Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

Semiparametric location estimation under non-random sampling†

Marc G. Genton[a,*], Mijeong Kim[a] and Yanyuan Ma[a]

Received 2 July 2012; Accepted 12 July 2012

We study a class of semiparametric skewed distributions arising when the sample selection process produces non-randomly sampled observations. Based on semiparametric theory and taking into account the symmetric nature of the population distribution, we propose both consistent estimators, i.e. robust to model mis-specification, and efficient estimators, i.e. reaching the minimum possible estimation variance, of the location of the symmetric population. We demonstrate the theoretical properties of our estimators through asymptotic analysis and assess their finite sample performance through simulations. We also implement our methodology on a real data example of ambulatory expenditures to illustrate the applicability of the estimators in practice. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: robustness; selection bias; semiparametric model; skewness; skew-symmetric distribution

1 Introduction

Suppose that in a general population, a certain trait X is symmetrically distributed with location μ, and we are interested in estimating the population location. We denote the probability density function of X in the population of interest as $f(x-\mu)$, $x \in \mathbb{R}$, where f is an even function. The most common practice is to assume f to be a normal density, however, here we do not impose any other assumption on f except that it is a symmetric density. Nevertheless, because of certain mechanisms involved in the data collection process, only a non-random, biased sample from the symmetric population is obtained. Taking the selection bias into account, the random but biased observations $X_1, \ldots, X_n$ are independent and identically distributed with density

$$2f(x-\mu)\pi(x-\mu;\beta), \quad x \in \mathbb{R}, \tag{1}$$

where π is decided by the selection mechanism. Here $\pi(x;\beta) \geqslant 0$ is usually named a skewing function and it satisfies $\pi(x;\beta)+\pi(-x;\beta)=1$ for any x. To allow additional flexibility, we allow π to contain an unknown parameter vector β. The skewing function captures the selection bias and we assume its functional form known. However, because no parametric form is assumed on the symmetric function f, (1) is a semiparametric model.

The special case of (1) where the density f and the skewing function π have parametric forms has been coined a skew-symmetric distribution by Wang et al. (2004). Furthermore, if $f=\phi$, the normal density, and $\pi(x;\beta)=\Phi(\beta x)$, $\beta \in \mathbb{R}$, with Φ the standard normal cumulative distribution function, then (1) reduces to the skew-normal distribution introduced by Azzalini (1985); see the book edited by Genton (2004) and the review by Azzalini (2005) for further

[a]Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA

*Email: genton@stat.tamu.edu

†Supporting information may be found in the online version of this article.

discussions of this distribution and related families. Arellano-Valle et al. (2006) have shown that distributions of the form (1) arise naturally when the observed data are obtained only from a selected portion of the population of interest; see also Arnold & Beaver (2002), Loperfido (2002), Copas & Li (1997), Rao (1985) and references therein.

For instance, consider two independent random variables X^* and Y, each symmetrically distributed around zero, with X^* having density f and Y having cumulative distribution function H. We observe X if and only if $Y < \beta X^*$, $\beta \in \mathbb{R}$, in which case we set $X = X^*$. Then $\mathrm{pr}(X \leqslant x) = \mathrm{pr}(X^* \leqslant x | Y < \beta X^*)$, implying that X has density (1) with $\pi(x;\beta) = H(\beta x)$. When $f = \phi$ and $H = \Phi$, X has the aforementioned skew-normal distribution.

A practical example where a distribution of type (1) arises is illustrated by an ambulatory expenditure data from the 2001 Medical Expenditure Panel Survey analyzed by Cameron & Trivedi (2010). The decision to spend is assumed to be related to the spending amount, hence the observations form a biased sample. Cameron & Trivedi (2010) considered a sample-selection model based on the assumption of normality, hence leading to a parametric skew-normal distribution, which corresponds to assuming f to be normal in (1). We, instead, suggest to eliminate the normal or any other distributional assumption on the symmetric density f. Hence we only require f to be symmetric but otherwise completely unspecified. In this relaxed model setting, we estimate its location μ, which represents the mean of ambulatory expenditures for the general population had there been no expenditure decision to be made.

The rest of the article is organized as follows. In Section 2, we adopt a semiparametric approach to construct a class of consistent estimators of μ that are robust to model mis-specification. We further illustrate how to construct the most efficient estimator via a modified kernel density estimation. We also establish the asymptotic properties of these estimators in this section. Simulation experiments are conducted in Section 3 to illustrate the finite sample performance of these estimators. We implement the proposed estimators to analyze a real data example in Section 4, and give a discussion in Section 5. Technical details are provided in the supporting information.

2 Estimation

2.1. Semiparametric derivation

Although the central interest in (1) is to estimate μ, because β is also unknown, we estimate β together with μ. To this end, we treat $\theta = (\mu, \beta^T)^T$ as the parameter of interest, and treat the unknown symmetric density function f as an infinite dimensional nuisance parameter.

A rich class of root-n consistent estimators for θ in the semiparametric model (1) is the locally efficient semiparametric estimators. Following Bickel et al. (1993) we view the space of all the mean zero, finite variance functions as a Hilbert space $\mathcal{H}$. We begin by finding two subspaces of $\mathcal{H}$, namely the nuisance tangent space Λ and its orthogonal complement $\Lambda^\perp$; see the supporting information for the description of Λ and $\Lambda^\perp$ and the locally efficient estimators. Tsiatis (2006) provides more detailed explanations of these concepts. From here on, we use a subindex $_0$ to denote the true values of the parameters or the true functions, and write the projection of h onto a space A as $\Pi(h|A)$.

For model (1), we establish in the supporting information that the nuisance tangent space Λ, corresponding to the unspecified symmetric probability density function f, and its orthogonal complement $\Lambda^\perp$, are respectively

$$\Lambda = \left\{ u(x-\mu) : u(t) = u(-t), \int_0^\infty u(t) f_0(t) dt = 0 \right\}, \quad \Lambda^\perp = \left\{ v(x-\mu) : v(t)\pi(t;\beta) + v(-t)\pi(-t;\beta) = 0 \right\}.$$

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

Any function in $\Lambda^{\perp}$ can be normalized to an influence function hence provides an estimation function. However, within this large class of estimation functions, the efficient score function is the most attractive because the corresponding estimator has the smallest estimation variance. The efficient score is defined as the orthogonal projection of the score function to $\Lambda^{\perp}$. Denote $g(x;\theta) = 2f(x-\mu)\pi(x-\mu;\beta)$. Calculating $\partial \log g(x;\theta)/\partial\theta$, we obtain the score function

$$S_\theta = \left(S_\mu, S_\beta^T\right)^T = \left\{-\frac{f_0'(x-\mu)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)}, \frac{\pi_\beta'(x-\mu;\beta)^T}{\pi(x-\mu;\beta)}\right\}^T,$$

with the notation $\pi_x'(x-\mu;\beta) = \partial\pi(x-\mu;\beta)/\partial x$ and $\pi_\beta'(x-\mu;\beta) = \partial\pi(x-\mu;\beta)/\partial\beta$. We decompose S_μ into

$$\begin{aligned} S_\mu = &\left[-\frac{f_0'(x-\mu)}{f_0(x-\mu)}\{\pi(x-\mu;\beta) - \pi(-x+\mu;\beta)\} - 2\pi_x'(x-\mu;\beta)\right] \\ &+ \left\{-\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi_x'(x-\mu;\beta)\right\}. \end{aligned}$$

We verify in the supporting information that

$$-\frac{f_0'(x-\mu)}{f_0(x-\mu)}\{\pi(x-\mu;\beta) - \pi(-x+\mu;\beta)\} - 2\pi_x'(x-\mu;\beta) \in \Lambda$$

and

$$-\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi_x'(x-\mu;\beta) \in \Lambda^{\perp}.$$

In addition, since $\pi(t;\beta) + \pi(-t;\beta) = 1$, $\pi_\beta'(t;\beta) + \pi_\beta'(-t;\beta) = 0$, therefore it indicates that $S_\beta \in \Lambda^{\perp}$. We thus obtain the efficient score vector for θ as

$$S_{\theta,\text{eff}}(x;\theta,f_0) = \left\{-\frac{f_0'(x-\mu)2\pi(-x+\mu;\beta)}{f_0(x-\mu)} - \frac{\pi_x'(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi_x'(x-\mu;\beta), \frac{\pi_\beta'(x-\mu;\beta)^T}{\pi(x-\mu;\beta)}\right\}^T.$$

2.2. Robust estimation family

The form of the efficient score depends on the true yet unknown population density f_0. Thus, it cannot be directly used to construct an estimating equation. However, a useful compromise is to take advantage of the efficient score function form to construct consistent estimators. Note that only the first component of the efficient score function relies on the unknown f_0 function. We find that for any symmetric density f^*, even if $f^* \neq f_0$, we would still have

$$\begin{aligned} &\left\{-\frac{f^{*'}(t)2\pi(-t;\beta)}{f^*(t)} - \frac{\pi_t'(t;\beta)}{\pi(t;\beta)} + 2\pi_t'(t;\beta)\right\}\pi(t;\beta) + \left\{-\frac{f^{*'}(-t)2\pi(t;\beta)}{f^*(-t)} - \frac{\pi_t'(-t;\beta)}{\pi(-t;\beta)} + 2\pi_t'(-t;\beta)\right\}\pi(-t;\beta) \\ &= -\frac{f^{*'}(t)2\pi(-t;\beta)\pi(t;\beta)}{f^*(t)} - \pi_t'(t;\beta) + 2\pi_t'(t;\beta)\pi(t;\beta) + \frac{f^{*'}(t)2\pi(t;\beta)\pi(-t;\beta)}{f^*(t)} - \pi_t'(t;\beta) + 2\pi_t'(t;\beta)\pi(-t;\beta) \\ &= 0. \end{aligned}$$

In the above we used $\pi(t;\beta) + \pi(-t;\beta) = 1$ and $\pi_t'(t;\beta) = \pi_t'(-t;\beta)$. Thus, according to the description of $\Lambda^{\perp}$, $S_{\theta,\text{eff}}(x;\theta,f^*)$ is still an element of $\Lambda^{\perp}$.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

The ISI's Journal for the Rapid Dissemination of Statistics Research

Based on the above observation, we propose to construct a simple consistent and robust estimator for θ as follows. We first postulate a symmetric density f^*. We then calculate the corresponding efficient score function

$$S_{\theta,\text{eff}}(x;\theta,f^*) = \left\{ \begin{array}{c} -\dfrac{f^{*'}(x-\mu)2\pi(-x+\mu;\beta)}{f^*(x-\mu)} - \dfrac{\pi'_x(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi'_x(x-\mu;\beta) \\ \dfrac{\pi'_\beta(x-\mu;\beta)}{\pi(x-\mu;\beta)} \end{array} \right\}. \tag{2}$$

We form the estimating equation

$$\sum_{i=1}^{n} S_{\theta,\text{eff}}(X_i;\theta,f^*) = 0$$

to solve for $\widehat{\mu}$ and $\widehat{\beta}$. In practice, a normal density or a Student's t density model for f^* are the obvious choices. If the postulated model f^* happens to be the same as f_0, we indeed obtain the efficient estimator for θ from this procedure. However, even if f^* is not the same as f_0, the construction still guarantees consistency. This means the estimator has a robustness property against the mis-specification of f_0.

In postulating a model f^* for f, the only constraint we have is $f^*(x) = f^*(-x)$. In other words, we can choose the variance of the density model arbitrarily. Intuitively, a variance choice that is close to the true variance of f may yield a more stable estimation while a drastically different variance choice could cause some loss on computational stability as well as affect the estimation variability of the final estimates for θ. Thus, a very natural alternative is to postulate a parametric model for f_0, instead of one particular density function. We denote the postulated density family $f^*(x;\gamma)$, where γ is a vector of additional parameters, such as the variance parameter of f^*. In terms of determining γ, we can simply augment the estimating function (2) with the score function for γ or plug in an estimated γ value to (2).

To be specific about estimating θ through augmenting or plugging-in when a more general model $f^*(x;\gamma)$ is postulated, we describe how to estimate γ. We write the model as $f^*(x;\gamma)$ whether or not it contains the truth f_0. Calculating $\partial \log g(x;\theta,\gamma,f^*)/\partial\gamma$ yields the nuisance score vector

$$S_\gamma(x;\theta,\gamma,f^*) = \frac{\partial f^*(x-\mu;\gamma)/\partial\gamma}{f^*(x-\mu;\gamma)}.$$

We can augment (2) with the above estimating function to solve for $\widehat{\gamma}$ and $\widehat{\theta}$ jointly. Alternatively, we can also iteratively use (2) with γ fixed at the current value and use $S_\gamma(x;\theta,\gamma,f^*)$ with θ fixed at the current value to obtain $\widehat{\gamma}$ and $\widehat{\theta}$.

In terms of the robustness and efficiency of this more general strategy, we find that if the posited model $f^*(\cdot;\gamma)$ contains the true f_0, then we obtain the efficient estimator. However, if the posited model $f^*(\cdot;\gamma)$ does not contain the true f_0, we still obtain a consistent estimator. Thus, this more general postulation strategy retains the robust and local efficient property of the simple postulation strategy. In addition, we find that the estimation of the additional parameter γ does not affect the estimation variance of θ. To make a distinction for the two postulation strategies, we write

$$S_{\theta,\text{eff}}(x;\theta,\gamma,f^*) = \left\{ \begin{array}{c} -\dfrac{f^{*'}(x-\mu;\gamma)2\pi(-x+\mu;\beta)}{f^*(x-\mu;\gamma)} - \dfrac{\pi'_x(x-\mu;\beta)}{\pi(x-\mu;\beta)} + 2\pi'_x(x-\mu;\beta) \\ \dfrac{\pi'_\beta(x-\mu;\beta)}{\pi(x-\mu;\beta)} \end{array} \right\}, \tag{3}$$

where $f^{*'}(t;\gamma) = \partial f^*(t;\gamma)/\partial t$. We summarize our discovery stated above in Theorem 1, after stating a useful lemma. The proofs of both Lemma 1 and Theorem 1 are provided in the supporting information.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

Lemma 1

Assume $\sqrt{n}(\widehat{\gamma}-\gamma^*)$ is bounded in probability. Then the two estimators obtained from solving the two estimating equations $\sum_{i=1}^n S_{\theta,eff}(X_i;\theta,\gamma^*,f^*)=0$ and $\sum_{i=1}^n S_{\theta,eff}(X_i;\theta,\widehat{\gamma},f^*)=0$ are asymptotically equivalent; that is, if the estimator $\widehat{\theta}_1$ is the solution of the first equation and $\widehat{\theta}_2$ is the solution of the second equation, then $\sqrt{n}(\widehat{\theta}_1-\widehat{\theta}_2)\to 0$ in probability.

Theorem 1

i) If the candidate family $f^*(x-\mu;\gamma)$ contains the truth f_0, i.e. there exists γ_0 such that $f^*(x-\mu;\gamma_0)=f_0(x-\mu)$, then $\sqrt{n}(\widehat{\theta}-\theta_0)\to N(0,V_{eff})$ in distribution when $n\to\infty$, where $V_{eff}=E\{S_{eff}(X;\theta_0,f_0)S_{eff}(X;\theta_0,f_0)^T\}^{-1}$ and $\widehat{\theta}$ solves the estimating equation $\sum_{i=1}^n S_{\theta,eff}(X_i;\theta,\widehat{\gamma},f^*)=0$. Here, $\widehat{\gamma}$ is a root-n consistent estimator for γ_0.

ii) If the candidate family $f^*(x-\mu;\gamma)$ does not contain the truth, i.e. $f^*(x-\mu;\gamma)\neq f_0(x-\mu)$ for any γ, then $\sqrt{n}(\widehat{\theta}-\theta_0)\to N(0,V)$ in distribution when $n\to\infty$, where $V=A^{-1}E\{S_{\theta,eff}(X;\theta_0,\gamma^*,f^*)S_{\theta,eff}^T(X;\theta_0,\gamma^*,f^*)\}(A^{-1})^T$, and

$$A=E\left\{\frac{\partial S_{\theta,eff}(X;\theta_0,\gamma^*,f^*)}{\partial\theta^T}\right\}.$$

Here $\widehat{\theta}$ solves the estimating equation $\sum_{i=1}^n S_{\theta,eff}(X_i;\theta,\widehat{\gamma},f^*)=0$, where $\widehat{\gamma}=\gamma^*$ or is a root-n consistent estimator of γ^*.

2.3. Efficient estimation

The efficiency of an estimator depends on how close the posited model f^* is to the true f_0. Although the various robust estimators proposed in Section 2.2. guarantee consistency, they only provide a possibility of achieving efficiency. That is, the estimation variability relies on the specific postulated model or the family of models. Only if the model happens to be true or the family happens to contain the true density f_0, then the optimal estimation variance is achieved, otherwise, all one can obtain is consistency.

To overcome this potential loss of efficiency, we propose to perform a nonparametric estimation of f using a modified procedure of the kernel density estimation. The explicit form of the modified kernel estimator for f is

$$\widehat{f}(t)=\frac{1}{2n}\sum_{i=1}^n\frac{1}{h}\left[K\left\{\frac{(X_i-\mu)-t}{h}\right\}+K\left\{\frac{(X_i-\mu)+t}{h}\right\}\right],$$

where K is a symmetric kernel function and h is a bandwidth. To see the rationale behind this estimation, we first ignore the semiparametric model (1). Then we can use the usual kernel density estimator at a given point x to obtain $(nh)^{-1}\sum_{i=1}^n K\left(X_i-x\right)$. Taking into account (1), we write the estimate as

$$2\widehat{f}(x-\mu)\pi(x-\mu;\beta)=\frac{1}{n}\sum_{i=1}^n\frac{1}{h}K\left(\frac{X_i-x}{h}\right)=\frac{1}{n}\sum_{i=1}^n\frac{1}{h}K\left\{\frac{(X_i-\mu)-(x-\mu)}{h}\right\}.$$

Since f is an even function, we require its estimator $\widehat{f}$ to be also even. Thus we have

$$2\widehat{f}(x-\mu)\pi(-x+\mu;\beta)=2\widehat{f}(-x+\mu)\pi(-x+\mu;\beta)=\frac{1}{n}\sum_{i=1}^n\frac{1}{h}K\left\{\frac{(X_i-\mu)-(-x+\mu)}{h}\right\}=\frac{1}{n}\sum_{i=1}^n\frac{1}{h}K\left(\frac{X_i+x-2\mu}{h}\right).$$

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

The ISI's Journal for the Rapid Dissemination of Statistics Research

Combining the above two equalities, we obtain

$$\widehat{f}(x-\mu)=\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{h}\left\{K\left(\frac{X_i-x}{h}\right)+K\left(\frac{X_i+x-2\mu}{h}\right)\right\}, \tag{4}$$

which yields the modified kernel density estimation for f.

The robust estimation described in Section 2.2. can be combined with the modified nonparametric kernel density estimation to yield a procedure that achieves the optimal semiparametric efficiency bound for θ. The estimation procedure is the following:

- Step 1. Choose any even density function f^*. From f^* and the known function π, obtain $\tilde{\theta}=(\tilde{\mu},\tilde{\beta}^T)^T$ through solving $\sum_{i=1}^{n}S_{\text{eff}}(X_i;\theta,f^*)=0$.
- Step 2. Choose a kernel function K and a bandwidth h, plug K, h and $\tilde{\mu}$ into (4) to obtain $\widehat{f}$.
- Step 3. Using the estimated $\widehat{f}$, obtain $\widehat{\theta}=(\widehat{\mu},\widehat{\beta}^T)^T$ through solving $\sum_{i=1}^{n}S_{\text{eff}}(X_i;\theta,\widehat{f})=0$.

It is worth pointing out that in the above procedure, it is not necessary to perform any iteration. The first order optimal asymptotic property is achieved via the simple one-step procedure. However, in practice, iterating between Steps 2 and 3, while using the previously obtained $\widehat{\mu},\widehat{\beta}$ to replace $\tilde{\mu},\tilde{\beta}$ in Step 2 is often recommended, especially when the sample size is moderate or small. A bandwidth h is needed in Step 2. Interestingly, there is no need to perform any under-smoothing in this step and the procedure is very insensitive to the bandwidth. Thus, a standard cross-validation procedure can be used on an initial kernel density estimation to obtain a bandwidth h, and one can then use this bandwidth throughout the estimation procedure. This practice is both theoretically justified and practically well behaved. In Theorem 2, we state the optimal property of the above one-step procedure. We use the notation $a^{\otimes 2}$ to denote aa^T, and the proof is given in the supporting information.

Theorem 2

Let $X_1,\ldots,X_n$ be independent and identically distributed with common density $2f_0(x-\mu_0)\pi(x-\mu_0;\beta_0)$. For any t, let

$$\widehat{f}(t;\tilde{\mu})=\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{h}\left\{K\left(\frac{X_i-\tilde{\mu}-t}{h}\right)+K\left(\frac{X_i+t-\tilde{\mu}}{h}\right)\right\},$$

$$\widehat{f}'(t;\tilde{\mu})=\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{h^2}\left\{K'\left(\frac{t-X_i+\tilde{\mu}}{h}\right)+K'\left(\frac{X_i+t-\tilde{\mu}}{h}\right)\right\},$$

where $\tilde{\mu}$ is estimated from Step 1. Assume $\widehat{\theta}=(\widehat{\mu},\widehat{\beta}^T)^T$ satisfies

$$\sum_{i=1}^{n}S_{\theta,\text{eff}}\{X_i;\widehat{\theta},\widehat{f}(\cdot;\tilde{\mu})\}=0.$$

It then follows that when $n\to\infty$, under the regularity conditions A6 (i)-(iv) listed in the supporting information, $\widehat{\theta}$ is the semiparametric efficient estimator and it satisfies $\sqrt{n}(\widehat{\theta}-\theta_0)\to N(0,V_{\text{eff}})$ in distribution when $n\to\infty$. Here $V_{\text{eff}}=[E\{S^{\otimes 2}_{\theta,\text{eff}}(X;\theta_0,f_0)\}]^{-1}$.

2.4. Population density estimation

A by-product of the efficient estimation described in Section 2.3. is the nonparametric estimation of the population density f itself. Because the only apriori information we have about f is its symmetry, hence it is not a surprise that

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

f has the typical nonparametric bias and variance properties. We point out here that the fact we do not know the location μ and had to estimate it does not affect the first order asymptotic property of estimating f. In other words, the first order asymptotic convergence properties of $\widehat{f}$ remain the same whether we know μ or not. We summarize the theoretical results in Theorem 3 and provide the necessary conditions and proofs in the supporting information.

Theorem 3
Let $c_2 = \int_{-1}^{1} s^2 K(s)ds$, $v_2 = \int_{-1}^{1} K^2(s)ds$. *Under the regularity conditions A6 (i)-(iv) given in the supporting information, the nonparametric estimation* $\widehat{f}$ *obtained from Step 2 satisfies the usual nonparametric bias and variance property*

$$bias\{\widehat{f}(t;\widetilde{\mu})\} = \frac{h^2}{2} f_0''(t) c_2 + o(h^2),$$

$$var\{\widehat{f}(t;\widetilde{\mu})\} = \frac{1}{nh}\left\{\frac{f_0(t)v_2}{2} + I(|t|<h)\int_{-1+\frac{|t|}{h}}^{1-\frac{|t|}{h}} K(s-t/h)K(s+t/h)w(hs)ds\right\} + o\left\{(nh)^{-1}\right\},$$

where $I(\cdot)$ *is the indicator function.*

Because the bias and variance properties have a similar form as in the usual nonparametric estimation, the subsequent mean squared error (MSE) and mean integrated squared error (MISE) results also remain in the standard form. We hence omit these results. Once a nonparametric estimation of f is obtained, it is straightforward to assemble $\widehat{f}$ and $\widehat{\theta}$ together to reconstruct an estimation $\widehat{g}$ of the density of the biased samples. This can provide a visual verification of the estimation in practice, see Section 4 for an illustration.

Estimating the population density function curve $\widehat{f}$ and the density $\widehat{g}$ of the biased samples is a nonparametric density estimation problem. Hence the bandwidth selection is important for the final performance. Here, the usual bandwidth selection procedure such as cross-validation and plug in methods are applicable. However, we recommend a more refined indirect cross-validation procedure, which allows us to use two different kernels: one is suitable for cross-validation purpose while the other is suitable for estimation purpose. The rationale of the indirect cross-validation method and suitable kernels are studied in Savchuk et al. (2010).

3 Simulations

We performed a set of simulation studies to investigate the finite sample performance of the various estimators we proposed. The data sets were generated from model (1) with true density

$$f_0(x) = \frac{3}{4}(1-x^2)I\{x \in [-1,1]\}, \tag{5}$$

and with a normal skewing function $\pi(x-\mu;\beta) = \Phi\{\beta(x-\mu)\}$. In our first simulation study, we assumed $\beta_0 = 1$ to be known, whereas in the second simulation study it was unknown. In each simulation, a sample size $n = 500$ was used and 1000 data sets were generated.

We implemented seven different estimators to illustrate their relative performance. The first estimator can be considered as an oracle estimator, where we solve $\sum_{i=1}^{n} S_{\theta,\text{eff}}(X_i;\theta,f_0) = 0$ to obtain $\widehat{\theta}$. Here we plug in the true density f_0 to the estimating equation, as if we would know the true f_0, hence the name oracle. The second and third estimators are very similar to the first, except that we plug in a wrong density. The estimating equation is explicitly $\sum_{i=1}^{n} S_{\theta,\text{eff}}(X_i;\theta,f^*) = 0$, where $f^* \neq f_0$. In particular, we plugged in a normal and a Student's t with 5 degrees of freedom density f^*. Our fourth, fifth and sixth estimators involve an additional parameter γ, which is the scale

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

The ISI's Journal for the Rapid Dissemination of Statistics Research

parameter of f^*. To be precise, in the fourth estimator, we augment $S_{\theta,\text{eff}}(X_i;\theta,f^*)$ with S_γ, where f^* is a correct model, i.e. density (5), with location parameter μ and scale parameter γ. While in the fifth and sixth estimators, we also augment $S_{\theta,\text{eff}}(X_i;\theta,f^*)$ with S_γ, but now f^* is a mis-specified model, where we used a normal model and a Student's t with 5 degrees of freedom model, with location parameter μ and scale parameter γ. Finally, the efficient estimator described in Section 2.3. is implemented as the seventh estimator.

The simulation results of the seven estimators are summarized in Table I. It can be seen that all the seven estimators in all simulations exhibit very small bias, providing evidence that regardless of whether the f^* function or f^* model is correctly specified or incorrectly specified, regardless of whether f^* is fully specified or partially specified or completely decided through data, the estimators remain consistent. In addition, the average of the estimated variability is very close to the sample version, hence the inference is reasonably reliable at sample size $n = 500$. Here, we point out that the reported estimated standard deviation is obtained via our asymptotic results. As expected from the theory, the estimators 1, 4, and 7 are more efficient than the estimators 2, 3, 5, and 6. In addition, the variabilities of the estimators 1 and 4 are fairly close to each other, which is expected because they are asymptotically equivalent to their first order approximation. Although the estimator 7 is also asymptotically efficient, the nonparametric estimation of f_0 causes some efficiency loss at $n = 500$. Because the estimator 7 still outperforms the estimators 2, 3, 5, 6, and it is a painless procedure, we recommend implementing it in practice, unless one is quite confident about postulating a correct f model. On the other hand, if a quick assessment of the parameters are needed, then one should feel comfortable to postulate a model and perform a Step 1 simple estimation. The simulation evidence strongly supports the consistency of such a procedure.

To further examine the performance of the additional nonparametric estimation procedure, we also plotted the estimated density curves of both the underlying true population and the population reflected by the biased selected sample in Figure 1 in the case of an unknown β. All the results are based on the Quartic kernel function and the bandwidth is selected via the indirect cross-validation procedure. As we can see, the estimation is satisfying when the true population is (5). The plots of the two estimated densities in the case of a known β are very similar to those for the case of an unknown β.

4 Data example

We now analyze the ambulatory expenditures data mentioned in the introduction. The data consists of $n = 2802$ observations and because the distribution of expenditures is highly skewed, the logarithmic scale was used. Following

Table I. Median of 1000 estimates of the location μ and its median absolute standard deviation (sd), and median of the estimated standard deviation ($\widehat{\text{sd}}$) for known and unknown β cases. True value is $\mu_0 = 4$. Results based on sample size $n = 500$.

	Known β			Unknown β		
Estimator	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$	$\widehat{\mu}$	sd	$\widehat{\text{sd}}$
Estimator 1	4.0020	0.0151	0.0142	4.0052	0.0214	0.0166
Estimator 2	4.0023	0.0226	0.0221	4.0062	0.0383	0.0353
Estimator 3	4.0025	0.0230	0.0225	4.0004	0.0470	0.0438
Estimator 4	4.0023	0.0167	0.0144	4.0035	0.0239	0.0166
Estimator 5	4.0023	0.0226	0.0221	4.0033	0.0390	0.0360
Estimator 6	4.0016	0.0252	0.0247	4.0009	0.0441	0.0417
Estimator 7	4.0012	0.0198	0.0186	3.9974	0.0363	0.0297

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

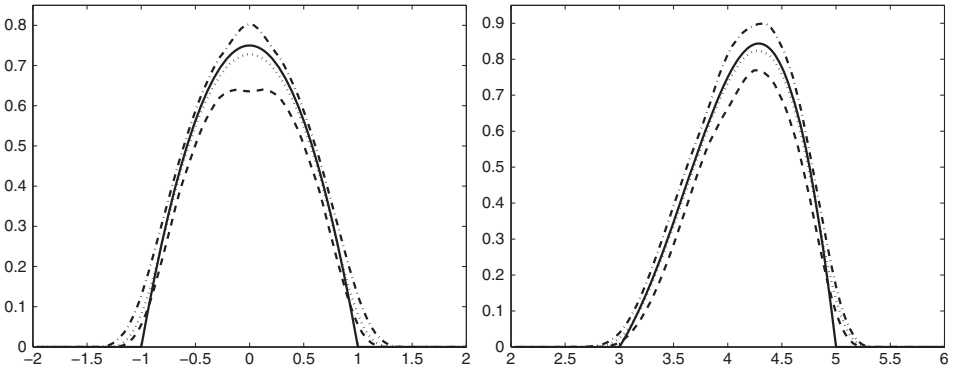

Figure 1. Pointwise quantile curves from simulation with unknown β. In each plot the solid curve is the true density and the other three curves are the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of all 1000 density estimates. The left and right panels correspond to the underlying population density, $f(x)$, and the observed selected subsample density, $g(x)$, respectively.

Cameron & Trivedi (2010), we fit model (1) with a normal skewing function $\pi(x-\mu;\beta) = \Phi\{\beta(x-\mu)\}$; see Section 1 for the motivation and justification of the choice of such a selection model. We computed the seven estimators of the location μ described in the previous section. Specifically, the estimator 1 posited a density (5) for f with a fixed standard deviation of 1.4107, which is the sample standard deviation. The estimators 2 and 3 posited a normal and Student's t with 5 degrees of freedom density for f with again a fixed standard deviation of 1.4107. The estimators 4, 5, and 6 are the same as the estimators 1, 2, and 3, respectively, but with an unknown standard deviation. Finally, the estimator 7 estimated f nonparametrically.

The results for our seven estimators are listed in Table II, as well as their estimated standard deviations. Our estimate of the location is $\widehat{\mu} = 7.95$, whereas other more stringent assumptions on model (1) lead to different estimates. In contrast, the sample mean, an estimator of the location μ that does not correct for the sample selection bias, is 6.56, significantly different from 7.95 at the 95% level according to Table II.

The estimated densities of the population distribution $\widehat{f}$ and the selected sample distribution $\widehat{g}$ are plotted in Figure 2. The estimated sample density curve is overlaid on the histogram of the observations and shows a good fit. The estimated density $\widehat{f}$ has a non-normal shape, hence confirming that it is wise to leave f completely unspecified.

5 Discussion

Throughout the article, our interest is the location of a population whose random sample is not available due to selection bias. The location of a population is well defined for a symmetric population, which is the case studied here. Sometimes, even if a population is not symmetric, one might still be interested in studying its location, where the location could be the mean, the median, the mode or any other suitably defined quantities. Interestingly, this will yield

Table II. Seven estimates $\widehat{\mu}$ of μ and their estimated standard deviations ($\widehat{\text{sd}}$) for the ambulatory expenditures data.

	Estimator 1	Estimator 2	Estimator 3	Estimator 4	Estimator 5	Estimator 6	Estimator 7
$\widehat{\mu}$	8.1913	8.0806	8.0716	8.0805	8.0762	8.0804	7.9456
$\widehat{\text{sd}}$	0.1221	0.1860	0.1821	0.1862	0.1849	0.1861	0.1395

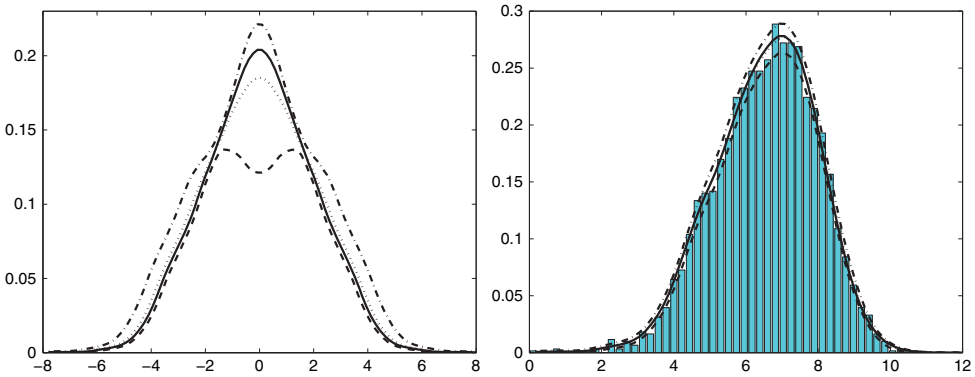

Figure 2. The estimated densities (solid curves) of the population distribution $\widehat{f}$ (left) and the selected sample distribution $\widehat{g}$ (right) for the ambulatory expenditures data. The estimated sample density curve is overlaid on the histogram of the observations, along with the median (dotted), 5% (dashed) and 95% (dot-dashed) quantile curves of the pointwise confidence bands.

a completely different type of semiparametric problems, in which the quantity of interest is a function of both the finite and infinite dimensional parameters. Such problems have been studied in Maity et al. (2007) for random samples, and it will be a challenging and worthy endeavor to investigate it further in the non-random sample case.

We have focused on a rather special selection process, which naturally yields a selection function π that satisfies $\pi(x) + \pi(-x) = 1$. This property has enabled us to derive consistent estimators that are robust to mis-specification of the symmetric part f of the model. Without this property, a consistent estimation of the population location generally requires estimating the population density f itself, and we will no longer be able to construct a robust estimator. In other words, if we still postulate a wrong density or wrong family of models for the density, then the subsequent estimation for the population location may no longer be consistent. However, as long as we are willing to perform nonparametric estimation procedures, possibly taking into account the additional symmetry property of the population distribution f and any characteristics of the selection procedure reflected in π, consistent and even efficient estimates for the population location may be still possible. How to best treat various selection mechanisms is something worth further investigation.

We have treated the case of model (1) where the density f is completely unspecified and the skewing function π is assumed to have a known parametric form due to a specific selection procedure. An alternative setting is when the density f has a known parametric form, whereas the selection mechanism is somewhat hidden, hence the skewing function π is unknown. These models have been investigated by Ma et al. (2005) and Ma & Hart (2007).

Acknowledgement

This research was partially supported by NSF grants DMS-0906341, DMS-1007504, DMS-1100492, NIH grant R01-NS073671 and by Award No. KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST).

References

Arellano-Valle, RB, Branco, MD & Genton, MG (2006), 'A unified view on skewed distributions arising from selections', *The Canadian Journal of Statistics,* **34**, 581–601.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.2

Arnold, BC & Beaver, RJ (2002), 'Skewed multivariate models related to hidden truncation and/or selective reporting', *Test,* **11**, 7–54.

Azzalini, A (1985), 'A class of distributions which includes the normal ones', *Scandinavian Journal of Statistics,* **12**, 171–178.

Azzalini, A (2005), 'The skew-normal distribution and related multivariate families (with discussion)', *Scandinavian Journal of Statistics,* **32**, 159–200.

Bickel, PJ, Klaassen, CAJ, Ritov, Y & Wellner, JA (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, *The Johns Hopkins University Press*, Baltimore.

Cameron, AC & Trivedi, PK (2010), *Microeconometrics using Stata*, Revised Edition, *Stata Press*, College Station, TX.

Copas, JB & Li, HG (1997), 'Inference from non-random samples (with discussion)', *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* **59**, 55–95.

Genton, MG (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality.* Edited Volume, *Chapman & Hall / CRC*, Boca Raton, FL.

Loperfido, N (2002), 'Statistical implications of selectively reported inferential results', *Statistics and Probability Letters,* **56**, 13–22.

Ma, Y, Genton, MG & Tsiatis, AA (2005), 'Locally efficient semiparametric estimators for generalized skew-elliptical distributions', *Journal of the American Statistical Association,* **100**, 980–989.

Ma, Y & Hart, J (2007), 'Constrained local likelihood estimators for semiparametric skew-normal distributions', *Biometrika,* **94**, 119–134.

Maity, A, Ma, Y & Carroll, RJ (2007), 'Efficient estimation of population-level summaries in general semiparametric regression models', *Journal of the American Statistical Association,* **102**, 123–139.

Newey, WK (1990), 'Semiparametric efficiency bounds', *Journal of Applied Econometrics,* **5**, 99–135.

Rao, CR (1985), 'Weighted distributions arising out of methods of ascertainment: what populations does a sample represent?', in Atkinson AC & Fienberg SE (eds), *A Celebration of Statistics: The ISI Centenary Volume, Springer-Verlag,* New York, 543–569.

Savchuk, OY, Hart, JD & Sheather, SJ (2010), 'Indirect cross-validation for density estimation', *Journal of the American Statistical Association,* **105**, 415–423.

Serfling, RJ (2002), *Approximation Theorems of Mathematical Statistics*, *Wiley*, New York.

Tsiatis, AA (2006), *Semiparametric Theory and Missing Data*, *Springer*, New York.

Wang, J, Boyer, J & Genton, MG (2004), 'A skew-symmetric representation of multivariate distributions', *Statistica Sinica,* **14**, 1259–1270.

