Potgieter *Journal of Statistical Distributions and Applications* (2020) 7:2
https://doi.org/10.1186/s40488-020-00103-y

Journal of Statistical Distributions and Applications

RESEARCH Open Access

Density deconvolution for generalized skew-symmetric distributions

Cornelis J. Potgieter[1,2]

Correspondence:
c.potgieter@tcu.edu
[1]Department of Mathematics, Texas Christian University, Forth Worth, TX, USA
[2]Department of Statistics, University of Johannesburg, Johannesburg, South Africa

Abstract

The density deconvolution problem is considered for random variables assumed to belong to the generalized skew-symmetric (GSS) family of distributions. The approach is semiparametric in that the symmetric component of the GSS distribution is assumed known, and the skewing function capturing deviation from the symmetric component is estimated using a deconvolution kernel approach. This requires the specification of a bandwidth parameter. The mean integrated square error (MISE) of the GSS deconvolution estimator is derived, and two bandwidth estimation methods based on approximating the MISE are also proposed. A generalized method of moments approach is also developed for estimation of the underlying GSS location and scale parameters. Simulation study results are presented including a comparing the GSS approach to the nonparametric deconvolution estimator. For most simulation settings considered, the GSS estimator is seen to have performance superior to the nonparametric estimator.

Keywords: Characteristic function, Kernel methods, Measurement error, Method of moments, Semiparametric estimation

Introduction

The density deconvolution problem arises when it is of interest to estimate the probability density function (pdf) $f_x(x)$ of a random variable X using observations contaminated by measurement error. Specifically, the observed sample consists of data $W_j = X_j + U_j$, $j = 1, \ldots, n$, where the X_j are independent and identically distributed (*iid*) random variables with pdf $f_x(x)$ and the U_j are *iid* measurement error variables with pdf $f_u(u)$. This paper presents a semiparametric approach for estimating $f_x(x)$ that assumes X belongs to the class of generalized skew-symmetric (GSS) distributions. The GSS deconvolution model for X specifies a base symmetric distribution, providing the basic structure for the model. Thereafer, kernel methodology is used to estimate a skewing function that captures the deviation from the specified symmetric distribution. This semiparametric GSS approach attempts to capture the best of a parametric and a nonparametric solution and provides a very flexible approach for modeling $f_x(x)$.

The problem of estimating $f_x(x)$ from a contaminated sample $W_1, \ldots, W_n$ was first considered by Carroll and Hall (1988) and Stefanski and Carroll (1990) who proposed a

fully nonparametric solution under the assumption of a fulle known measurement error distribution $f_u(u)$. Since then, much work on the topic has followed. Fan (1991a); Fan (1991b) considered the theoretical properties of the density deconvolution estimator, and Fan and Truong (1993) extended the methodology to nonparametric regression. Diggle and Hall (1993) and Neumann and Hössjer (1997) considered the case of the measurement error distribution being unknown, and assumed that an external sample of error data was available to estimate the measurement error distribution. Delaigle et al. (2008) considered how replicate data can be used to estimate the characteristic function of the measurement error. The nonparametric estimator requires the selection of a bandwidth parameter. The two-stage plug-in bandwidth of Delaigle and Gijbels (2002) has become the gold-standard in application; Delaigle and Gijbels (2004) provides an overview of several popular bandwidth selection approaches. Delaigle and Hall (2008) considered the use of simulation-extrapolation (SIMEX) for bandwidth selection in a variety of measurement error problems.

Two more recent papers considered the deconvolution problem in new and novel ways. Delaigle and Hall (2014) considered parametrically-assisted nonparametric density deconvolution, while the groundbreaking work of Delaigle and Hall (2016) made use of the empirical phase function to estimate the pdf $f_x(x)$ with the measurement error having unknown distribution and without the need for replicate data. The phase function approach imposes the restrictions that X has no symmetric component and that the characteristic function of U is real-valued and strictly positive.

The GSS family of distributions that is the basis for estimation in this paper dates back to Azzalini (1985), the first publication discussing a so-called *skew-normal* distribution. There has been a great deal of activity since with the monographs by Genton (2004) and Azzalini (2013) providing a good overview of the existing literature on the topic. Much of the GSS research has been theoretical in nature. While this theoretical work is important for understanding the statistical properties of GSS distributions, the applied value of this family has not often been realized in the literature. Notable exceptions that have used GSS distributions in application include the modeling of pharmacokinetic data, see Chu et al. (2001), the redistribution of soil in tillage, see Van Oost et al. (2003), and the retrospective analysis of case-control studies, see Guolo (2008). All of these authors considered fully parametric models. Arellano-Valle et al. (2005) considered a fully parametric measurement error model assuming both X and U follow skew-normal distributions. Lachos et al. (2010) modeled X using a scale-mixture of skew-normal distributions while assuming U is a mixture of normals. Furthermore, both Kim et al. (2016) and Wang et al. (2017) consider factor analysis models using skew-symmetric distributions. Most recently, Kahrari et al. (2019) developed linear mixed models using a skew-normal-Cauchy distribution and Arellano-Valle et al. (2020) considered the measurement error problem using a two-piece normal distribution to allow for skewness. No other work applying GSS distributions in the measurement error context was found.

The present paper is structured as follows. In the next section, the GSS deconvolution estimator is developed and some of its theoretical properties derived. In the subsequent section, bandwidth estimation methods for the skewing function are considered. Thereafer, a generalized method of moments (GMM) approach for estimating the GSS location and scale parameters is developed. The penultimate section presents simultion results,

and the paper concludes with two real-data applications. An Appendix contains both some technical arguments and additional simulation results.

Generalized skew-symmetric deconvolution

Derivation of the GSS estimator

Consider the problem of estimating the probability density function (pdf) $f_x(x)$ associated with random variable X based on a sample contaminated by additive measurement error, $W_j = X_j + U_j, j = 1, \ldots, n$. Here, the X_j are the true measurements of interest, and the W_j and U_j represent, respectively, the contaminated observation and the measurement error. It is assumed that the X_j are *iid* $f_x(x)$, the U_j are *iid* $f_u(u)$, and X_j and U_j are mutually independent for all j. Furthermore, the U_j are assumed to have a symmetric distribution with mean 0 and variance σ_u^2. As is typical in the deconvolution literature, the distribution of U_j is assumed fully known. Auxiliary data, when available, would make it possible to relax this assumption and estimate $f_u(u)$; see for example Delaigle et al. (2008).

The deconvolution estimator developed here assumes that $f_x(x)$ belongs to the GSS class of distributions. That is, $X = \xi + \omega Z$ with $\xi \in \mathbb{R}$ and $\omega > 0$ denoting location and scale parameters, and with Z having pdf

$$f_z(z) = 2f_0(z)\pi(z), z \in \mathbb{R} \tag{1}$$

with $f_0(z)$ a pdf symmetric around 0 and $\pi(z)$, hereafter referred to as the skewing function, satisfying the inequality constraint $0 \leq \pi(z) = 1 - \pi(z) \leq 1$. In fact, any function satisfying this inequality constraint can be paired with any symmetric pdf $f_0(z)$ and will result in (1) being a valid pdf. The corresponding pdf of X is $f_x(x) = (2/\omega)f_0[(x - \xi)/\omega]\,\pi[(x - \xi)/\omega]$.

The approach considered here is semiparametric in nature. The symmetric pdf $f_0(z)$ is assumed known, but no parametric assumptions are made regarding the skewing function $\pi(z)$. (In fact, if symmetric component $f_0(z)$ were not assumed known, pdf $f_z(z)$ would not be identifiable; see Appendix A.1 for details). The base density $f_0(z)$ provides the basic strucuture of the model, and the skewing function $\pi(z)$ captures the deviation from the base model. Thus, the approach attempts to capture the best of a parametric and a nonparametric solution, and the GSS family provides a very flexible approach for modeling $f_z(z)$.

GSS random variables have an invariance property under even transformations that is central to the development of the deconvolution estimator in the remainder of this section. Let Z be GSS according to (1) and let Z_0 have symmetric pdf $f_0(z)$. For any even function $t(z)$, it holds that $t(Z) \stackrel{d}{=} t(Z_0)$ with $\stackrel{d}{=}$ denoting equality in distribution; see Proposition 1.4 in Azzalini (2013). Thus, the distribution of $t(Z)$ depends only on $f_0(z)$ and not on $\pi(z)$. Now, let $\psi_z(t)$ denote the characteristic function of Z, and let $c_0(t) = \mathrm{Re}[\psi_z(t)]$ and $s_0(t) = \mathrm{Im}[\psi_z(t)]$ denote the real and imaginary components of $\psi_z(t)$. The real component can be expressed as $c_0(t) = \mathrm{E}[\cos(tZ)]$. By the property of even transformation, it follows that $c_0(t) = \mathrm{E}[\cos(tZ_0)]$ which is the characteristic function associated with $f_0(z)$.

Now, assume (ξ, ω) are known, and define $W^* = (W - \xi)/\omega$. Furthermore, observe that $W^* = Z + \omega^{-1}U$and therefore has characteristic function $\psi_{W^*}(t) = \psi_z(t/\omega)\psi_u(t)$ where $\psi_u(t)$ is the real-valued characteristic function of U. It follows that

$$\mathrm{Re}\{\psi_{w^*}(t)\} = c_0(t)\psi_u(t/\omega) \tag{2}$$

and

$$\mathrm{Im}\{\psi_{w^*}(t)\} = s_0(t)\psi_u(t/\omega). \tag{3}$$

The functions $c_0(t)$ and $\psi_u(t)$ in (2) and (3) are known while $s_0(t)$ is unknown. Noting that $f_z(z)$ can be expressed as

$$f_z(z) = f_0(z) + \frac{1}{2\pi}\int_{\mathbb{R}} \sin(tz)s_0(t)dt, \tag{4}$$

it follows that an estimator of $s_0(t)$ can be used to construct an estimator of $f_z(z)$. To this end, for random sample $W_1, \ldots, W_n$, let $W_j^* = (W_j - \xi)/\omega$ for $j = 1, \ldots, n$, and define

$$\tilde{s}_0(t) = \frac{1}{\psi_u(t/\omega)}\frac{1}{n}\sum_{1\le j\le n} \sin\left(tW_j^*\right).$$

This empirical estimator, while unbiased for $s_0(t)$, is not suitable for estimating $f_z(z)$ when substituted in (4) as the integral diverges. This is attributable to the tail behavior of $\tilde{s}_0(t)$. While $s_0(t)$ converges to 0 as $|t| \to \infty$ for any continuous distribution, $\tilde{s}_0(t)$ corresponds to an empirical measure and diverges as $|t| \to \infty$. This follows upon noting that the bounded periodic function $n^{-1}\sum_j \sin(tW_j^*)$ is divided by $\psi_u(t/\omega)$, with the latter decreasing to 0 as $|t|$ increases.

Next, consider the "smoothed" estimator

$$\hat{s}_0(t) = \frac{\psi_k(ht)}{\psi_u(t/\omega)}\frac{1}{n}\sum_{1\le j\le n} \sin\left(tW_j^*\right) \tag{5}$$

where $\psi_k(t)$ is a non-negative weight function and h is a bandwidth parameter. This estimator has expectation $\mathrm{E}[\hat{s}_0(t)] = \psi_k(ht)s_0(t)$ and therefore is biased for $s_0(t)$. However, it also has some desirable properties. Firstly, it is an odd function, $\hat{s}_0(-t) = -\hat{s}_0(t)$ for all $t \in \mathbb{R}$. Secondly, substitution of (5) into (4) results in the well-defined estimator for $f_z(z)$,

$$\hat{f}_z(z) = f_0(z) + \frac{1}{2\pi}\int_{\mathbb{R}} \sin(tz)\hat{s}_0(t)dt, \tag{6}$$

provided $\psi_k(t)$ is chosen such that $|\psi_k(ht)/\psi_u(t/\omega)| \to 0$ as $|t| \to \infty$. Choosing $\psi_k(t)$ to be 0 outside a bounded interval will trivially satisfy this requirement.

Estimator (6) suffers from the same drawback as the usual nonparametric deconvolution estimator in that it may be negative in parts. In practice, the negative parts can be truncated and the resulting function rescaled to integrate to 1. To circumvent this ad-hoc fix, combine Eqs. (1) and (4) to obtain

$$\pi(z) = \frac{1}{2} - \frac{1}{4\pi f_0(z)}\int_R \sin(tz)\, s_0(t)dt. \tag{7}$$

Substitution of (5) in (7), along with the identity $\sin(tz) = \left(e^{itz} - e^{-itz}\right)/(2i)$, gives

$$\hat{\pi}(z) = \frac{1}{2} + \frac{1}{8f_0(z)}\left\{\tilde{f}_{w^*}(z) - \tilde{f}_{w^*}(-z)\right\} \tag{8}$$

where $\tilde{f}_{w^*}(z) = (nh\omega)^{-1}\sum K_{h\omega}[(z - W_j^*)/(h\omega)]$ is the well-studied nonparametric deconvolution density estimator of Carroll and Hall (1988) with deconvolution kernel $K_h(y) = (2\pi)^{-1}\int_{\mathbb{R}} e^{-ity}\psi_k(t)/\psi_u(t/h)dt$. The potential for (6) being negative in parts is reflected in (8) not being range-respecting. Specifically, it is possible to have $\hat{\pi}(z) \notin [0, 1]$ for a set z with nonzero measure. A range-corrected skewing function estimator is $\tilde{\pi}(z) =$

$\max\left[0, \min\left\{1, \hat{\pi}\ (z)\right\}\right]$. The estimated density function of X based on the range-corrected skewing function is

$$\tilde{f}\ (x|\xi,\omega) = \frac{1}{\omega} f_0\left(\frac{x-\xi}{\omega}\right)\tilde{\pi}\left(\frac{x-\xi}{\omega}\right). \tag{9}$$

Use of the range-corrected skewing function estimate ensures that (9) is always a valid pdf. There is no need for any additional truncation of negative values and subsequent rescaling as would be the case with direct implementation of (6).

Some properties of the estimator

The range-corrected estimator $\tilde{\pi}(z)$ is asymptotically equivalent to $\hat{\pi}(z)$ in (8) on any closed subset of $\mathbb{R}$. As such, the latter will be used to evaluate the properties of the GSS deconvolution estimator. Firstly, note that using the known expected value of the nonparametric deconvolution estimator $\tilde{f}_{w^*}(z)$, it follows from (8) that

$$E\left[\hat{\pi}(z)\right] - \pi(z) = \frac{c_k}{4}\frac{f_z''(z) - f_z''(-z)}{f_0(z)} \cdot h^2 + O\left(h^3\right)$$

with constant c_k depending only on the kernel function $\psi_k(t)$. Thus, for an appropriately chosen bandwidth h, $\hat{\pi}(z)$ is consistent for $\pi(z)$, and the density estimator $\tilde{f}(x|\xi,\omega)$ in (9) is also consistent for $f_x(x)$.

The mean integrated square error (MISE), derived in Appendix A.2, is

$$\text{MISE}(h) = (2\pi)^{-1}\int_{\mathbb{R}}\left\{\frac{\psi_k^2(ht)}{n}\left[\frac{1 - c_0(2t)\psi_u(2t/\omega)}{2\psi_u^2(t/\omega)} - s_0^2(t)\right] + \left[\psi_k(ht) - 1\right]^2 s_0^2(t)\right\} dt. \tag{10}$$

When the distribution Z is symmetric, i.e. $\pi(z) = 1/2$ for all z so that $s_0(t) = 0$ for all t, and letting MISE_{sym} denotes the MISE calculated under symmetry,

$$\begin{aligned}\text{MISE}_{\text{sym}}(h) &= (4\pi)^{-1}\int_{\mathbb{R}}\frac{\psi_k^2(ht)}{n}\left[\frac{1 - c_0(2t)\psi_u(2t/\omega)}{\psi_u^2(t/\omega)}\right] dt \\ &\leq (2\pi n)^{-1}\int_{\mathbb{R}}\frac{\psi_k^2(ht)}{\psi_u^2(t/\omega)} dt.\end{aligned}$$

Here the inequality follows upon noting that $|1 - c_0(2t)\psi_u(2t/\omega)| \leq 2$ for all t. This upper bound of MISE_{sym} is proportional to the asymptotic MISE of the nonparametric deconvolution estimator, see equation (2.7) in Stefanski & Carroll (1990). Thus, in the symmetric case, one would expect the GSS deconvolution estimator to perform no worse than the nonparametric deconvolution estimator for a correctly specified symmetric component $c_0(t)$. In fact, since this is an upper bound, large gains in efficiency may be possible. Our simulation results presented in a later section are congruent with this statement.

Bandwidth selection

Implementation of the GSS deconvolution estimator requires a bandwidth paramter h to be specified. Two methods for selecting this bandwidth are developed in this section. The first method uses cross-validation (CV) to approximate the integrated square error (ISE), and the second method approximates the MISE in (10).

A cross-validation bandwidth

For the GSS deconvolution estimator, the density-based ISE is proportional to the ISE for the imaginary component $s_0(t)$ of the characteristic function,

$$\int_{\mathbb{R}} \left[\tilde{f}_z(z) - f_z(z)\right]^2 dz \propto \int_{\mathbb{R}} \left[\hat{s}_0(t) - s_0(t)\right]^2 dt. \tag{11}$$

This follows from Parseval's identity and recalling that the real component $c_0(t)$ is known. Let $C(h)$ denote the expression obtained by expanding the square on the right-hand side of (11) and keeping only terms involving the estimator $\hat{s}_0(t)$,

$$C(h) = \int_{\mathbb{R}} \hat{s}_0^2(t)dt - 2\int_{\mathbb{R}} \hat{s}_0(t)s_0(t)dt. \tag{12}$$

Now, note that the second integral in (12) can be written as

$$\int_{\mathbb{R}} \hat{s}_0(t)s_0(t)dt = \sum_{i=1}^{n} \int_{\mathbb{R}} \frac{\psi_k(ht)\sin(tW_i^*)}{\psi_u(t/\omega)} s_0(t)dt. \tag{13}$$

Define $\tilde{s}_{(i)}(t)$ to be an estimate of $s_0(t)$ excluding the ith observation,

$$\tilde{s}_{(i)}(t) = \frac{(n-1)^{-1}\sum_{j\neq i}\sin(tW_j^*)}{\psi_u(t/\omega)}.$$

This quantity is unbiased for $s_0(t)$ for all i, and $\tilde{s}_{(i)}(t)$ is independent of W_i. The CV score follows by substitution of $\tilde{s}_{(i)}(t)$ in (13) for each i in the summand, giving

$$\hat{C}(h) = \int_{\mathbb{R}} \frac{\psi_k(ht)}{\psi_u^2(t/\omega)} \left[\psi_k(ht)\left\{\frac{1}{n}\sum_{j=1}^{n}\sin(tW_j^*)\right\}^2 - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sin(tW_i^*)\sin(tW_j^*)\right]. \tag{14}$$

This result is similar to that of Stefanski and Carroll (1990) in the nonparametic setting, but here only requires estimating the imaginary component of the characteristic function. The CV bandwidth is defined to be the value $\tilde{h}$ that minimizes $\hat{C}(h)$.

An MISE bandwidth

Consider the MISE in (10), and note that the only unknown quantity therein is $s_0^2(t)$. Furthermore, observe that $\mathrm{E}\left[\sin(tW_j^*)\sin(tW_k^*)\right] = \psi_u^2(t/\omega)s_0^2(t)$ whenever $j \neq k$. Thus, $s_0^2(t)$ can be estimated by

$$\hat{s}_2(t) = \max\left\{0, \frac{1}{n(n-1)\psi_u^2(t/\omega)}\sum_{j=1}^{n}\sum_{k\neq j}\sin(tW_j^*)\sin(tW_k^*)\right\}\mathcal{I}(|t|\leq\kappa), \tag{15}$$

where $\mathcal{I}(\cdot)$ is the indicator function and κ is some positive constant. The constant κ can be thought of as a smoothing parameter which ensures that the estimator $\hat{s}_2(t)$ behaves well for large values of $|t|$. Ideally, κ should be chosen in a data-dependent way and development of this approach is ongoing. However, based on extensive simulation work, it has been found that values $\kappa \in [3, 5]$ work reasonably well for a wide range of underlying

GSS distributions considered. Now, taking (10), substituting $\hat{s}_2(t)$ for $s_0^2(t)$, and ignoring components that do not depend on the bandwidth, gives MISE approximation score

$$\hat{M}(h) = \frac{1}{h}\int_{\mathbb{R}} \left\{ \frac{\psi_k^2(t)}{n\psi_u^2[\,t/(h\omega)]} \left[\frac{1-\psi_u[\,2t/(h\omega)]\,c_0(2t/h)}{2}\right] + \left[\frac{n-1}{n}\psi_k(t) - 2\right]\psi_k(t)\hat{s}_2(t/h)\right\} dt. \tag{16}$$

The MISE-approximation bandwidth is defined to be the value $\tilde{h}$ that minimizes $\hat{M}(h)$.

Location and scale estimation

Generalized method of moments

Up to this point, the location and scale parameters ξ and ω have been treated as known quantities. This is unrealistic in practice. Estimation of the GSS parameters for a known symmetric component has been considered in the literature, see Ma et al. (2005); Azzalini et al. (2010), and Potgieter and Genton (2013). However, none of these authors considered the presence of measurement error. Here, a Generalized Method of Moments (GMM) approach accounting for measurement error is developed. Recall that $W_j = X_j + U_j = \xi + \omega Z_j + U_j, j = 1,\ldots,n$. Let $M \geq 2$ be a positive integer and assume that the Z_j and the U_j have at least $2M$ finite moments. Let T_k denote the $(2k)$th centered moment,

$$T_k := T_k(\xi,\omega) = n^{-1}\sum_{j=1}^{n}\left(\frac{W_j-\xi}{\omega}\right)^{2k}. \tag{17}$$

This variable has expectation $\mathrm{E}\,[T_k] = \mathrm{E}\left[\left(Z+\omega^{-1}U\right)^{2k}\right]$ and admits expansion

$$\mathrm{E}\,[T_k] = \sum_{j=0}^{k}\binom{2k}{2j}\omega^{-2(k-j)}\mathrm{E}\left[Z^{2j}\right]\mathrm{E}\left[U^{2(k-j)}\right]. \tag{18}$$

By the GSS property of even transformations, $\mathrm{E}[\,Z^{2j}] = \mathrm{E}[\,Z_0^{2j}]$ for $j = 1,\ldots,M$ with Z_0 a random variable with pdf $f_0(z)$. Furthermore, the evaluation of the moments of U pose no problem as this distribution is assumed known. Thus, $E[\,T_k]$ can easily be evaluated using (18).

Now, define quadratic form $D(\xi,\omega) = n\mathbf{T}_M^\top\Sigma^{-1}\mathbf{T}_M$ with $\mathbf{T}_M$ denoting the vector $\mathbf{T}_M = (T_1 - \mathrm{E}[\,T_1],\ldots,T_M - \mathrm{E}[\,T_M])^\top$ with covariance matrix Σ. The covariance matrix has entries $\Sigma_{ij} = n^{-1}\left(\mathrm{E}\left[T_{i+j}\right] - \mathrm{E}\,[T_i]\,\mathrm{E}\left[T_j\right]\right)$. The GMM estimators are defined to be the minimizer of $D(\xi,\omega)$. In evaluating $D(\xi,\omega)$, both the expectations $\mathrm{E}\,[T_k]$, $k = 1,\ldots,M$ and the covariance matrix Σ are functions of the parameter ω, but not of ξ.

Selection from multiple GMM solutions

One difficulty encountered with the GMM approach is that the statistic $D(\xi,\omega)$ frequently has multiple minima, and the global minimum does not always corresponds to the "correct" solution. This equivalent problem also occurs in the non-measurement error setting and is an artifact of the skewing function being unknown; see Section 7.2.2 in Azzalini (2013) for an overview and illustration. Solutions considered there range from selecting the model with the smallest squared integral of the second derivative of the estimated skewing function, to selecting a solution based on matching model-based and empirical skewness coefficients.

Now, assume that $D(\xi, \omega)$ has J local minima occurring at $(\hat{\xi}_j, \hat{\omega}_j)$, $j = 1, \ldots, J$. Furthermore, let $\tilde{f}_j(x|\hat{\xi}_j, \hat{\omega}_j)$ denote the GSS density deconvolution estimator in (9) obtained using solution $(\hat{\xi}_j, \hat{\omega}_j)$. Thus, J different GSS deconvolution estimators are calculated. Using the jth estimated density, define the kth model-implied moment,

$$\tilde{\mu}_{j,k} = \int_{\mathbb{R}} x^k \tilde{f}_j(x|\hat{\xi}_j, \hat{\omega}_j) dx, \tag{19}$$

and model-implied characteristic function,

$$\tilde{\phi}_j(t) = \int_{\mathbb{R}} \exp(itx) \tilde{f}_j(x|\hat{\xi}_j, \hat{\omega}_j) dx. \tag{20}$$

Based on these quantities, two different selection methods are now proposed. Throughout, it will be assumed that measurement error U has distribution symmetric about 0.

Skewness matching: In model $W = X + U$, the skewness of X can be estimated by $\hat{\gamma}_x = \left[\hat{\sigma}_w^2/(\hat{\sigma}_w^2 - \sigma_u^2)^{3/2}\right] \hat{\gamma}_w$ where $\hat{\sigma}_w^2$ and $\hat{\gamma}_w$ denote the sample variance and skewness of *iid* random variables $W_1, \ldots, W_n$. Now, for the jth solution pair $(\hat{\xi}_j, \hat{\omega}_j)$, the GSS model-implied skewness is given by $\hat{\gamma}_j = \left(\tilde{\mu}_{j,3} - 3\tilde{\mu}_{j,2}\tilde{\mu}_{j,1} + 2\tilde{\mu}_{j,1}^3\right) / \left(\tilde{\mu}_{j,2} - \tilde{\mu}_{j,1}^2\right)$ with $\tilde{\mu}_{j,k}$ as defined in (19). The selected solution is the one with implied skewness closest to the empirical skewness. Specifically, letting $d_j = |\hat{\gamma}_x - \hat{\gamma}_j|$, $j = 1, \ldots, J$, the selected solution is $(\hat{\xi}_{j^*}, \hat{\omega}_{j^*})$ with $j^* = \arg\min_{1 \le j \le J} d_j$.

Phase function matching: The phase function, a normalized version of the characteristic function, is a recent tool employed in density deconvolution – see Delaigle and Hall (2016) and Nghiem and Potgieter (2018) for further details. Let $\rho_w(t)$ and $\rho_x(t)$, denote the phase functions of X and $W = X + U$. For U having strictly positive characteristic function, these phase functions are equal, $\rho_w(t) = \rho_x(t)$ for all t. The empirical estimate of the phase function of X is $\hat{\rho}_x(t) = \hat{\psi}_w(t)/|\hat{\psi}_w(t)|$ with $\hat{\psi}_w(t)$ the empirical characteristic function of W, and $|z| = (z\bar{z})^{1/2}$ and $\bar{z}$ denoting the complex norm and cojucate of z. For the jth GMM solution $(\hat{\xi}_j, \hat{\omega}_j)$, the model-implied phase function is given by $\tilde{\rho}_j(t) = \tilde{\phi}_j(t)/|\tilde{\phi}_j(t)|$ with $\tilde{\phi}_j(t)$ as defined in (20). Now, letting $w(t)$ denote a non-negative weight function symmetric around 0, define distance metric $R_j = \int_{\mathbb{R}} |\hat{\rho}_x(t) - \tilde{\rho}_j(t)| w(t) dt$ for $j = 1, \ldots, J$. The selection solution is $(\hat{\xi}_{j^*}, \hat{\omega}_{j^*})$ with $j^* = \arg\min_{1 \le j \le J} R_j$. That is, the selected solution has minimum phase function distance. In this paper, weight function $w(t) = [1 - (t/t^*)^2]^3 \; \mathcal{I}(|t| \le t^*)$ will be used with t^* the smallest $t > 0$ such that $|\hat{\psi}_w(t)| \le n^{-1/4}$ as per Delaigle and Hall (2016).

Simulation studies

The performance of the GSS deconvolution estimator was evaluated using extensive simulations. Letting $\phi(z)$ and $\Phi(z)$ denote the standard normal density and distribution functions, data $X_1, \ldots, X_n$ were generated from GSS distributions with symmetric component $f_0(z) = \phi(z)$ and using three different skewing functions, $\pi_0(z) = 1/2$, $\pi_1(z) = \Phi(9.9625z)$ and $\pi_2(z) = \Phi(z^3 - 2z)$. The location and scale parameters were taken to be $\xi = 0$ and $\omega = 1$. Figure 1 illustrates the three resulting pdfs $f_x(x) = (2/\omega)\phi[(x - \xi)/\omega]\, \pi_k[(x - \xi)/\omega]$, $k = 0, 1, 2$. Note that the skewing function $\pi_0(z)$ does not introduce any deviation from symmetry and corresponds to simulating from a normal distribution. Additionally, the skewing function $\pi_1(z)$ results in a positive skew distribution, while $\pi_2(z)$ results in a bimodal distribution.

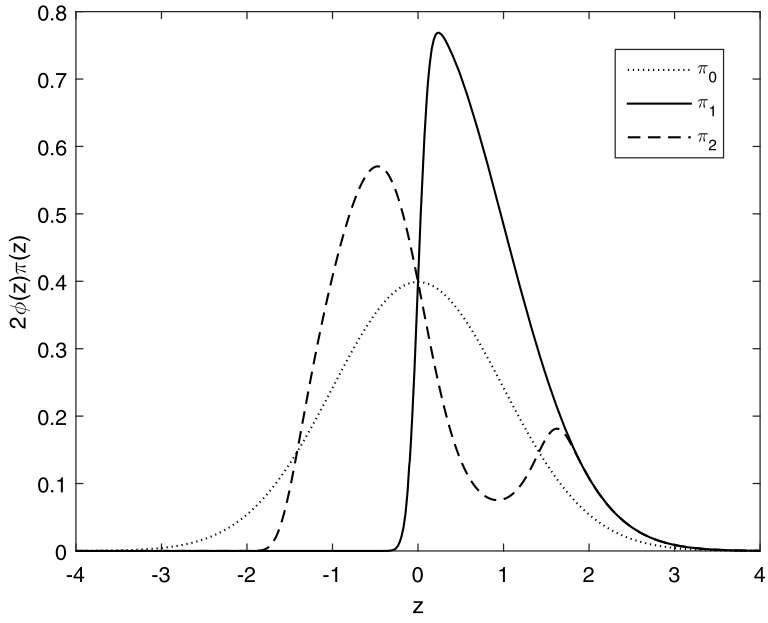

Fig. 1 Skew-symmetric densities used in simulation study

Two measurement error distributions were considered with $U_1, \ldots, U_n$ being either Normal or Laplace with mean 0 and variance chosen to have noise-to-signal ratio NSR $= \sigma_u^2/\sigma_x^2$ either 0.2 or 0.5. Samples $W_j = X_j + U_j$, $j = 1, \ldots, n$, with $n \in \{50, 100, 200, 500\}$ were generated from each of the possible simulation configurations described.

Comparison of oracle estimators

The first simulation study presented compares the proposed GSS estimator to the established nonparametric estimator of Carroll and Hall (1988), and assumes the existence of an *oracle* that selects the "best" possible bandwidth for each of the estimators. Specifically, for a sample $W_1, \ldots, W_n$, let $\tilde{f}_{\text{gss}}(x|h)$ and $\tilde{f}_{\text{np}}(x|h)$ denote, respectively, the GSS and nonparametric estimators with bandwidth h. The ISE is defined as

$$\text{ISE}_{\text{m}}(h) = \int_{\mathbb{R}} \left[\tilde{f}_{\text{m}}(x|h) - f_x(x)\right]^2 dx$$

where m $\in$ {gss, np}. Then, the "best" bandwidth is the value that minimizes the ISE between the estimated and true densities. Furthermore, when GMM results in more than one solution for the GSS location and scale parameters, the oracle also selects the solution that result in smallest ISE. In practice, no oracle exists to do these selections. Even so, comparing the estimators under such idealized conditions speaks to the best possible performance of these methods.

For each simulation configuration, $N = 1000$ samples were generated. Due to the occasional occurrence of very large outliers in ISE, the median ISE (rather than mean ISE) is reported. The first and third quartiles of ISE are also reported. Results for $n \in \{200, 500\}$ are summarized in Table 1, and for $n \in \{50, 100\}$ are presented in Table 6 in Appendix A.5.

Inspection of Table 1 shows how well the GSS estimator can perform relative to the nonparametric estimator. In the symmetric case with skewing function $\pi_0(z)$, the reduction in median ISE is most dramatic and exceeds 50% in all cases. For skewing functions

Table 1 Median of 100 × ISE, as well as first and third quartiles $[Q_1, Q_3]$ for the oracle GSS and nonparametric (NP) deconvolution estimators

		$n = 200$		$n = 500$	
π	(NSR, U)	GSS	NP	GSS	NP
π_0	(0.2, N)	0.131	0.442	0.070	0.282
		[0.055, 0.263]	[0.256, 0.709]	[0.032, 0.148]	[0.186, 0.418]
	(0.5, N)	0.199	0.817	0.122	0.596
		[0.084, 0.405]	[0.532, 1.228]	[0.048, 0.296]	[0.409, 0.845]
	(0.2, L)	0.113	0.273	0.058	0.147
		[0.053, 0.241]	[0.140, 0.476]	[0.027, 0.117]	[0.079, 0.236]
	(0.5, L)	0.148	0.327	0.076	0.169
		[0.074, 0.323]	[0.165, 0.603]	[0.040, 0.158]	[0.086, 0.308]
π_1	(0.2, N)	1.690	2.453	1.400	1.875
		[1.271, 2.188]	[1.855, 3.173]	[1.031, 1.775]	[1.434, 2.419]
	(0.5, N)	2.277	4.079	2.034	3.514
		[1.729, 2.956]	[3.116, 5.275]	[1.547, 2.645]	[2.716, 4.352]
	(0.2, L)	1.200	1.701	0.712	1.096
		[0.832, 1.658]	[1.223, 2.258]	[0.422, 1.112]	[0.818, 1.463]
	(0.5, L)	1.542	2.353	1.025	1.615
		[1.054, 2.162]	[1.671, 3.176]	[0.652, 1.469]	[1.206, 2.105]
π_2	(0.2, N)	1.410	1.768	1.004	1.289
		[0.918, 2.082]	[1.251, 2.465]	[0.689, 1.406]	[0.971, 1.719]
	(0.5, N)	3.068	3.896	2.483	3.153
		[1.976, 4.542]	[2.731, 5.241]	[1.602, 3.504]	[2.302, 4.174]
	(0.2, L)	0.638	0.754	0.315	0.434
		[0.358, 1.060]	[0.494, 1.250]	[0.190, 0.515]	[0.272, 0.650]
	(0.5, L)	1.413	1.310	0.667	0.707
		[0.728, 2.472]	[0.763, 2.112]	[0.381, 1.199]	[0.439, 1.114]

$\pi_1(z)$ and $\pi_2(z)$, the reduction in median ISE is also seen to be as large as 40%. There is one instance where median ISE of the nonparametric estimator is smaller than that of the GSS estimator – skewing function $\pi_2(z)$ with NSR $= 0.5$, Laplace measurement error, and sample size $n = 200$. (The same holds true for sample sizes $n = 50$ and 100 in Table 6.) However, the equivalent scenario with sample size $n = 500$ has the GSS estimator with smaller median ISE. This possibly indicates the effect of estimating the location and scale parameters in smaller samples and when large amounts of heavier-tailed-than-normal measurement error is present. Overall, the GSS deconvolution estimator performs very well. Thus, the additional structure being imposed through the a priori specification of the symmetric pdf $f_0(z)$ can result in a large decrease in ISE.

Bandwidth estimation

The next simulation study investigated the two proposed bandwidth estimation approaches. Specifically, the CV and MISE bandwidths as well as the two-stage plug-in (PI) bandwidth of Delaigle and Gijbels (2002), originally developed for nonparametric deconvolution, were implemented. For each simulated sample, the ISE was calculated. When necessary, GMM solution selection with phase-function matching was used. The nonparametric deconvolution estimator with PI bandwidth was also calculated; corresponding results are included for reference purposes. The median ISE values for the

Table 2 Median of 100 × ISE for the GSS deconvolution estimators with CV, MISE, and PI bandwidths, and the nonparametric (NP) estimator with PI bandwidth. Sample size $n = 200$

π	(NSR, U)	CV	MISE	PI	NP
π_0	$(0.2, N)$	0.409	0.370	0.294	0.535
	$(0.5, N)$	0.652	0.701	0.492	1.039
	$(0.2, L)$	0.409	0.407	0.299	0.433
	$(0.5, L)$	0.574	0.630	0.435	0.653
π_1	$(0.2, N)$	2.217	2.116	2.399	2.709
	$(0.5, N)$	3.193	3.032	3.983	4.601
	$(0.2, L)$	1.645	1.494	1.712	1.998
	$(0.5, L)$	2.299	2.116	2.274	2.848
π_2	$(0.2, N)$	2.138	1.593	1.755	1.956
	$(0.5, N)$	4.648	4.175	3.785	4.375
	$(0.2, L)$	1.359	1.230	0.894	1.044
	$(0.5, L)$	2.872	2.633	2.786	1.752

methods are summarized in Tables 2 and 3 for sample sizes $n \in \{200, 500\}$, and in Tables 7 and 8 in Appendix A.5 for sample sizes $n \in \{50, 100\}$.

In Tables 2 and 3, it is seen that there isn't a consisent "best" bandwidth method. For skewing functions π_0 (the symmetric case) and π_2, the PI bandwidth generally has smallest median ISE. In these same scenarios, MISE frequently (but by no means consistently) outperforms CV. For $\pi_1(z)$ the MISE bandwidth performs best. In all simulation settings, there is a GSS bandwidth method that results in better performance than the nonparametric estimator. These same conclusions broadly hold for sample sizes $n \in \{50, 100\}$ in Appendix A.5.

The results presented above were restricted to phase-function matching for the GMM estimators, as it was found to generally have better performance that skewness matching. For details of the simulation comparing the two GMM matching methods, see Appendix A.3.

Table 3 Median of 100 × ISE for the GSS deconvolution estimators with CV, MISE, and PI bandwidths, and the nonparametric (NP) estimator with PI bandwidth. Sample size $n = 500$

π	(NSR, U)	CV	MISE	PI	NP
π_0	$(0.2, N)$	0.190	0.180	0.160	0.334
	$(0.5, N)$	0.356	0.382	0.297	0.728
	$(0.2, L)$	0.186	0.202	0.152	0.233
	$(0.5, L)$	0.295	0.350	0.226	0.401
π_1	$(0.2, N)$	1.885	1.788	2.027	2.064
	$(0.5, N)$	2.781	2.640	3.350	3.810
	$(0.2, L)$	0.897	0.784	0.991	1.271
	$(0.5, L)$	1.264	1.039	1.304	1.929
π_2	$(0.2, N)$	1.492	1.158	1.173	1.401
	$(0.5, N)$	3.746	3.147	2.967	3.456
	$(0.2, L)$	0.845	0.873	0.471	0.636
	$(0.5, L)$	1.752	1.640	1.376	1.048

GMM estimation

One other simulation study was performed, and considered the choice of M (the number of even moment to use) when evaluating the GMM estimators of (ξ, ω). These simulations results are presented in Appendix A.4. In summary, the larger value $M = 5$ was generally seen to outperform $M = 2$ for $\pi_1(z)$ and $\pi_2(z)$. In the symmetric $\pi_0(z)$ case, $M = 2$ performed slightly better than $M = 5$. In all instances, root mean square error (RMSE) was used as criterion.

Data applications

Coal abrasiveness index data

Data from an industrial application, first considered by Lombard (2005), are analyzed here. The data were obtained by taking batches of coal, splitting them in two, and randomly allocating each of the half-batches to one of two methods used to measure the abrasiveness index (AI) of coal, a measure of the quality of coal. The observed data consist of 98 pairs (w_{1i}, w_{2i}) assumed to be from a population with $W_{1i} = X_i + U_{1i}$ and $W_{2i} = \mu + \sigma(X_i + U_{2i})$. Here, X_i denotes the true AI of the ith batch, U_{1i} and U_{i2} denote measurement error, and constants μ and σ account for the two AI measurement methods being on different scales. Of interest is estimating $f_x(x)$, the true density of AI. However, the data (w_{1i}, w_{2i}) first need to be combined in a sensible way.

To this end, let $\mu_{w,k}$ and $\sigma^2_{w,k}$ denote the mean and variance of the W_{ki}, $k = 1, 2$, and let μ_x and σ^2_x denote the mean and variance of the X_i. Note that $\mu_{w,1} = \mu_x$, $\mu_{w,2} = \mu + \sigma\mu_x$, $\sigma^2_{w,1} = \sigma^2_x + \sigma^2_u$, and $\sigma^2_{w,2} = \sigma^2\left(\sigma^2_x + \sigma^2_u\right)$. By replacing the population moments with their sample counterparts, estimators $\hat{\sigma} = s_{w,2}/s_{w,1} = 0.679$ and $\hat{\mu} = \bar{w}_2 - \hat{\sigma}\bar{w}_1 = 59.503$ are obtained. Here, $(\bar{w}_1, s_{w,1})$ denote the sample mean and standard deviation of the observed w_1-data with similar definitions holding for the w_2-quantities. Now, the paired observations are combined as $w_i = 0.5w_{1i} + 0.5\left(w_{2i} - \hat{\mu}\right)/\hat{\sigma}$. At the population level this corresponds to $W_i \approx X_i + 0.5\,(U_{1i} + U_{2i}) := X_i + \varepsilon_i$. An estimate of the measurement error variance σ^2_ε is obtained by calculating $\hat{\sigma}^2_u = (2n)^{-1}\sum\left[W_{1i} - \left(W_{2i} - \hat{\mu}\right)/\hat{\sigma}\right]^2 = 174.6$ and noting that $\hat{\sigma}^2_\varepsilon = 174.6/2 = 87.3$. This corresponds to the W_i having noise-to-signal ratio NSR = 16.35%.

The GSS deconvolution estimator for $f_x(x)$ is now calculated assuming a normal symmetric component, $f_0(z) = \phi(z)$, along with a Laplace distribution for the measurement error ε. (The equivalent estimator assuming normal measurement was also calculated and is nearly identical in shape.) GMM with $M = 5$ gives solution pairs $(\hat{\xi}_1, \hat{\omega}_1) = (192.88, 29.90)$ and $(\hat{\xi}_2, \hat{\omega}_2) = (230.41, 32.43)$. For each of these, the corresponding skewing function estimate $\tilde{\pi}_j(z)$ and phase function distance R_j was calculated, the latter using weight function $w(t) = [1 - (t/t^*)^2]^3$ for $t \in [-t^*, t^*]$ and $t^* = 0.06$. Here, $R_1 = 0.023 < 0.046 = R_2$ and therefore solution $(\hat{\xi}_1, \hat{\omega}_1)$ with estimated skewing function $\tilde{\pi}_1(z)$ was selected. Skewness matching resulted in selection of the same solution. Figure 2 shows a kernel density estimator of $f_w(w)$, the density of the contaminated W_i, as well as the GSS deconvolution estimator of $f_x(x)$ with MISE bandwidth $\tilde{h} = 0.102$.

This application illustrates one of the less appealing aspects of the GSS approach sometimes encountered in smaller samples. Note the sharp "edge" in the GSS estimator around $x = 225$. This is an artefact of the hard truncation applied when calculating the range-respecting skewing function estimate $\tilde{\pi}(z)$. The resulting density estimate is

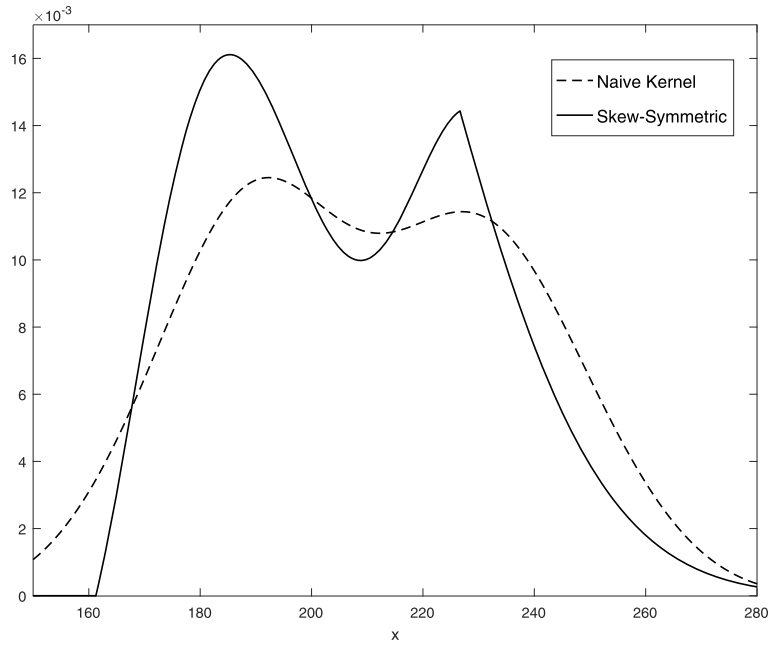

Fig. 2 Abrasiveness Index Density Estimation

not differentiable at this point. This is equivalent to non-differentiable points in the nonparametric deconvolution estimator when it is truncated to be positive.

Systolic blood pressure data

The data here are a subset of $n = 1615$ male participants in the Framingham Heart Study, see for example Carroll et al. (2006) for more detail. The data consist of systolic blood pressure measurements from two patient exams (the second and third exams in the study). At each exam, two replicate measurements were obtained giving data $(\text{SBP}_{21}, \text{SBP}_{22}, \text{SBP}_{31}, \text{SBP}_{32})$. Let $P_1 = (\text{SBP}_{21} + \text{SBP}_{22})/2$ and $P_2 = (\text{SBP}_{31} + \text{SBP}_{32})/2$ denote the average systolic blood pressure observed at each of the exams, and calculate transformed variables $W_j = \log(P_j - 50)$, $j = 1, 2$, as suggested by Carroll et al. (2006). This is done to adjust large skewness present in the data. The measurement $W = (W_1 + W_2)/2 = X + U$ is a surrogate for the true long-term average systolic blood pressure X (on the transformed logarithmic scale). Using the replicates (W_1, W_2), estimate standard deviations $\hat{\sigma}_x = 0.1976$ and $\hat{\sigma}_u = 0.0802$ are obtained.

The GSS deconvolution estimator assuming a Laplace distribution for the measurement error U and using a normal reference density $f_0(z) = \phi(z)$ was computed. GMM with $M = 5$ resulted in only one solution, $(\hat{\xi}, \hat{\omega}) = (4.429, 0.210)$, and therefore no selection was needed. Figure 3 displays both the GSS deconvolution estimator and the nonparametric deconvolution estimator, both with PI bandwidths.

The nonparametric deconvolution estimator has previously been applied to the Framingham Heart Study. It is therefore reassuring that the GSS estimator is not dissimilar in appearance.

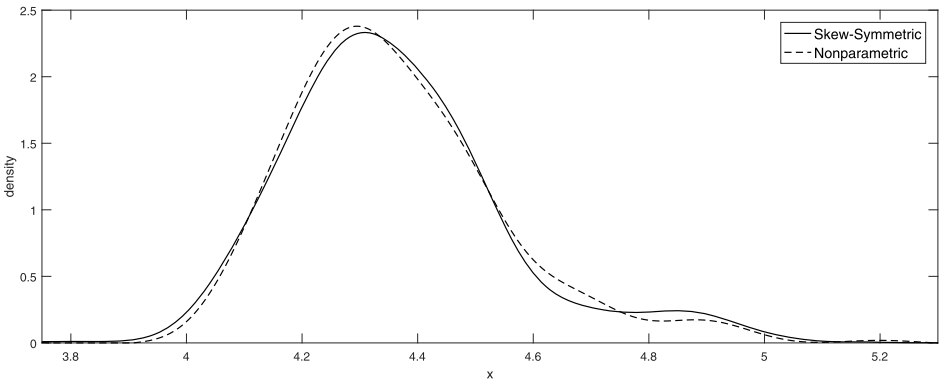

Fig. 3 Density deconvolution estimators of log(SBP-50)

Conclusion

In this paper, the density deconvolution problem is considered for variables belonging to the family of generalized skew-symmetric (GSS) distributions. Implementation requires both the estimation of location and scale parameters (ξ, ω), and the estimation of a skewing function $\pi(z)$. Estimation methods are proposed for both of these quantities, and extensive simulation studies are performed. In simulation studies performed, the GSS deconvolution estimator is generally seen to result in large improvements over the nonparametric deconvolution estimator (using median ISE as criterion).

There are still several questions related to GSS deconvolution that can be considered. Firstly, the estimator requires the specification of a known symmetric component $f_0(z)$. While this is done to ensure model identifiability, it would be possible to consider several candidate symmetric densities and choose the "best" among these. The related goodness-of-fit testing problem for a specified symmetric component can also be explored. Secondly, it should be noted that the contaminated W also has a GSS distribution. An alternative modeling approach could therefore estimate the pdf of W directly and then recover the pdf of X. Lastly, it was observed in the simulation study that the nonparametric deconvolution kernel in a few isolated instances had superior performance to the GSS estimator under selection, while GSS had better under oracle conditions for the same simulation configurations. This suggests that further refinement of the bandwidth calculation and solution selection procedure may be possible, and related work is ongoing.

Appendix

A.1 Generalized skew-symmetric representation

Here, it is established that any continuous random variable has a non-unique representation as a GSS distribution. This motivates, in part, the need to assume a parametric form for pdf $f_0(z)$ when doing estimation. Let Y be a continuous random variable with pdf $f_y(y)$ and let ξ be a real number. Furthermore, let B be a Bernoulli$(p = 0.5)$ random variable, and define new random variables $D_\xi = |Y - \xi|$ and $T = BD_\xi - (1 - B)D_\xi$. The random variable T is symmetric about 0 and has pdf $f_t(t) = (1/2)\left[f_y(\xi + t) + f_y(\xi - t)\right]$. Next, define

$$\pi_t(t) = \frac{1}{2}\frac{f_y(\xi + t)}{f_t(t)} = \frac{f_y(\xi + t)}{f_y(\xi + t) + f_y(\xi - t)}$$

and note that $\pi_t(t)$ satisfies $0 \leq \pi_t(t) = 1 - \pi_t(-t) \leq 1$. By construction, it follows that $f_y(y)$ can be expressed as $f_y(y) = 2f_t(y-\xi)\pi_t(y-\xi)$. Assuming that Y has finite variance, the variance of T is given by $\omega_\xi^2 = \int_{\mathbb{R}} t^2 f_t(t)dt$. Then, letting $f_\xi(t) = f_t(t/\omega_\xi)/\omega_\xi$ and $\pi_\xi(t) = \pi_t(t/\omega_\xi)$, it is possible to write

$$f_y(y) = \frac{2}{\omega_\xi} f_\xi\left(\frac{y-\xi}{\omega_\xi}\right)\pi_\xi\left(\frac{y-\xi}{\omega_\xi}\right).$$

This representation does not depend on a specific value for ξ and, as such, holds for every ξ. However, each value of ξ is associated with a different symmetric component $f_\xi(z)$ and skewing function $\pi_\xi(z)$. As such, there is a family of distributions $f_\xi(z)$ symmetric about 0 and with unit variance such that the random variable Y can be expressed as a GSS distribution with symmetric component belonging to this family. The work in this paper is motivated by the assumption that it is possible to correctly specify one symmetric distribution in the family $f_\xi(z)$.

A.2 MISE derivation

To derive an expression for the mean integrated square error (MISE), considering the estimator $\hat{s}_0(t)$ defined in (5). Recall that $\mathrm{E}\left[\hat{s}_0(t)\right] = \psi_k(ht)s_0(t)$. Additionally, it has covariance structure

$$\mathrm{Cov}\left[\hat{s}_0(t_1), \hat{s}_0(t_2)\right] = \frac{\psi_k(ht_1)\psi_k(ht_2)}{n} \times \left[\frac{c_0(t_1-t_2)\psi_u[(t_1-t_2)/\omega] - c_0(t_1+t_2)\psi_u[(t_1+t_2)/\omega)]}{2\psi_u(t_1/\omega)\psi_u(t_2/\omega)} - s_0(t_1)s_0(t_2)\right].$$

The integrated squared error (ISE) of the GSS estimator can now be expressed in terms of $\hat{s}_0(t)$,

$$\begin{aligned}\mathrm{ISE} &= \int_{\mathbb{R}} \left[\tilde{f}_z(z) - f_z(z)\right]^2 dz \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \left|\hat{\psi}_z(t) - \psi_z(t)\right|^2 dt \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \left[\hat{s}_0(t) - s_0(t)\right]^2 dt\end{aligned}$$

where the first equality is an application of Parseval's identity, and the second follows upon noting that the estimated characteristic function $\hat{\psi}_z(t)$ and true characteristic function $\psi_z(t)$ have common real component $c_0(t)$ which therefore cancels out, leaving only the estimated and true imaginary components. Also note that ISE is a function of the bandwidth h through $\hat{s}_0(t)$. Now, MISE $=$ E[ISE] can be evaluated using the expectation and covariance functions associated with $\hat{s}_0(t)$, in the latter setting $t_1 = t_2 = t$. Eq. 10 follows.

A.3 GMM estimators simulation

The performance of GMM estimation of (ξ, ω) was evaluated in a simulation study. Data were simulated as described in the main paper. For each simulated dataset, the estimators minimizing $D(\xi, \omega)$ were obtained for both $M = 2$ and $M = 5$ even moments. While the sixth, eight and tenth sample moments used for the $M = 5$ setting arguably contain additional information, there is a great deal of added variability introduced when estimating these higher order moments. This simulation explored the benefits, if any, of doing so. In simulated samples where multiple solutions $(\hat{\xi}_j, \hat{\omega}_j)$, $j = 1, \ldots, J$ were obtained, the existence of an oracle able to choose the solution

Table 4 RMSE for GMM estimators, N = Normal, L = Laplace

			$M=2$		$M=5$	
π	n	(NSR, U)	RMSE($\hat{\xi}$)	RMSE($\hat{\omega}$)	RMSE($\hat{\xi}$)	RMSE($\hat{\omega}$)
π_0	200	$(0.2, N)$	0.400	0.116	0.404	0.127
		$(0.5, N)$	0.454	0.140	0.452	0.153
		$(0.2, L)$	0.409	0.120	0.414	0.133
		$(0.5, L)$	0.494	0.157	0.483	0.168
	500	$(0.2, N)$	0.370	0.094	0.383	0.105
		$(0.5, N)$	0.415	0.113	0.431	0.128
		$(0.2, L)$	0.377	0.097	0.389	0.109
		$(0.5, L)$	0.453	0.133	0.453	0.136
π_1	200	$(0.2, N)$	0.131	0.112	0.092	0.091
		$(0.5, N)$	0.177	0.138	0.151	0.121
		$(0.2, L)$	0.139	0.117	0.092	0.093
		$(0.5, L)$	0.195	0.154	0.152	0.124
	500	$(0.2, N)$	0.080	0.069	0.055	0.057
		$(0.5, N)$	0.103	0.084	0.079	0.071
		$(0.2, L)$	0.083	0.072	0.055	0.058
		$(0.5, L)$	0.118	0.097	0.079	0.073
π_2	200	$(0.2, N)$	0.133	0.055	0.096	0.058
		$(0.5, N)$	0.234	0.071	0.185	0.068
		$(0.2, L)$	0.153	0.058	0.093	0.059
		$(0.5, L)$	0.334	0.109	0.194	0.088
	500	$(0.2, N)$	0.081	0.034	0.059	0.037
		$(0.5, N)$	0.135	0.037	0.112	0.039
		$(0.2, L)$	0.093	0.035	0.057	0.037
		$(0.5, L)$	0.219	0.061	0.124	0.054

closest to the true value $(0, 1)$ (as measured using Euclidean distance) was assumed. A total of $N = 1000$ samples were generated for each simulation configuration. Root mean square error (RMSE) was used as criterion, and the results are shown in Table 4.

In the setting with X normal, i.e. using $\pi_0(z)$, using $M = 5$ moments results in a small increase in RMSE compared to the case $M = 2$. The average increase in RMSE for ξ is 1.2% and for ω is 9.5% across the settings considered. On the other hand, the simulation results for skewing functions $\pi_1(z)$ and $\pi_2(z)$ look very different. Here, the RMSE for ξ decreases for both skewing functions, and the RMSE for ω decreases for skewing function $\pi_2(z)$. Also, the average RMSE of ω for $\pi_2(z)$ remains unchanged across the simulation settings considered. One possible reason for the increase in RMSE in the symmetric case is that the underlying distribution is normal and therefore higher-order moments do not contain any "extra" information about the distribution. On the other hand, for $\pi_1(z)$ and $\pi_2(z)$ there is a substantive departure from normality and the higher-order sample moments, despite their large variability, do contain useful information about the underlying distribution. As the increase in RMSE in the symmetric case is relatively small compared to the decrease in the asymmetric cases, the paper uses the GMM estimators with $M = 5$ in all other simulations.

Table 5 Median of 100 × ISE for GSS estimator with MISE bandwidth

		$n = 200$					$n = 500$				
π	(NSR, *U*)	MIN	SKW	PHS	RND	NP	MIN	SKW	PHS	RND	NP
π_0	(0.2, *N*)	0.278	0.460	0.370	0.455	0.535	0.143	0.234	0.180	0.215	0.334
	(0.5, *N*)	0.559	0.943	0.701	1.015	1.039	0.320	0.503	0.382	0.529	0.728
	(0.2, *L*)	0.343	0.411	0.407	0.469	0.433	0.173	0.196	0.202	0.216	0.233
	(0.5, *L*)	0.533	0.629	0.630	0.701	0.653	0.286	0.317	0.350	0.384	0.401
π_1	(0.2, *N*)	1.923	2.407	2.116	2.322	2.709	1.545	1.832	1.788	1.812	2.064
	(0.5, *N*)	2.829	3.730	3.032	3.744	4.601	2.474	3.166	2.640	3.052	3.810
	(0.2, *L*)	1.309	1.637	1.494	1.721	1.998	0.671	1.016	0.784	1.234	1.271
	(0.5, *L*)	1.789	2.216	2.116	2.524	2.848	0.992	1.431	1.039	1.671	1.929
π_2	(0.2, *N*)	1.593	1.612	1.593	4.513	1.956	1.158	1.158	1.158	5.050	1.401
	(0.5, *N*)	4.115	4.175	4.175	7.120	4.375	3.147	3.147	3.147	6.273	3.456
	(0.2, *L*)	1.229	1.230	1.230	4.098	1.044	0.873	0.873	0.873	4.401	0.636
	(0.5, *L*)	2.520	2.529	2.633	4.838	1.752	1.631	1.631	1.640	3.925	1.048

A.4 Solution selection simulation

The simulation results comparing the performance of the skewness matching and phase function distance solution selection mechanisms follow here. Data were generated as described in the "Simulation studies" section of the main paper. For each simulated sample, all GMM solutions $(\hat{\xi}_j, \hat{\omega}_j)$, $j = 1, \ldots, J$ were obtained. Solution selection was then implemented for both skewness matching and phase function matching. These techniques require a bandwidht to be selected. The simulation implemented CV, MISE, and PI bandwidth selection. However, the conclusions with regards to selection methods were very similar for these and therefore only MISE bandwidth results are included here. To contextualize these results from selection, results corresponding to an oracle able to choose the solution with smallest ISE are also reported, as well as a *blind selection* approach randomly selecting one of the GMM solutions.

The simulation results are summarized in Table 5. In this table, the median ISE of skewness matching and phase function distance are given in the columns SKW and PHS. The column MIN contains the median ISE for the oracle selecting the solution with smallest ISE, while RND contains the median ISE of randomly selecting one of the GMM solutions. Finally, the median ISE of the nonparametric deconvolution estimator with PI bandwidth is given in column NP for reference purposes.

Inspection of Table 5 shows that estimation under both the skewness and phase function matching generally performs better than the fully nonparametric estimator, with the exception being the combination of skewing function $\pi_2(z)$ and Laplace measurement error. However, as the GSS estimator outperformed the nonparametric estimator under an oracle bandwidth as seen in Table 1 of the main paper, this does suggest that further improvement of the GSS estimator may still be possible by refining parameter estimation and bandwidth selection – this is ongoing work. Further inspection of Table 5 shows that estimation under both the skewness and phase function matching performs better than random selection, with the exception that random selection outperforms the skewness matching for $\pi_1(z)$ and normal measurement error. While there are a few instances where skewness matching outperformed phase function matching, the latter generally has very good performance and comes close to the best possible performance of the minimum ISE under oracle selection.

Table 6 Median of 100 × ISE, as well as first and third quartiles $[Q_1, Q_3]$ for the oracle GSS and nonparametric (NP) deconvolution estimators, sample sizes $n = 50, 100$

		$n = 50$		$n = 100$	
π	(NSR, U)	GSS	NP	GSS	NP
π_0	$(0.2, N)$	0.315	0.865	0.199	0.591
		[0.126, 0.712]	[0.473, 1.489]	[0.086, 0.400]	[0.351, 0.975]
	$(0.5, N)$	0.431	1.400	0.288	1.040
		[0.177, 0.899]	[0.802, 2.254]	[0.115, 0.594]	[0.626, 1.621]
	$(0.2, L)$	0.322	0.687	0.203	0.445
		[0.141, 0.715]	[0.350, 1.243]	[0.086, 0.425]	[0.222, 0.795]
	$(0.5, L)$	0.421	0.835	0.255	0.548
		[0.185, 1.019]	[0.401, 1.635]	[0.114, 0.592]	[0.245, 0.991]
π_1	$(0.2, N)$	2.290	3.830	1.984	3.094
		[1.622, 3.180]	[2.713, 5.300]	[1.383, 2.680]	[2.249, 4.064]
	$(0.5, N)$	2.830	5.576	2.512	4.803
		[2.059, 3.932]	[4.037, 7.598]	[1.817, 3.504]	[3.510, 6.312]
	$(0.2, L)$	2.088	3.370	1.608	2.389
		[1.435, 2.880]	[2.303, 4.641]	[1.142, 2.273]	[1.678, 3.297]
	$(0.5, L)$	2.410	4.077	1.977	3.138
		[1.742, 3.391]	[2.864, 5.798]	[1.405, 2.755]	[2.248, 4.427]
π_2	$(0.2, N)$	2.658	3.031	1.900	2.305
		[1.464, 4.274]	[1.737, 4.675]	[1.138, 2.873]	[1.500, 3.287]
	$(0.5, N)$	4.482	5.360	3.924	4.682
		[2.566, 6.831]	[3.193, 7.536]	[2.344, 5.602]	[2.954, 6.372]
	$(0.2, L)$	1.968	2.101	1.132	1.264
		[1.052, 3.412]	[1.228, 3.458]	[0.606, 1.878]	[0.785, 1.995]
	$(0.5, L)$	3.333	3.121	2.498	2.080
		[1.880, 5.838]	[1.862, 5.038]	[1.276, 4.039]	[1.172, 3.255]

A.5 supplemental simulation results

This subsection contains two sets of supplemental simulation results. The first of these, found in Table 6, pertains to a comparison of oracle estimators for sample sizes $n = \{50, 100\}$. The second of these, found in Tables 7 and 8, pertains to comparing bandwidth estimation methods for sample sizes $n = \{50, 100\}$. The conclusions that can be drawn from these results are consistent with those discussed in the "Simulation studies" section of the main paper, and are included here for completeness.

Table 7 Median of 100 × ISE for the GSS deconvolution estimators with CV, MISE, and PI bandwidths, and the nonparametric (NP) estimator with PI bandwidth. Sample size $n = 50$

π	(NSR, U)	CV	MISE	PI	NP
π_0	$(0.2, N)$	1.415	1.291	0.875	1.230
	$(0.5, N)$	2.068	1.974	1.194	1.896
	$(0.2, L)$	1.333	1.298	0.984	1.195
	$(0.5, L)$	1.836	1.817	1.313	1.815
π_1	$(0.2, N)$	3.416	3.214	3.927	4.503
	$(0.5, N)$	4.732	4.952	7.256	6.536
	$(0.2, L)$	3.517	3.348	4.277	4.176
	$(0.5, L)$	4.244	4.171	7.529	5.418
π_2	$(0.2, N)$	4.267	3.537	3.691	3.574
	$(0.5, N)$	7.220	7.256	6.177	6.051
	$(0.2, L)$	3.464	3.255	3.070	2.783
	$(0.5, L)$	6.004	5.776	5.949	4.058

Table 8 Median of 100 × ISE for the GSS deconvolution estimators with CV, MISE, and PI bandwidths, and the nonparametric (NP) estimator with PI bandwidth. Sample size $n = 100$

π	(NSR, U)	CV	MISE	PI	NP
π_0	$(0.2, N)$	0.727	0.630	0.526	0.810
	$(0.5, N)$	1.162	1.154	0.771	1.452
	$(0.2, L)$	0.761	0.736	0.576	0.792
	$(0.5, L)$	0.978	1.011	0.755	1.110
π_1	$(0.2, N)$	2.946	2.726	3.191	3.553
	$(0.5, N)$	3.887	3.896	5.940	5.443
	$(0.2, L)$	2.695	2.556	3.220	2.898
	$(0.5, L)$	3.382	3.356	5.613	4.050
π_2	$(0.2, N)$	2.916	2.360	2.410	2.612
	$(0.5, N)$	5.987	5.751	4.918	5.330
	$(0.2, L)$	2.190	1.843	1.663	1.732
	$(0.5, L)$	4.407	4.143	4.856	2.769

Abbreviations

CV: Cross-validation; GMM: Generalized method of noments; GSS: Generalized skew-symmetric; iid: Independent and identically distributed; ISE: Integrated squared error; MISE: Mean integrated squared error; pdf: Probability density function; PI: Plug-in; RMSE: Root mean square error

Acknowledgements

Not applicable.

Author's contributions

This is a single author paper and all research and writing is was conducted by the author. The author read and approved the final manuscript.

Funding

The author has no funding sources to declare.

Availability of data and materials

The coal abrasiveness index data are discussed in Lombard (2005). These data are proprietary, and cannot be released publically. The systolic blood pressure data are discussed in Carroll et al. (2006) and constitute a subset of the Framingham Heart Study. The subset of data used in this paper is publically available in the `R` package `decon`.

Competing interests

The author declares that he has no competing interests.

Received: 16 September 2019 Accepted: 9 July 2020
Published online: 23 July 2020

References

Arellano-Valle, R. B., Ozan, S., Bolfarine, H., Lachos, V.: Skew normal measurement error models. J. Scand. J. Stat. **96**(2), 265–281 (2005)
Arellano-Valle, R. B., Azzalini, A., Ferreira, C. S., Santoro, K.: A two-piece normal measurement error model. Comput. Stat. Data Anal. **144**, 106863 (2020)
Azzalini, A.: A class of distributions which includes the normal ones. Scand. J. Stat. **12**, 171–178 (1985)
Azzalini, A., Genton, M. G., Scarpa, B.: Invariance-based estimating equations for skew-symmetric distributions. Metron. **68**(3), 275–298 (2010)
Azzalini, A.: The Skew-normal and Related Families. Cambridge University Press, New York (2013)
Carroll, R. J., Hall, P.: Optimal rates of convergence for deconvolving a density. J. Am. Stat. Assoc. **83**(404), 1184–1186 (1988)
Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M.: Measurement Error in Nonlinear Models: a Modern Perspective. CRC press, Boca Raton (2006)
Chu, K. K., Wang, N., Stanley, S., Cohen, N. D.: Statistical evaluation of the regulatory guidelines for use of furosemide in race horses. Biometrics. **57**(1), 294–301 (2001). https://doi.org/10.1111/j.0006-341x.2001.00294.x
Delaigle, A., Gijbels, I.: Estimation of integrated squared density derivatives from a contaminated sample. J. R. Stat. Soc. Ser. B Stat. Methodol. **64**(4), 869–886 (2002)
Delaigle, A., Gijbels, I.: Practical bandwidth selection in deconvolution kernel density estimation. Comput. Stat. Data Anal. **45**(2), 249–267 (2004)
Delaigle, A., Hall, P.: Using simex for smoothing-parameter choice in errors-in-variables problems. J. Am. Stat. Assoc. **103**(481), 280–287 (2008)
Delaigle, A., Hall, P., Meister, A.: On deconvolution with repeated measurements. Ann. Stat. **36**(2), 665–685 (2008). https://doi.org/10.1214/009053607000000884

Delaigle, A., Hall, P.: Parametrically assisted nonparametric estimation of a density in the deconvolution problem. J. Am. Stat. Assoc. **109**(506), 717–729 (2014)
Delaigle, A., Hall, P.: Methodology for non-parametric deconvolution when the error distribution is unknown. J. R. Stat. Soc. Ser. B Stat. Methodol. **78**(1), 231–252 (2016)
Diggle, P. J., Hall, P.: A fourier approach to nonparametric deconvolution of a density estimate. J. R. Stat. Soc. Ser. B Methodol. **55**(2), 523–531 (1993). https://doi.org/10.1111/j.2517-6161.1993.tb01920.x
Fan, J.: Asymptotic normality for deconvolution kernel density estimators. Sankhyā: Indian J. Stat. Ser. A. **53**(1), 97–110 (1991a)
Fan, J.: On the optimal rates of convergence for nonparametric deconvolution problems. Ann. Stat. **19**(3), 1257–1272 (1991b)
Fan, J., Truong, Y. K.: Nonparametric regression with errors in variables. Ann. Stat. **21**(4), 1900–1925 (1993). https://doi.org/10.1214/aos/1176349402
Genton, M. G. E. .: Skew-elliptical Distributions and Their Applications: a Journey Beyond Normality. CRC Press, Boca Raton (2004)
Guolo, A.: A flexible approach to measurement error correction in case–control studies. Biometrics. **64**(4), 1207–1214 (2008)
Kahrari, F., Ferreira, C., Arellano-Valle, R.: Skew-normal-cauchy linear mixed models. Sankhya B. **81**(2), 185–202 (2019)
Kim, H.-M., Maadooliat, M., Arellano-Valle, R. B., Genton, M. G.: Skewed factor models using selection mechanisms. J. Multivar. Anal. **145**, 162–177 (2016)
Lachos, V., Labra, F., Bolfarine, H., Ghosh, P.: Multivariate measurement error models based on scale mixtures of the skew–normal distribution. Statistics. **44**(6), 541–556 (2010)
Lombard, F.: Nonparametric confidence bands for a quantile comparison function. Technometrics. **47**(3), 364–371 (2005)
Ma, Y., Genton, M. G., Tsiatis, A. A.: Locally efficient semiparametric estimators for generalized skew-elliptical distributions. J. Am. Stat. Assoc. **100**(471), 980–989 (2005)
Neumann, M. H., Hössjer, O.: On the effect of estimating the error density in nonparametric deconvolution. J. Nonparametric Stat. **7**(4), 307–330 (1997)
Nghiem, L., Potgieter, C. J.: Density estimation in the presence of heteroscedastic measurement error of unknown type using phase function deconvolution. Stat. Med. **37**(25), 3679–3692 (2018)
Potgieter, C. J., Genton, M. G.: Characteristic function-based semiparametric inference for skew-symmetric models. Scand. J. Stat. **40**(3), 471–490 (2013)
Stefanski, L. A., Carroll, R. J.: Deconvolving kernel density estimators. Statistics. **21**(2), 169–184 (1990)
Van Oost, K., Van Muysen, W., Govers, G., Heckrath, G., Quine, T., Poesen, J.: Simulation of the redistribution of soil by tillage on complex topographies. Eur. J. Soil Sci. **54**(1), 63–76 (2003)
Wang, W.-L., Liu, M., Lin, T.-I.: Robust skew-t factor analysis models for handling missing data. Stat. Methods Appl. **26**(4), 649–672 (2017)

Publisher's Note