aligned}$$

It is uniformly bounded from above by

$$\begin{aligned}
O(h_n^{m+\alpha+1}) & \\
& + O(F(\\
& -M_n(1-h_n))) \\
& + O(M_n^{-m-1}).
\end{aligned}$$

On the other hand, the variance of $\hat{F}_{n,Fourier}(x_0)$ is

$$
\begin{aligned}
Var[\widehat{F}_{n,Fourier}(x_0)] &= \sum_{j=1}^{n} \frac{1}{(2\pi)^2} Var\left[\int_{-M_n}^{x_0}\int_{-\infty}^{\infty} e^{-itx}\varphi_K(t\cdot h_n)e^{itY_j}\psi_j(t)dtdx\right] \\
&\leq (M_n+|x_0|)\sum_{j=1}^{n}\frac{1}{(2\pi)^2}\int_{-\infty}^{\infty}|\varphi_K(t\cdot h_n)|^2|\psi_j(t)|^2dt \\
&= \frac{M_n+|x_0|}{4\pi^2}\int_{-\infty}^{\infty}|\varphi_K(t\cdot h_n)|^2\left[\sum_{k=1}^{n}|\varphi_{U_k}(t)|^2\right]^{-1}dt \\
&= \frac{M_n+|x_0|}{4\pi^2}\int_{-1/h_n}^{1/h_n}|\varphi_K(t\cdot h_n)|^2\left[\sum_{k=1}^{n}|\varphi_{U_k}(t)|^2\right]^{-1}dt \\
&= \frac{M_n+|x_0|}{2\pi^2}\int_{0}^{1/h_n}|\varphi_K(t\cdot h_n)|^2\left[\sum_{k=1}^{n}|\varphi_{U_k}(t)|^2\right]^{-1}dt \\
&= \frac{M_n+|x_0|}{2\pi^2}\int_{0}^{1}|\varphi_K(s)|^2\left[\sum_{k=1}^{n}|\varphi_{U_k}(s/h_n)|^2\right]^{-1}h_n^{-1}ds,
\end{aligned}
$$

which is uniformly bounded from above by

$$
O\left(M_n\cdot\left[h_n\sum_{k=1}^{n}|\varphi_{U_k}(1/h_n)|^2\right]^{-1}\right).
$$

Thus, the conclusion follows.

Proof of Proposition 1

Under the conditions (a) ~ (c),

$$
\begin{aligned}
E\left[\widehat{G}^*(x,\lambda)\right] &= \frac{1}{n}\sum_{j=1}^{n}E\left[\Phi\left(\frac{x-Y_j}{\sigma_j\lambda^{1/2}}\right)\right] = \frac{1}{n}\sum_{j=1}^{n}E\left[\Phi\left(\frac{x-X_j-\sigma_jZ_j}{\sigma_j\lambda^{1/2}}\right)\right] \\
&= \frac{1}{n}\sum_{j=1}^{n}\int\int\left[\Phi\left(\frac{x-u-\sigma_jz}{\sigma_j\lambda^{1/2}}\right)\right]f(u)\varphi(z)dudz \\
&= \frac{1}{n}\sum_{j=1}^{n}\int\left[-\int\Phi(v)dF(x-\sigma_j\lambda^{1/2}v-\sigma_jz)\right]\varphi(z)dz \\
&= \frac{1}{n}\sum_{j=1}^{n}\int\int F(x-\sigma_j\lambda^{1/2}v-\sigma_jz)\varphi(v)\varphi(z)dudz \\
&= \frac{1}{n}\sum_{j=1}^{n}E[F(x-(1+\lambda)^{1/2}\sigma_jZ)]
\end{aligned}
$$

where Z is a standard normal random variable. By Lebesgue dominated convergence theorem, $\lim_{\lambda\to-1} E[\hat{G}^*(x,\lambda)] = F(x)$.

Using Taylor expansion, if, $\max(\sigma_j^4)$ is small, the above expectation can be written as

$$
\begin{aligned}
E\left[\widehat{G}^*(x,\lambda)\right] =& F(x)+\frac{1}{2n}F''(x)(\lambda+1)\sum_{j=1}^{n}\sigma_j^2 \\
&+\frac{3}{4!n}F^{(4)}(x)(\lambda+1)^2\sum_{j=1}^{n}\sigma_j^4+o\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j^4\right).
\end{aligned} \qquad (8)
$$

Hence the expectation of $\hat{G}^*(x, \lambda)$ is well-approximated by a quadratic function of λ at a given x with the approximation error $o\left(\frac{1}{n}\sum_{j=1}^{n}\sigma_j^4\right)$. *i.e.* $E[\hat{G}^*(x, \lambda)] \approx \beta_0 + \beta_1\lambda + \beta_2\lambda^2$.

Therefore,

$$\begin{aligned} E\left[\widehat{F}(x,\lambda)\right] =E\left[\widehat{\beta}_0-\widehat{\beta}_1+\widehat{\beta}^2\right] &= \lim_{\lambda\to -1} E\left[\widehat{\beta}_0-\widehat{\beta}_1\lambda+\widehat{\beta}_2\lambda^2\right] \\ &\approx \lim_{\lambda\to -1} E\left[\widehat{G}^*(x,\lambda)\right]=F(x). \end{aligned}$$

By analogous argument as above, the variance of $\hat{G}^*(x, \lambda)$ is

$$\begin{aligned} Var(\widehat{G}^*(t,\lambda)) &= \frac{1}{n^2}\sum_{j=1}^{n} Var(\Phi(\frac{x-Y_j}{\lambda^{1/2}\sigma_j})) \\ &= \frac{1}{n^2}\sum_{j=1}^{n}[E\{2\Phi(V)F[x-(1+\lambda)^{1/2}\sigma_j Z]\} \\ &\quad -\{E|F(x-(1+\lambda)^{1/2}\sigma_j Z)]\}^2] \end{aligned}$$

where V is a standard normal random variable,

$$Z=\frac{V+\lambda^{1/2}W}{(1+\lambda)^{1/2}},$$

which has standard normal distribution. W is a standard normal random variable, independent of V.

By Lebesgue dominated convergence theorem,

$$\begin{aligned} &\lim_{\lambda\to -1} Var\left[\widehat{G}^*(x,\lambda)\right] \\ &\qquad =F(x)(1-F(x))/n, \end{aligned}$$

thus, it can then be estimated by $\hat{F}_{X,SIMEX}(x)(1 - \hat{F}_{X,SIMEX}(x))/n$. The proof is complete.

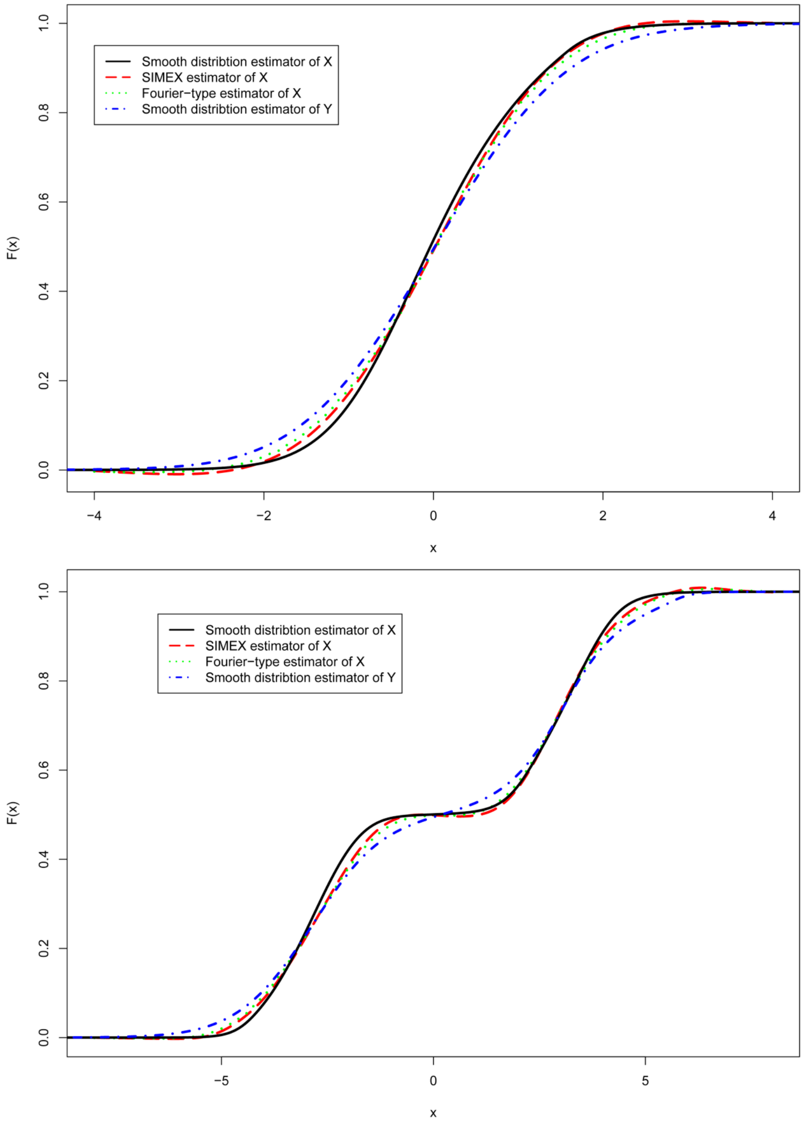

Figure 1.

Distribution estimation for data contaminated with heteroscedastic errors: In the upper plot, the true distribution is standard normal, $N(0, 1)$. The measurement errors are from $N(0, \sigma_j^2)$ and $\sigma_j \sim U(0.4, 0.6)$, $j = 1, \cdots, n$ with sample size $n = 500$. In the lower plot, the true distribution is a normal mixture $0.5N(-3, 1) + 0.5N(3, 1)$. The measurement errors are from $N(0, \sigma_j^2)$ and $\sigma_j \sim U(0.8, 1)$, $j = 1, \cdots, n$ with sample size $n = 500$. Solid line – the smooth distribution estimator from X; dashed line – the SIMEX estimator; dotted line – the Fourier estimator; dot-dashed line – the naive estimator.

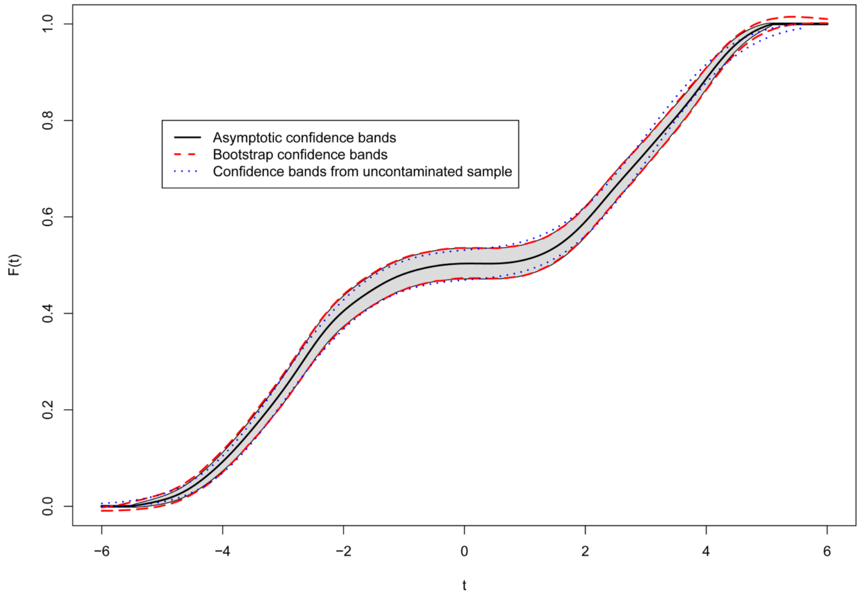

Figure 2.

The SIMEX estimator and its associated 95% confidence bands of the distribution function: the true distribution is normal mixture 0.5*N*(−3, 1) + 0.5*N*(3, 1). The measurement errors are from $N(0,\sigma_j^2)$ and $\sigma_j \sim U(0.8, 0.9)$, $i = 1, \cdots, n$ with sample size $n = 500$. Solid line – the SIMEX estimator and its associated asymptotic confidence bands; dashed line – the nonparametric bootstrap confidence bands; dotted line – the estimated confidence bands from uncontaminated sample *X*.

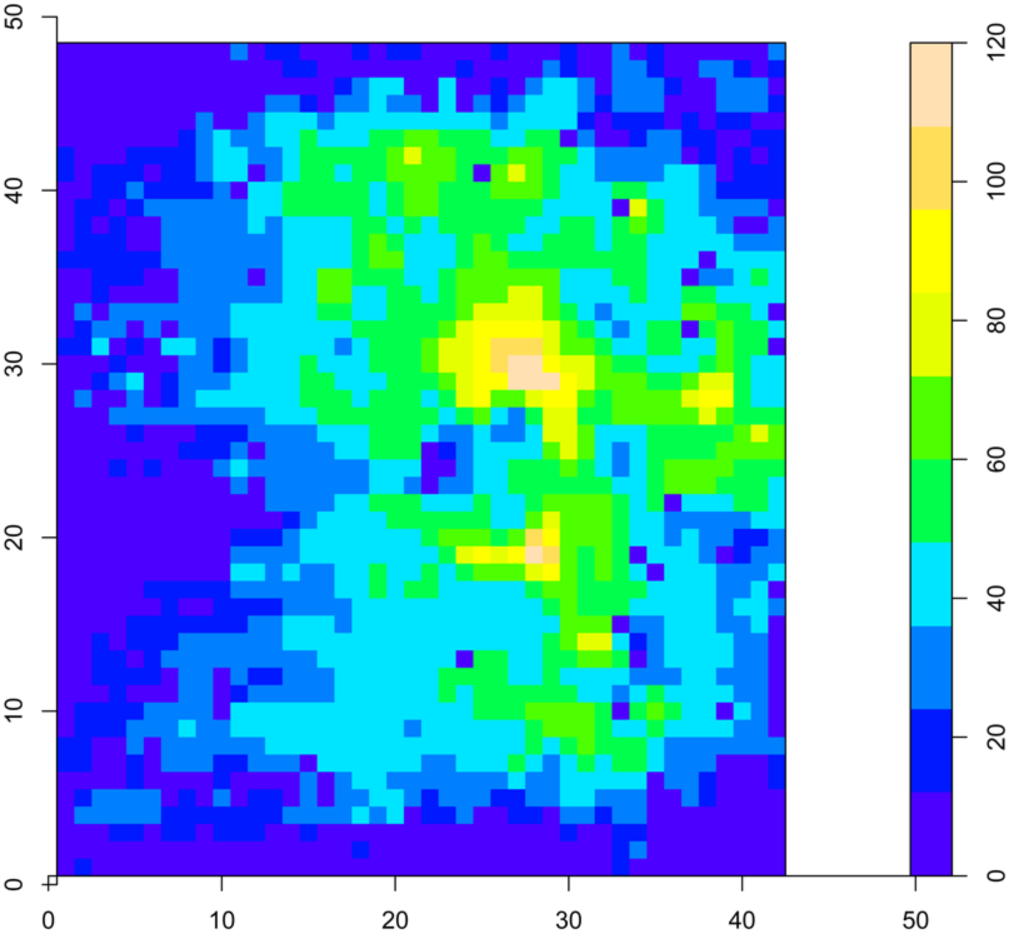

Figure 3.

An example of pressure intensity data at the seating interface for one subject: each pressure intensity at the seating area corresponds to a color-scale rectangular segment in the image. The color bar indicates the mapping from intensity values to colors.

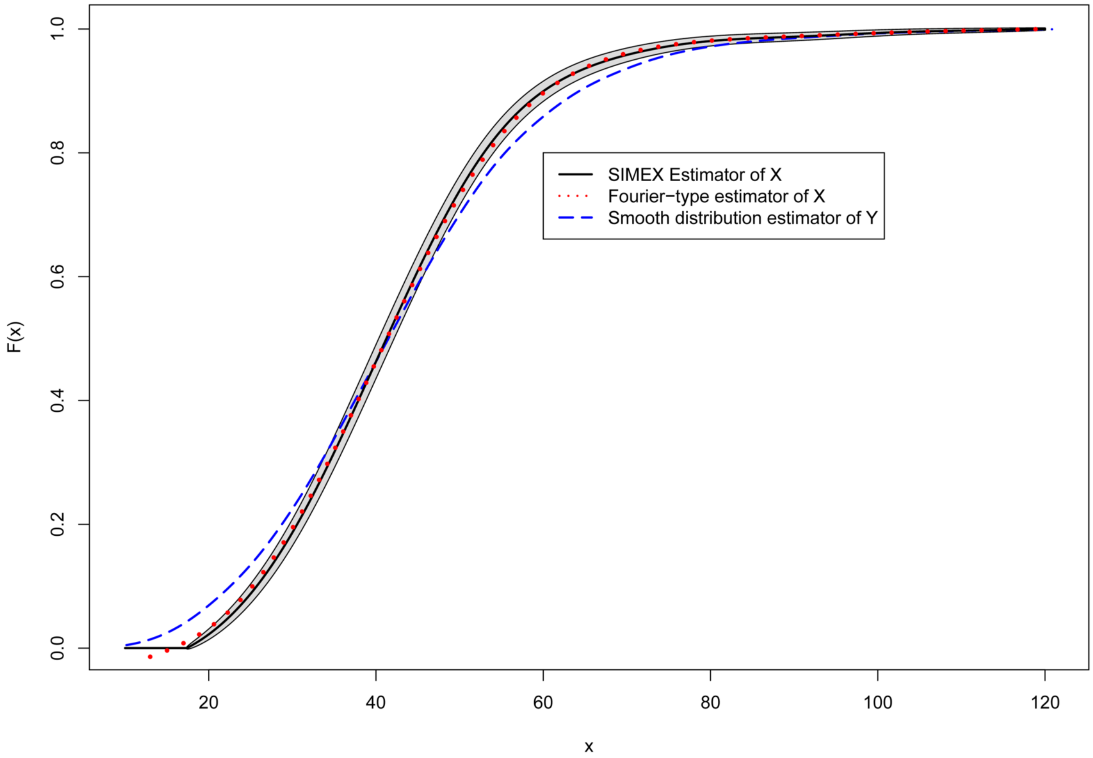

Figure 4.

Smooth distribution estimation of pressure intensities in the NMES study. Three estimators are considered here: the SIMEX estimator (solid line); the Fourier-type estimator (dotted line); the Naive estimator, *i.e.* ignoring the measurement errors (dashed line). The gray region is the 95% confidence region of the SIMEX estimator.

Table 1

The average ISEs for the case of heteroscedastic Gaussian errors in the simulation study.

			ISE			
Error	**True dist**[*]	**Sample size**	**Fourier**	**SIMEX**	**Naive**	$\hat{F}_X(x)$
$\sigma_U \sim U(0.4, 0.6)$	Normal	50	0.0119	0.0127	0.0151	0.0095
		100	0.0056	0.0058	0.0086	0.0045
		500	0.0018	0.0016	0.0043	0.0010
	Mixture	50	0.0124	0.0119	0.0215	0.0103
		100	0.0078	0.0074	0.0171	0.0061
		500	0.0032	0.0031	0.0091	0.0014
	Gamma	50	0.0184	0.0157	0.0282	0.0112
		100	0.0107	0.0090	0.0218	0.0057
		500	0.0039	0.0041	0.0158	0.0016
$\sigma_U \sim U(0.8, 1)$	Normal	50	0.0241	0.0228	0.0291	0.0091
		100	0.0165	0.0138	0.0257	0.0044
		500	0.0073	0.0059	0.0166	0.0011
	Mixture	50	0.0252	0.0218	0.0316	0.0102
		100	0.0171	0.0147	0.0286	0.0059
		500	0.0089	0.0069	0.0197	0.0018
	Gamma	50	0.0262	0.0230	0.0395	0.0109
		100	0.0193	0.0187	0.0303	0.0062
		500	0.0097	0.0072	0.0211	0.0017

[*] Normal = $N(0, 1)$; Mixture = $0.5N(-3, 1)+0.5N(3, 1)$; Gamma = $\Gamma(2, 1)$.