Biometrika (2007), **94**, 1, *pp*. 119–134 doi:10.1093/biomet/asm020

Printed in Great Britain

Constrained local likelihood estimators for semiparametric skew-normal distributions

BY YANYUAN MA AND JEFFREY D. HART

Department of Statistics, Texas A&M University, College Station,Texas 77843, U.S.A.
ma@stat.tamu.edu hart@stat.tamu.edu

SUMMARY

A local likelihood estimator for a nonparametric nuisance function is proposed in the context of semiparametric skew-normal distributions. Constraints imposed on such functions result in a nonparametric estimator with a different target function for maximization from classical local likelihood estimators. The optimal asymptotic semiparametric efficiency bound on parameters of interest is achieved by using this estimator in conjunction with an estimating equation formed by summing efficient scores. A generalized profile likelihood approach is also proposed. This method has the advantage of providing a unique estimate in cases where an estimating equation has multiple solutions. Our nonparametric estimator of the nuisance function leads to an estimator of the semiparametric skew-normal density. Both the estimating equation and profile likelihood approaches are applicable to more general skew-symmetric distributions.

Some key words: Efficient score; Local estimator; Profile likelihood; Semiparametric model.

1. INTRODUCTION

Since the classical skew-normal distribution was formally introduced by Azzalini (1985), it has been generalized extensively to accommodate different data patterns. Among this broad class of distributions, the semiparametric skew-symmetric distributions (Wang et al., 2004) allow maximum flexibility of the data distribution and maximum uncertainty in the data-generation scheme.

The probability density function of the semiparametric skew-symmetric distribution has the form

$$f(x) = 2|\Sigma|^{-1/2} g\{\Sigma^{-1/2}(x-\xi)\}\pi\{\Sigma^{-1/2}(x-\xi)\}, \quad x \in \mathbb{R}^d, \tag{1}$$

where g is a standardized symmetric probability density function, and π is a nonnegative function satisfying $\pi(y) + \pi(-y) = 1$ for all $y \in \mathbb{R}^d$. This article concentrates on the case where $d = 1$, $g = \phi$, the standard normal density, and the skewing function π is twice differentiable. When $d = 1$, we use the notation $\Sigma = \sigma^2$.

One motivation for semiparametric skew-symmetric distributions involves selection models. Suppose that a random sample from a population of interest is not available, but instead a 'selected' biased sample is available. The probability density function of each observation has the form

$$f(x; \beta, \alpha) = g(x, \beta)\frac{w(x, \beta, \alpha)}{E\{w(X, \beta, \alpha)\}},$$

where g is the probability density function of the population of interest, and the weight function w reflects the selection mechanism; see Rao (1985) and references therein.

A certain class of selection mechanisms leads to the skew-symmetric model. Let X^* and Y be independent random variables, each of which is symmetrically distributed about 0, with X^* having density g and Y having cumulative distribution function H. One observes X if and only if $Y < p(X^*)$, in which case $X = X^*$. The function p is odd, often equal to cx for some constant c. Then $\mathrm{pr}(X \leqslant x) = \mathrm{pr}\{X^* \leqslant x | Y < p(X^*)\}$, implying that X has density $2g(x)H\{p(x)\}$ (Arnold & Beaver, 2002). This selection criterion thus leads to $w(x) = H\{p(x)\}$, the same type of function as π in (1). Varying H and/or p generates different distributions, some of which are considered in Azzalini (1985), Genton (2004) and Genton & Loperfido (2005). If H and p are not modelled separately, then $H\{p(\cdot)\}$ is precisely π in our model. In the absence of an explicit selection mechanism, a skew-symmetric model may capture the effect of a latent selection mechanism that causes a variable's distribution in a subpopulation to differ from the distribution in the parent population. A more thorough model motivation is given in Ma et al. (2005) and references therein.

In the remainder of the article, we use ξ_0, σ_0 and π_0 to denote the true values of ξ, σ and π. Inference for semiparametric skew-symmetric distributions has concentrated on ξ and σ^2, which represent the mean and variance of a general population; the skewing function π is usually considered a nuisance. When $d = 1$, a locally efficient estimating equation for $\beta = (\xi, \sigma)^{\mathrm{T}}$ is known (Ma et al., 2005) to be

$$\sum_{i=1}^{n} \left[\frac{X_i - \xi}{\sigma} \left\{ 2\pi \left(\frac{X_i - \xi}{\sigma} \right) - 1 \right\} - 2\pi' \left(\frac{X_i - \xi}{\sigma} \right) \right] = 0 \qquad (2)$$

$$\sum_{i=1}^{n} \left\{ \frac{(X_i - \xi)^2}{\sigma^2} - 1 \right\} = 0,$$

where $\pi'(y) = d\pi(y)/dy$. Note that

$$S_{\mathrm{eff}}(X_i, \beta, \pi) = \sigma^{-1} \left[\frac{X_i - \xi}{\sigma} \left\{ 2\pi \left(\frac{X_i - \xi}{\sigma} \right) - 1 \right\} - 2\pi' \left(\frac{X_i - \xi}{\sigma} \right), \frac{(X_i - \xi)^2}{\sigma^2} - 1 \right]^{\mathrm{T}}$$

is in fact a locally efficient score function; that is, if we plug the true skewing function π_0 into (2), the estimator achieves the optimal efficiency. If a skewing function other than π_0 is plugged into (2), then the resulting estimator is still consistent for β, though not necessarily efficient. This consistency property is studied in Ma et al. (2005), wherein Remarks 1–3 are of particular interest.

As a result of the presence of a skewing function π in the estimator, it has been proposed that one take π to be either a parametric model or a cumulative distribution function of a symmetric distribution evaluated at an odd polynomial (Ma & Genton, 2004). In the first approach, if the parametric model is misspecified, the optimal semiparametric efficiency bound is not achieved. In the second approach, a polynomial of fairly high degree might be needed to approximate π_0 sufficiently well. This could cause computational problems that might entail less than optimal semiparametric efficiency in practice.

Fully nonparametric estimation of π is necessary to guarantee optimal semiparametric efficiency and consistent estimation of f. The task is complicated by the constraints $0 \leqslant \pi(y) \leqslant 1$ and $\pi(y) + \pi(-y) = 1$ for all $y \in \mathbb{R}$. We take advantage of an equivalent representation of $\pi(y)$, $H\{p(y)\}$, and estimate p nonparametrically. Here H is an arbitrary fixed function that satisfies the same constraints as π, and p is an odd function. In

the literature, H has been required to be the cumulative distribution function of a symmetric distribution, but it can be more general. For example, $H(y) = \sin(y)/2 + 1/2$ also suffices. Note that the range of p need not be restricted. Taking into consideration the different nature of oddness, a remote property, and kernel-type nonparametric estimation, a local procedure, our nonparametric method uses data local to y and other data local to $-y$ to obtain estimates of π in the neighbourhood of y, and hence $-y$ as well. The implementation modifies locally parametric nonparametric estimation studied in Hjort & Jones (1996), Loader (1996) and Eguchi & Copas (1998), among others. Such nonparametric estimation of the nuisance π is shown to provide an asymptotically efficient estimator for β when used in (2). We could also insert the nonparametric estimator of π in the likelihood and maximize the likelihood with respect to β, hence obtaining a generalized profile likelihood estimator (Severini & Wong, 1992). We argue rigorously that our nonparametric estimator of π is within $o_p(n^{-1/4})$ of π_0. This suggests, but does not prove, that the generalized profile likelihood estimator of β is asymptotically optimal. In addition, the estimators of β and π lead to a nonparametric estimator of the density itself that has typical bias and variance properties.

2. Constrained local likelihood estimator

The univariate semiparametric skew-normal distribution has the probability density function

$$f(x) = \frac{2}{\sigma}\phi\left(\frac{x-\xi}{\sigma}\right)\pi\left(\frac{x-\xi}{\sigma}\right), \tag{3}$$

where ξ and σ are the mean and standard deviation of the population, which is assumed to be normally distributed, and the skewing function π is determined by a selection criterion unknown to the data analyst. Assume that we observe independent and identically distributed observations $X_1, \ldots, X_n$ from f. Our main interest is in estimating the population parameters $\beta = (\xi, \sigma)^{\mathrm{T}}$. A previous study in Ma et al. (2005) has shown that the efficiency of the estimator of β depends on the estimation of π, an infinite-dimensional nuisance parameter subject to the two previously mentioned constraints.

We first focus on the estimation of π, using the representation $\pi(y) = H\{p(y)\}$ to enforce the constraints. For convenience, we require H to be monotone, so that H^{-1} is well defined, and estimate the odd function $p(y) = H^{-1}\{\pi(y)\}$. Note that the monotonic constraint on H is not essential because it suffices to estimate any $p(y)$ that satisfies $\pi(y) = H\{p(y)\}$.

A convenient tool for estimating p is the local polynomial estimator. In a neighbourhood of y_0, we approximate $p(y)$ with a kth-degree polynomial $p_k(y)$, and estimate the coefficients of p_k using the observations close to y_0. To ensure that the resulting function is odd, we estimate $p(y_0)$ for $y_0 > 0$ by $\hat{p}(y_0)$ and then simply define $\hat{p}(-y_0) = -\hat{p}(y_0)$. Note that observations near $-y_0$ contain information about $p(y_0)$, since $p(-y_0) = -p(y_0)$. Thus, a natural way to construct the estimator of $p(y_0)$ is to use data near either y_0 or $-y_0$.

Let $Y_i = (X_i - \xi)/\sigma$ for $i = 1, \ldots, n$. If we assume that (ξ, σ) is the truth, $Y_1, \ldots, Y_n$ have the semiparametric skew-normal density $f(y) = 2\phi(y)\pi(y)$. Along the lines of Hjort & Jones (1996), local likelihood estimators of $\pi(y_0)$ and $\pi(-y_0)$ are obtained by maximizing

$$\frac{1}{n}\sum_{i=1}^{n} K_h(Y_i - y_0)\log[2\phi(Y_i)H\{p_k(Y_i)\}] - \int K_h(t - y_0)2\phi(t)H\{p_k(t)\}dt$$

and

$$\frac{1}{n}\sum_{i=1}^{n} K_h(Y_i + y_0)\log[2\phi(Y_i)H\{p_k(Y_i)\}] - \int K_h(t + y_0)2\phi(t)H\{p_k(t)\}dt$$

respectively, where K is a kernel function and $K_h(\cdot)$ stands for $h^{-1}K(\cdot/h)$. If K is symmetric about 0, then summing the last two expressions yields

$$\frac{1}{n}\sum_{i=1}^{n}\{K_h(Y_i - y_0) + K_h(Y_i + y_0)\}\log H\{p_k(Y_i)\} + C. \tag{4}$$

Here C is a term free of the polynomial p_k. Our estimator of $\pi(y_0)$ is obtained by maximizing (4) with respect to the parameters in p_k. To ensure that $\hat{p}(y) = -\hat{p}(-y)$, the coefficients of the even-order terms in $p_k(y)$ will have opposite signs at y_0 and $-y_0$. A typical $p_k(y)$ has the form $p_k(y) = \alpha_0\text{sign}(y) + \alpha_1 y + \alpha_2\text{sign}(y)y^2 + \alpha_3 y^3 + \ldots$, where $\text{sign}(y) = 1, -1$ or 0 when $y > 0$, $y < 0$ or $y = 0$, respectively. Note that one is not obliged to include all the terms in the local polynomial model p_k. For example, $p_0(y) = \alpha_0\text{sign}(y)$, $p_1(y) = \alpha_1 y$ or $p_1(y) = \alpha_0\text{sign}(y) + \alpha_1 y$ are all applicable. The sign change of the even-order terms in $p_k(y)$ does cause a discontinuity of the estimator at $y = 0$; however, when n is large, the discontinuity tends to diminish. To see why, consider $p_0(y) = \alpha_0\text{sign}(y) = \alpha_0$ at $y > 0$. The resulting estimator $\hat{\pi}(y)$ is given in (8) and satisfies

$$\lim_{y\to 0}\hat{\pi}(y) = \frac{n^{-1}\sum_{i=1}^{n} K_h(Y_i)I_{[0,\infty)}(Y_i)}{n^{-1}\sum_{i=1}^{n} K_h(Y_i)} \doteq \frac{E\{K_h(Y_i)I_{[0,\infty)}(Y_i)\}}{E\{K_h(Y_i)\}} \doteq \frac{\phi(0)\pi_0(0)}{2\phi(0)\pi_0(0)} = \pi_0(0).$$

The order of the polynomial only plays a role in the bias of the nonparametric estimator; using a constant or linear polynomial yields a bias of $O(h^2)$. Since this is sufficient for the purpose of semiparametric estimation, these low-order polynomials are the ones most often used in practice.

Two possible approaches exist for incorporating estimation of π into the semiparametric estimation of β. One is to plug the estimated nuisance function into (2), and the other is to use generalized profile likelihood estimators. In principle, the first approach is robust, in that β is consistently estimated even when π is misspecified, while the second one is not. However, when local polynomials are used, the 'correctness' of π is guaranteed, and hence robustness is not an issue. Compared with generalized profile likelihood, the estimating equation approach has two drawbacks: it requires estimation of π', and treatment of multiple solutions in some cases. The latter drawback often causes some ad hoc techniques to be used in selecting an estimate, whereas in general profile likelihood the global maximizer is an unequivocal estimate. An advantage of the efficient estimating equation approach is that derivation of its asymptotic properties is not too difficult. Implementation of either of the two approaches is straightforward.

Step 1. Choose an appropriate H, a polynomial form p_k and starting values $\tilde{\xi}$ and $\tilde{\sigma}$.

Step 2. Form $Y_i = (X_i - \tilde{\xi})/\tilde{\sigma}$, where $\tilde{\xi}$ and $\tilde{\sigma}$ are the current estimates of ξ and σ. For each $i = 1, \ldots, n$, obtain the polynomial coefficients $\hat{\alpha}_{i0}, \ldots, \hat{\alpha}_{ik}$ by maximizing

$$\sum_{j=1}^{n}\{K_h(Y_j - Y_i) + K_h(Y_j + Y_i)\}\log H\{p_{ik}(Y_j)\}$$

with respect to $\alpha_{i0}, \ldots, \alpha_{ik}$. Here we use p_{ik} to emphasize that the polynomials are local to each X_i.

Step 3. In the efficient estimating equation approach, approximate $\hat{\pi}'$ using numerical differences, plug the resulting $\hat{\pi}$ and $\hat{\pi}'$ into (2) and solve the estimating equations to obtain $\hat{\beta}$. In the generalized profile likelihood approach, update $\hat{\xi}$ and $\hat{\sigma}$ by maximizing

$$\sum_{i=1}^{n} \log\left[\frac{2}{\sigma}\phi\left(\frac{X_i-\xi}{\sigma}\right) H\left\{\hat{p}_{ik}\left(\frac{X_i-\xi}{\sigma}\right)\right\}\right]$$

with respect to ξ and σ.

Step 4. Repeat Steps 2 and 3 until the estimates converge.

Convenient starting values for ξ and σ can be obtained using (2), with $\pi(y) = H(y)$. Note that this guarantees a starting value with root-n consistency. From the propositions in §3, Step 4 in the above algorithm is not really needed; the estimates of ξ and σ after one iteration should already have the desired first-order asymptotic properties. However, in practice with small or moderate sample sizes, one almost always needs several iterations to obtain results agreeing with the theory. The variance of the resulting estimator can be estimated from the inverse of the empirical Fisher information matrix, using either

$$\hat{\mathrm{var}}(\hat{\beta})^{-1} = \sum_{i=1}^{n} S_{\mathrm{eff}}(X_i, \hat{\beta}, \hat{\pi})^{\otimes 2}, \tag{5}$$

when the efficient estimating equation approach is taken, or

$$\hat{\mathrm{var}}(\hat{\beta})^{-1} = \sum_{i=1}^{n} \left(\frac{\partial}{\partial\beta}\log\left[\frac{2}{\sigma}\phi\left(\frac{X_i-\xi}{\sigma}\right) H\left\{\hat{p}_{ik}\left(\frac{X_i-\xi}{\sigma},\beta\right)\right\}\right]\right)^{\otimes 2}\Bigg|_{\xi=\hat{\xi},\sigma=\hat{\sigma}}, \tag{6}$$

when the likelihood approach is taken. Here $A^{\otimes 2}$ represents the operation AA^{T}. Note that in addition to the fact that we evaluate $\hat{p}_{ik}$ at $(X_i-\xi)/\sigma$, which involves β, the form of the function $\hat{p}_{ik}$ itself also depends on β. To emphasize this dependence on β and to avoid confusion, we use the notation $\hat{p}(\cdot,\beta)$ in (6). Consequently, the partial derivative in (6) is not simply with respect to the β appearing in $(X_i-\xi)/\sigma$. We therefore propose to approximate the partial derivative in (6) using numerical differences of the two different $\hat{\pi}$ functions, $\hat{\pi}\{(X_i-\xi)/\sigma,\beta\}$, estimated using different β's. If the profile likelihood approach achieves the optimal semiparametric efficiency, then the partial derivative in (6) estimates S_{eff}.

Since we actually estimate the nuisance function $\pi(y)$, we have a density estimator. It is not difficult to see that this estimator has asymptotic properties typical of a nonparametric density estimator; that is, it has bias of order h^2 and variance of order $(nh)^{-1}$.

As pointed out by a referee, a much simpler estimator for π exists. Suppose we ignore the constraints on f and compute a consistent nonparametric estimator $\hat{f}(x)$. Treating (ξ,σ) as the truth, we have $\hat{f}(x) \doteq 2\sigma^{-1}\phi\{(x-\xi)/\sigma\}\pi\{(x-\xi)/\sigma\}$, or equivalently $\hat{f}(\xi+\sigma y) \doteq 2\sigma^{-1}\phi(y)\pi(y)$. Since this relationship holds for any y, $\hat{f}(\xi-\sigma y) \doteq 2\sigma^{-1}\phi(y)\pi(-y)$, and thus

$$\hat{\pi}(y;\xi,\sigma) = \frac{\hat{f}(\xi+\sigma y)}{\hat{f}(\xi+\sigma y)+\hat{f}(\xi-\sigma y)} \doteq \pi(y). \tag{7}$$

So long as $\hat{f}$ is nonnegative, it is easily verified that $\hat{\pi}(\cdot;\xi,\sigma)$ satisfies the requisite constraints for each (ξ,σ). One may now maximize, with respect to (ξ,σ), a likelihood obtained by substituting $\hat{\pi}(\cdot;\xi,\sigma)$ for π. Note that here the functional form of the likelihood requires no initial estimates of ξ_0 and σ_0. The existence of such a functional form is a blessing of the skew-symmetric model. In most semiparametric models, it is not possible

to obtain an explicit nuisance functional form that is free of the parameters of interest. Hence, while (7) has the virtue of simplicity, it does not necessarily generalize to more complicated settings as does the local likelihood approach. Even in our skew-symmetric setting, local likelihood has an advantage over (7), as will be seen in §5.

3. Asymptotic properties

In this section, we prove the semiparametric efficiency property of the estimating equation approach and ostensibly take a step towards proving the same property for the generalized profile likelihood approach. Proposition 1 below states that a generally desirable $o_p(n^{-1/4})$ rate of the nuisance parameter holds for an estimator of π that differs only slightly from the local likelihood estimator with polynomial p_k equal to a constant. The results stated in this section are specific to this particular estimator, but similar results can be established for higher-order local polynomial models, albeit at the expense of more complicated proofs.

Our local likelihood estimator of $\pi(y)$ for p_k identical to a constant α is a solution of the equation

$$\sum_{i=1}^{n}\{K_h(Y_i - y) + K_h(Y_i + y)\}H_1\{\alpha\text{sign}(Y_i)\}\text{sign}(Y_i) = 0,$$

where $H_1(\cdot)$ is the derivative of $\log H(\cdot)$. It is easily verified that, for $y > 0$, the unique solution to this equation is

$$\hat{\pi}(y) = H(\hat{\alpha}) = \frac{n^{-1}\sum_{i=1}^{n}\{K_h(Y_i - y) + K_h(Y_i + y)\}I_{[0,\infty)}(Y_i)}{n^{-1}\sum_{i=1}^{n}\{K_h(Y_i - y) + K_h(Y_i + y)\}}. \tag{8}$$

The denominator of this estimator is the sum of two density estimators, and estimates $2\phi(y)\pi_0(y) + 2\phi(-y)\pi_0(-y) = 2\phi(y)$. This suggests another estimator,

$$\tilde{\pi}(y) = \frac{1}{2\phi(y)nh}\sum_{i=1}^{n}\left\{K\left(\frac{Y_i - y}{h}\right) + K\left(\frac{Y_i + y}{h}\right)\right\}I_{[0,\infty)}(Y_i).$$

The estimator $\tilde{\pi}$ is the subject of Proposition 1, in which we show that the norm

$$||\tilde{\pi} - \pi_0|| = \left[\int_0^\infty\{\tilde{\pi}(y) - \pi_0(y)\}^2\phi(y)\,dy\right]^{1/2}$$

is $o_p(n^{-1/4})$. This result and the subsequent Corollary 1 allow us to prove efficiency of the estimating equations approach.

PROPOSITION 1. *Let $X_1, \ldots, X_n$ be independent and identically distributed with common density $2\phi\{(x - \xi_0)/\sigma_0\}\pi_0\{(x - \xi_0)/\sigma_0\}/\sigma_0$. Let $b \in (0, 1)$ be a constant, let $m = [nb]$, the integer part of nb, and suppose that $\tilde{\xi}$ and $\tilde{\sigma}$ are estimators constructed from the observations $X_{m+1}, \ldots, X_n$ and satisfying $\tilde{\xi} - \xi_0 = O_p(n^{-1/2})$ and $\tilde{\sigma} - \sigma_0 = O_p(n^{-1/2})$. Define $\tilde{\pi}(y)$ to be the estimator*

$$\tilde{\pi}(y) = \frac{1}{2\phi(y)mh}\sum_{i=1}^{m}\left\{K\left(\frac{Y_i - y}{h}\right) + K\left(\frac{Y_i + y}{h}\right)\right\}I_{[0,\infty)}(Y_i), \tag{9}$$

where $Y_i = (X_i - \tilde{\xi})/\tilde{\sigma}$, $i = 1, \ldots, m$. Assume also that the following conditions hold:

(i) *the kernel K integrates to 1, is symmetric about 0, has support $(-1, 1)$ and is continuous on $[-1, 1]$;*

(ii) *the bandwidth h is such that $C^{-1}n^{-1/5} < h < Cn^{-1/5}$ for all n, where $C > 1$ is a constant and can be arbitrarily large;*

(iii) *the skewing function π_0 is twice differentiable and has bounded first and second derivatives.*

It then follows that $||\tilde{\pi} - \pi_0|| = o_p(n^{-1/4})$.

All proofs are given in the Appendix. Note that the proposition assumes that the initial estimates $\tilde{\xi}$ and $\tilde{\sigma}$ are computed from a subset of the data, and $\tilde{\pi}$ is computed from the complementary observations. This is done only to make the proof simpler. Clearly, if both $(\tilde{\xi}, \tilde{\sigma})$ and $\tilde{\pi}$ are computed from all the observations, then the estimator $\tilde{\pi}$ will be even more efficient.

It is also important to note that the result holds for bandwidths of the order $n^{-1/5}$. In practice, methods such as crossvalidation are often used to choose h, and most such methods will yield a bandwidth of the asymptotically optimal order, i.e. $n^{-1/5}$.

The following corollary is a direct application of Proposition 1.

COROLLARY 1. *Under the same conditions as in Proposition* 1*, if we approximate π_0' using $\tilde{\pi}'(y) \equiv \{\tilde{\pi}(y + n^{-1/4}) - \tilde{\pi}(y - n^{-1/4})\}/(2n^{-1/4})$, then $||\tilde{\pi}' - \pi_0'|| = o_p(1)$.*

Proposition 1 is in fact much stronger than what we need for our proof of semiparametric efficiency of the estimating equation approach. We only need $||\hat{\pi} - \pi_0|| = o_p(1)$ and $||\hat{\pi}' - \pi_0'|| = o_p(1)$. However, the $o_p(n^{-1/4})$ rate is often desirable in semiparametric methods, and so we hope that in future work it will facilitate a proof of efficiency of the generalized profile likelihood approach.

Under regularity conditions, Proposition 1 leads to the efficiency of the semiparametric estimator of β, as given in Proposition 2.

PROPOSITION 2. *Let $\hat{\beta}$ be the estimator obtained from* (2) *using observations $X_1, \ldots, X_m$, where π is replaced by the estimator* (9) *computed from $X_{m+1}, \ldots, X_n$. If $m = n - n^{\epsilon}$ for some $0 < \epsilon < 1$, and appropriate regularity conditions hold, then $\hat{\beta}$ is semiparametric efficient and*

$$\sqrt{n}(\hat{\beta} - \beta_0) \to N\left\{0, \sigma_0^2 E\left(\left[\frac{X_i - \xi_0}{\sigma_0}\left\{2\pi_0\left(\frac{X_i - \xi_0}{\sigma_0}\right) - 1\right\} - 2\pi_0'\left(\frac{X_i - \xi_0}{\sigma_0}\right), \frac{(X_i - \xi_0)^2}{\sigma_0^2} - 1\right]^{\otimes 2}\right)^{-1}\right\}$$

in distribution as $n \to \infty$.

Note that the variance matrix in Proposition 2 is the inverse of the efficient Fisher information matrix, which is estimated in (5). The regularity conditions mainly ensure the nonsingularity of the information matrix and its inverse, and sufficient smoothness of the functions involved in the asymptotic expansion. We omit these technicalities and instead refer the reader to Newey (1990) for details. Note that, again, the separation of the observations into two groups, one for estimating π and the other for estimating β, is to avoid technical complexity in the proof. In practice, it is certainly not necessary to use only part of the data to compute either of the estimates.

We find a similar semiparametric efficiency result for the generalized profile likelihood approach harder to derive. If a strict profile likelihood method is used, then an $o_p(n^{-1/4})$

rate for the estimated nuisance function generally suffices for efficiency of the parameter of interest (Murphy & van der Vaart, 2000). However, most nonparametric estimators are not, strictly speaking, maximum likelihood estimators, and so the result cannot be used directly. There are many possibilities for relaxing the strict maximization requirement on the nuisance parameter, resulting in different versions of generalized profile likelihood, such as a sieve maximum likelihood estimator. However, a general treatment of arbitrary estimators of the nuisance function in profile likelihood seems unavailable, and hence the efficiency analysis seems to be a case-by-case exercise. In our problem, the local maximum likelihood estimator converges to a standard maximum likelihood estimator when $n \to \infty$, and hence $h \to 0$, and so we believe that the performance of the local maximum likelihood estimator will closely mimic the maximum likelihood estimator when h is sufficiently small. However, we have not found a rigorous proof.

In contrast, it is relatively easy to obtain asymptotic properties of the estimator based on (7). Since the form of $\hat{\pi}$ does not depend on the estimates of ξ and σ, the resulting estimator is a direct maximizer of $\sum_{i=1}^{n} \log[2\sigma^{-1}\phi\{(x_i - \xi)/\sigma\}\hat{\pi}\{(x_i - \xi)/\sigma\}]$. Treating $\hat{\pi}$ as a known function that approximates π, we can easily verify that the resulting estimators of ξ and σ are consistent and have the same asymptotic variance as given in Proposition 2; that is, they are efficient estimators. The complete proof of the asymptotic efficiency is available at the webaddress in the Appendix.

In contrast to the usual concern with bandwidth selection in nonparametric estimation, the bandwidth h is not nearly so important in estimating β. In fact, h does not enter into the first-order asymptotics of the final estimator for β. Thus, any h that yields a convergence rate of $o_p(n^{-1/4})$ for the nuisance parameter guarantees the asymptotic efficiency of $\hat{\beta}$. However, if one is interested in estimating the density itself, then the bandwidth h is as important as in ordinary nonparametric density estimation, and the usual bandwidth of order $n^{-1/5}$ is needed to minimize the mean squared error. Crossvalidation can be used to obtain a bandwidth that estimates an optimal h, and we implement such an approach in our real-data example of §5.

4. SIMULATION STUDIES

We undertook a simulation study to investigate the performance of our estimators of ξ, σ and π. We generated 500 datasets with sample size $n = 100$ from a distribution with density $2\phi\{(x - \xi)/\sigma\}\Phi[\sin\{c(x - \xi)/\sigma\}]/\sigma$, where $\xi = 3$, $\sigma = 1$ and $c = -3$. Eight different estimators, (a)–(h), were implemented. Estimator (a) uses the estimating equation (2), with a misspecified skewing function $\pi(x) = H(\alpha x) = 1/\{1 + \exp(-\alpha x)\}$. The parameter α is obtained from an initial estimating step where ξ, σ and α are estimated jointly through maximum likelihood. Estimator (b) is the local linear estimator proposed in §2, with $p_k(x) = \alpha x$ and the same H as in estimator (a). Step 3 in (b) is carried out through solving the estimating equation (2). Estimator (c) is very similar to (b), except that the Step 3 in (c) is carried out using the generalized profile likelihood approach. Thus, estimators (b) and (c) can be viewed as local versions of (a). Estimators (d)–(f) are the same as (a)–(c), respectively, except that they use a different choice of H, namely $H(\alpha x) = \sin(\alpha x)/2 + 1/2$. Estimator (g) is the maximum likelihood estimator, where π is estimated by (8). Estimator (h) is constructed in the same way as (a), except that we use the true model for the skewing function, i.e. we adopt

Table 1. *Simulation study. Means, empirical standard errors* (emp. se.)*, estimated standard errors* (est. se.) *and empirical coverage probabilities* (cov.) *of nominal* 95% *confidence intervals for each of the estimators* (a)–(h). *Five hundred simulations were performed for each sample size*

	$\hat{\xi}(3)$				$\hat{\sigma}(1)$			
	mean	est. se.	emp. se.	cov.	mean	est. se.	emp. se.	cov.
				$n = 100$				
(a)	3·0082	0·3448	0·5897	0·9720	1·0437	0·1107	0·1994	0·9680
(b)	3·0143	0·0967	0·0778	0·9340	0·9918	0·0705	0·0755	0·9560
(c)	2·9920	0·0744	0·0653	0·8920	0·9890	0·0688	0·0720	0·9400
(d)	2·9840	0·2637	0·1922	0·8340	1·0124	0·1463	0·1021	0·9340
(e)	3·0542	0·1186	0·0861	0·8500	0·9968	0·0725	0·0779	0·9600
(f)	2·9849	0·0754	0·0648	0·8880	0·9890	0·0691	0·0674	0·9180
(g)	2·9978	0·0747	0·0705	0·9400	0·9892	0·0687	0·0731	0·9540
(h)	2·9983	0·0665	0·0663	0·9340	0·9888	0·0686	0·0706	0·9400
				$n = 200$				
(a)	2·9627	0·2845	0·6112	0·9820	1·0420	0·0693	0·8219	0·9680
(b)	3·0049	0·0489	0·0491	0·9420	1·0029	0·0495	0·0515	0·9520
(c)	2·9980	0·0497	0·0457	0·9160	1·0028	0·0497	0·0513	0·9500
(d)	2·9931	0·2098	0·1689	0·8760	1·0183	0·0936	0·0710	0·9320
(e)	3·0154	0·0578	0·0519	0·9020	1·0040	0·0496	0·0518	0·9580
(f)	2·9923	0·0484	0·0472	0·9300	1·0026	0·0498	0·0497	0·9400
(g)	3·0012	0·0497	0·0476	0·9240	1·0029	0·0496	0·0515	0·9520
(h)	3·0011	0·0449	0·0455	0·9460	1·0027	0·0496	0·0503	0·9440

$H(\alpha x) = \Phi[\sin\{\alpha(x - \xi)/\sigma\}]$. Thus, estimator (h) is the optimal semiparametric estimator in the sense that it is the most efficient. A bandwidth $h = 0{\cdot}5n^{-1/5}$ is used in the local linear estimators.

The same simulation was repeated with sample size $n = 200$. We are interested in comparing the performance of estimators (a)–(g), with estimator (h) as a benchmark. Based on our asymptotic results, we expect that results for estimators (b), (c), (e), (f) and (g) will be similar to those for (h). The results of the simulation are given in Table 1.

Several remarks are worth making. First, none of the estimators of ξ or σ exhibited substantial bias, and, except for scheme (a), the empirical standard errors decreased from $n = 100$ to $n = 200$. Secondly, the estimated variances of the global linear estimators (a) and (d) do not match the empirical variance well. However, when $n = 200$, the two variances match reasonably well for the other estimators where the nuisance function is estimated nonparametrically. Thirdly, the variances of the local linear estimators (b), (c), (e) and (f) are rather insensitive to the choice of the function H, whereas the global estimators (a) and (b) depend heavily on this choice. In reality, it is certainly not trivial to find a 'good' H. As a result, one would have to go to higher-order polynomials, i.e. the flexible skew-normal distributions proposed in Ma & Genton (2004), to get a better estimator. Fourthly, with $n = 200$, the variances of all estimators based on nonparametric estimators of π are fairly similar to the optimal one. Hence, the asymptotic efficiency of the local linear estimator apparently does not require a huge sample size to exhibit itself. Finally, estimator (g) has the best performance among all the estimators, possibly benefiting from its computational simplicity.

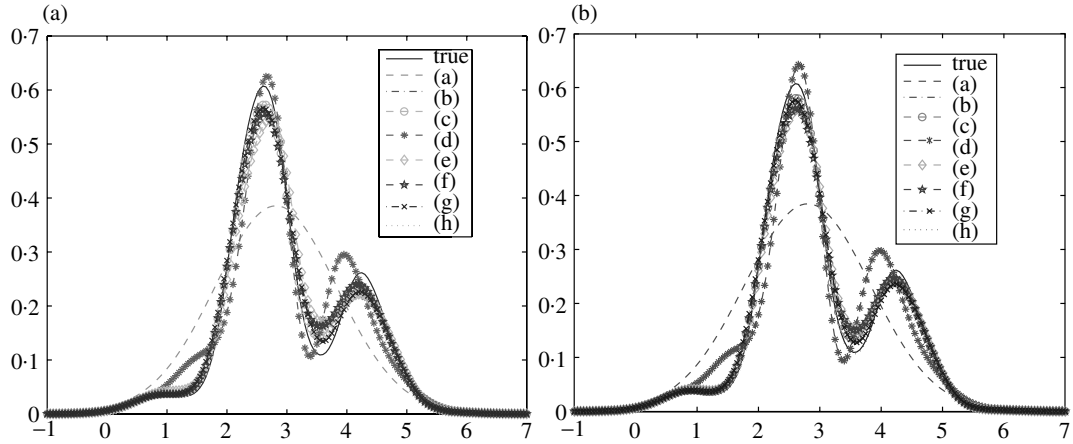

Fig. 1. Simulation study. The average estimated densities using different estimators (a)–(h), together with the true density, based on 500 simulations, for (a) $n = 100$ and (b) $n = 200$.

Figure 1 displays the averages of the estimated densities from estimators (a)–(h). The resulting density of the global linear estimators (a) and (d) does not bear much resemblance to the true density. Of course we do not expect them to; in fact, the result from (d) is surprisingly close to the true density. The estimated densities for all the other estimators are reasonably close to the truth, especially when $n = 200$, with some bias at the peaks and troughs. When the sample size increases from $n = 100$ to $n = 200$, the improved density estimation is not evident from inspecting the plotted curves, except for a small improvement around the smaller mode.

The software used in the simulation is available at the webaddress in the Appendix.

5. AN EXAMPLE

Here we apply the local linear estimators proposed in §3 to datasets consisting of white cell counts of Australian athletes. Histograms of cell counts for 102 male and 100 female athletes are shown in Fig. 2.

Under the assumption that skew-normal models are appropriate, our goal is to infer from each of these datasets the parameters ξ and σ and also the underlying densities for the male and female populations of athletes. Under the further assumption that the samples are selectively chosen, as discussed in §1, from the populations of all adult males and females in Australia, the parameters ξ and σ have the interpretation of being the mean and standard deviation, respectively, of each of the latter populations.

We consider two types of estimator of π in our analysis, namely a local likelihood estimator obtained by maximizing an expression of the form (4), and an estimator as in (7) with $\hat{f}$ equal to a kernel density estimator. A version of L_2 crossvalidation was used to select the bandwidth of our local likelihood estimator. Given a random sample $X_1, \ldots, X_n$ of observations from f and an estimator $\hat{f}_h$ with smoothing parameter h, an L_2 crossvalidation curve is defined (Bowman, 1984) by

$$\mathrm{CV}(h) = \int_{-\infty}^{\infty} \hat{f}_h(x)^2 \, dx - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_h^i(X_i), \quad h > 0, \tag{10}$$

where $\hat{f}_h^i$ is an estimator computed from the $n - 1$ observations that exclude X_i, $i = 1, \ldots, n$.

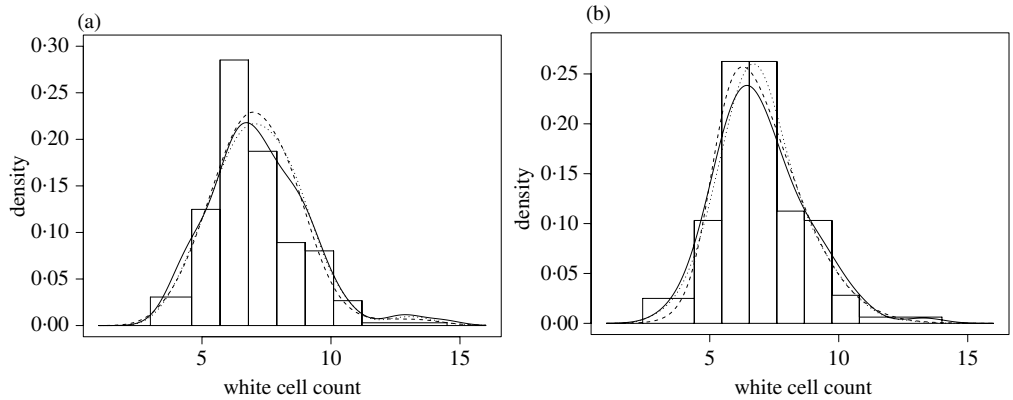

Fig. 2. White cell counts data. Histogram and density of white cell counts for (a) 102 male and (b) 100 female Australian athletes. The solid lines are kernel estimates, the dotted lines are skew-normal estimates that estimate π via (7), and the dashed lines are skew-normal estimates that use local likelihood to estimate π.

In our setting, for given ξ and bandwidth h, write

$$\hat{f}_h(x;\xi) = \frac{2}{\sigma(\xi)}\phi\left(\frac{x-\xi}{\sigma(\xi)}\right)\hat{\pi}_h(x-\xi), \tag{11}$$

where $\hat{\pi}_h(x-\xi)$ is a local likelihood estimator of $\pi(x-\xi)$ and $\sigma^2(\xi) = \sum_{i=1}^n (X_i-\xi)^2/n$. Note that the π function here is defined slightly differently in that we absorb the standard deviation σ into π. Now define

$$\text{CV}(h,\xi) = \sum_{i=1}^{n} \hat{f}_h(X_{(i)};\xi)^2(S_i - S_{i-1}) - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_h^i(X_i;\xi), \tag{12}$$

where $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ are the ordered X_i's, $S_i = (X_{(i)} + X_{(i+1)})/2$, $i = 0, 1, \ldots, n$, with $X_{(0)} = 2X_{(1)} - X_{(2)}$ and $X_{(n+1)} = 2X_{(n)} - X_{(n-1)}$, and $\hat{f}_h^i(\cdot\,;\xi)$ is an estimator computed from the $n-1$ observations excluding X_i, $i = 1, \ldots, n$. The first sum on the right-hand side of (12) is an approximation of $\int_{-\infty}^{\infty} \hat{f}_h(x;\xi)^2\,dx$. Our choice of bandwidth is $\hat{h}$, where $(\hat{h}, \hat{\xi})$ minimizes $\text{CV}(h,\xi)$ with respect to (h,ξ).

In applying (4), we took K to be a standard normal density, $p_k(y) \equiv \alpha y$, and H to be a logistic curve, i.e. $H(y) \equiv (1+e^{-y})^{-1}$. The global minimizer of (12) was $(h,\xi) = (1{\cdot}03, 8{\cdot}26)$ for the male athletes and $(h,\xi) = (\infty, 5{\cdot}13)$ for the women. For the women's data, the resulting density estimate is essentially a parametric estimate of the form

$$\hat{f}(x) = \frac{2}{\hat{\sigma}}\phi\left(\frac{x-\hat{\xi}}{\hat{\sigma}}\right)H\{\hat{\alpha}(x-\hat{\xi})\},$$

where $\hat{\alpha}$ is the same for all x.

Profile likelihood curves were computed using estimates of π based on the bandwidths chosen by our crossvalidation scheme. These curves have the form

$$L(\xi) = \prod_{i=1}^{n} \hat{f}_{\hat{h}}(X_i;\xi),$$

where $\hat{f}_h(\cdot\,;\xi)$ is defined as in (11). The maximizers of $L(\xi)$ were $\hat{\xi}_M = 8{\cdot}12$ for the men and $\hat{\xi}_F = 5{\cdot}11$ for the women. These agree quite well with the values of ξ that minimized the crossvalidation curves.

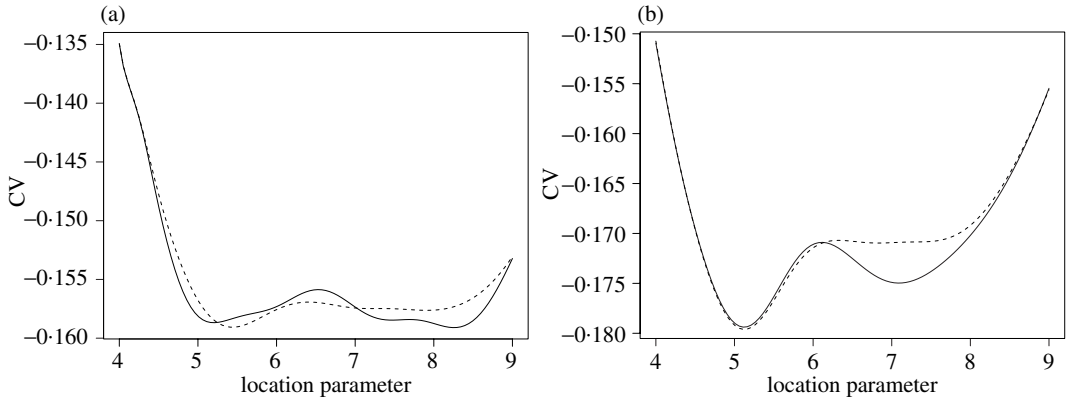

Fig. 3. White cell count data. Crossvalidation curves for (a) male and (b) female athletes. Each curve is a function of ξ for a fixed bandwidth h. The bandwidth is fixed at 1·03 for the males, and 1·5 for the females respectively for the solid curves, and at 100 for the dashed curves.

The sample means for the men and women are 7·22 and 6·99, respectively. The previous results thus seem to indicate that the two datasets are quite different, in that $\hat{\xi}_M > 7{\cdot}22$, whereas $\hat{\xi}_F < 6{\cdot}99$. However, a closer inspection of the two crossvalidation surfaces shows that the results are not as different as they appear to be. Plots of $\mathrm{CV}(1{\cdot}03, \xi)$ and $\mathrm{CV}(100, \xi)$ for the male athletes are shown in Fig. 3(a). What is interesting here is that the local minimum at $(h, \xi) = (100, 5{\cdot}4)$ is only slightly larger than the global minimum. The difference seems small enough that the two estimates of ξ should be deemed almost equally credible. Indeed, comparing the two density estimates, $(1{\cdot}03, 8{\cdot}26)$ and $(100, 5{\cdot}4)$, via a penalized likelihood approach actually favours the estimate with $\hat{\xi} = 5{\cdot}4$. Crossvalidation curves for the female athletes are shown in Fig. 3(b). Here an estimate of about 5·11 is unambiguously supported by either curve. Furthermore, the value of ξ that maximizes $L(\xi)$ at $h = 1{\cdot}5$ is also 5·11.

To estimate π as in (7), we took $\hat{f}$ to be a kernel estimate with a Gaussian kernel and bandwidth chosen to minimize the crossvalidation function (10). The optimal bandwidths were 0·71 and 0·76 for the men and women, respectively. The respective maximizers of the profile likelihood (7) were 8·08 and 7·36. The fact that the latter estimate differs markedly from the estimate of 5·11 obtained in our previous analysis raises the question as to which estimate is better supported by the data. A simple likelihood analysis that takes into account the fact that each density estimate uses a different effective number of parameters gives stronger support to the estimate of ξ equal to 5·11.

Various estimates of the densities are shown in Fig. 2, and estimates of π are shown in Fig. 4. Note that the two skew-normal density estimates for the women in Fig. 2(b) are not remarkably different, and yet the corresponding estimates of π in Fig. 4(b) are quite different. The difference in the two estimates of π undoubtedly explains the large difference in the corresponding estimates of ξ.

Our analysis illustrates an advantage of the local likelihood method over the simpler approach based on estimating π as in (7). The former method collapses to a parametric version of the skew-normal model when the bandwidth is large. Such a property is also possessed by the methods of Hjort & Jones (1996) and others. The density estimator based on (7) collapses to a Gaussian curve for large h, since (7) converges to 1/2 as the bandwidth of $\hat{f}$ tends to ∞. This means that an estimate based on (7) can only account for skewness through a nonparametric estimate of π. When a parametric skew-normal model for the

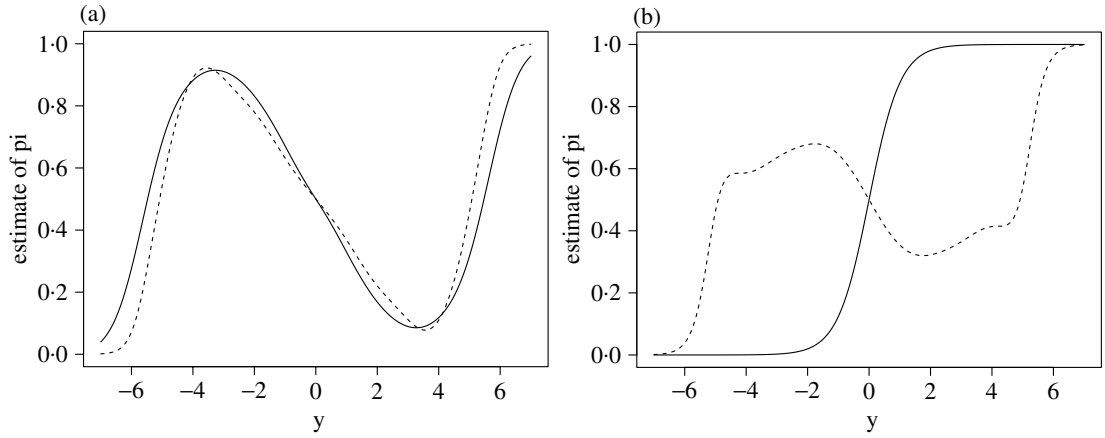

Fig. 4. White cell counts data. Estimates of π for the (a) male and (b) female athletes' data. The solid curves are local likelihood estimates that optimize the crossvalidation function (12), while the dashed estimates have the form (7) with $\hat{f}$ a crossvalidated kernel estimate.

data is warranted, as appears to be the case for the female athletes' data, the local likelihood estimate can thus produce a more efficient estimate of the underlying density.

Acknowledgement

This work was supported by grants from the U.S. National Cancer Institute and the U.S. National Science Foundation.

Appendix

Technical proofs

We briefly outline essential parts of the proofs here. Detailed proofs are available at http://www.stat.tamu.edu/~ma.

Outline Proof of Proposition 1. Since the kernel K has support $(-1, 1)$, $\hat{\pi}(y) = 0$ for all $y \geqslant Y_{(m)} + h$, where $Y_{(m)}$ is the largest of $Y_1, \ldots, Y_m$. The square of the norm of interest is thus

$$\begin{aligned}&\int_0^\infty \{\hat{\pi}(y) - \pi_0(y)\}^2 \phi(y)\, dy \\ &= \int_0^h \{\hat{\pi}(y) - \pi_0(y)\}^2 \phi(y)\, dy + \int_h^{Y_{(m)}+h} \{\hat{\pi}(y) - \pi_0(y)\}^2 \phi(y)\, dy + \int_{Y_{(m)}+h}^\infty \pi_0^2(y)\phi(y)\, dy.\end{aligned}$$

We analyze the three terms on the right-hand side of the last equation separately. It can be shown that the last term is bounded by $\int_{Y_{(m)}}^\infty 2\pi_0(y)\phi(y)\, dy/2$, and

$$\begin{aligned}&\mathrm{pr}\left\{\int_{Y_{(m)}}^\infty 2\pi_0(y)\phi(y)dy > \frac{\epsilon}{\sqrt{n}}\right\} \leqslant \mathrm{pr}\left\{Y^*_{(m)} < \eta F^{-1}\left(1 - \frac{\epsilon}{\sqrt{n}}\right)\right\} \\ &+\mathrm{pr}\left\{\frac{\xi_0 - \tilde{\xi}}{\sigma_0} < -\frac{\eta - 1}{2} F^{-1}\left(1 - \frac{\epsilon}{\sqrt{n}}\right)\right\} + \mathrm{pr}\left(1 - \frac{\tilde{\sigma}}{\sigma_0} < -\frac{\eta - 1}{2}\right), \qquad \text{(A1)}\end{aligned}$$

where F is the cumulative distribution function of the density $f(y) = 2\pi_0(y)\phi(y)$, ϵ is an arbitrary positive constant, $Y^*_{(m)}$ is the largest of $(X_i - \xi_0)/\sigma_0$, $i = 1, \ldots, m$, and $\eta > 1$. Letting $\eta - 1$ have the

form $\delta/\sqrt{n}$ for some constant δ, we can show that the right-hand side of (A1) can be made smaller than any positive number for n sufficiently large.

For any sequence of positive constants C_m, we have

$$\begin{aligned}&\mathrm{pr}\left[\int_h^{Y_{(m)}+h}\{\hat\pi(y)-\pi_0(y)\}^2\phi(y)\,dy>\frac{\epsilon}{\sqrt{n}}\right]\\ &\leqslant \mathrm{pr}\left[\int_h^{C_m}\{\hat\pi(y)-\pi_0(y)\}^2\phi(y)\,dy>\frac{\epsilon}{\sqrt{n}}\right]+\mathrm{pr}(Y_{(m)}+h\geqslant C_m).\end{aligned}$$

Obviously, $\mathrm{pr}(Y_{(m)}+h\geqslant C_m)=o(1)$ if $C_m=(C\log n)^{1/2}$ for $C>4$, a is chosen close enough to 1, and $n\to\infty$. In the sequel, E_0 denotes expectation with respect to the joint distribution of $\tilde\xi$ and $\tilde\sigma$, while pr* and E^* denote probability and expectation with respect to the conditional distribution of $X_1,\ldots,X_n$ given $\tilde\xi$ and $\tilde\sigma$. Defining

$$G_n=G_n(\epsilon)=\mathrm{pr}^*\left[\int_h^{C_m}\{\hat\pi(y)-\pi_0(y)\}^2\phi(y)\,dy>\frac{\epsilon}{\sqrt{n}}\right],$$

we have

$$\mathrm{pr}\left[\int_h^{C_m}\{\hat\pi(y)-\pi_0(y)\}^2\phi(y)\,dy>\frac{\epsilon}{\sqrt{n}}\right]=E_0G_n.$$

By the theorem in Section 1·3·6 of Serfling (1980), it suffices to show that G_n converges in probability to 0 for each $\epsilon>0$. By Markov's inequality and Fubini's theorem,

$$G_n\leqslant\frac{\sqrt{n}}{4\epsilon}\int_h^{C_m}\frac{1}{\phi^2(y)}E^*\{\hat f_h(y)-f(y)\}^2\phi(y)\,dy,$$

where $\hat f_h(y)=(mh)^{-1}\sum_{i=1}^m K\{(y-Y_i)/h\}$. Defining

$$f_n(y)=2\left(\frac{\tilde\sigma}{\sigma}\right)\phi\left(\frac{\tilde\sigma y+\tilde\xi-\xi}{\sigma}\right)\pi_0\left(\frac{\tilde\sigma y+\tilde\xi-\xi}{\sigma}\right),$$

we only need to show that

$$\sqrt{n}\int_h^{C_m}E^*\{\hat f_h(y)-f_n(y)\}^2\{\phi(y)\}^{-1}\,dy \tag{A2}$$

and

$$\sqrt{n}\int_h^{C_m}\{f_n(y)-f(y)\}^2\{\phi(y)\}^{-1}\,dy \tag{A3}$$

converge in probability to 0. It is easily verified that (A3) is $O_p(n^{-1/2})$. Standard results from density estimation (Silverman, 1986) yield

$$\begin{aligned}E^*\{\hat f_h(y)-f_n(y)\}^2=&\frac{1}{mh}\int_{-1}^1K^2(z)f_n(y-hz)\,dz-\frac{1}{m}\left\{\int_{-1}^1K(z)f_n(y-hz)\,dz\right\}^2\\&+\frac{h^4}{4}\left\{\int_{-1}^1z^2K(z)f_n''(\hat y_n)\,dz\right\}^2,\end{aligned}$$

where $\hat y_n$ is between y and $y-hz$. Each of the three terms on the right-hand side of the last equation can be shown to be $o_p(1)$, and hence (A2) holds.

The only remaining part of the proof is to show that $\sqrt{n}\int_0^h\{\hat\pi(y)-\pi_0(y)\}^2\phi(y)\,dy\to0$ in probability. Since this integral is over an interval of length h, the only potential problem here is that the bias, conditional on $\tilde\xi$ and $\tilde\sigma$, of the kernel estimator of $f(y)$ on $[0,h]$ is of order h, rather than

h^2 as when $y > h$. The resulting integral of the squared bias is $O_p(h^3)$, which is still small enough since $\sqrt{n}h^3 = O(n^{-1/10})$. This completes the proof of Proposition 1. □

Proof of Corollary 1. Approximating the derivative π_0' by a difference, we obtain

$$\begin{aligned}
&\lim_{n\to\infty} ||\tilde{\pi}_n'(y) - \pi_0'(y)|| \\
&= \lim_{n\to\infty} ||\frac{\tilde{\pi}_n(y+n^{-1/4}) - \tilde{\pi}_n(y-n^{-1/4})}{2n^{-1/4}} - \lim_{n\to\infty} \frac{\pi_0(y+n^{-1/4}) + \pi_0(y-n^{-1/4})}{2n^{-1/4}}|| \\
&= \lim_{n\to\infty} ||\tilde{\pi}_n(y+n^{-1/4}) - \tilde{\pi}_n(y-n^{-1/4}) - \pi_0(y+n^{-1/4}) + \pi_0(y-n^{-1/4})||/(2n^{-1/4}) \\
&\leqslant \lim_{n\to\infty} ||\tilde{\pi}_n(y+n^{-1/4}) - \pi_0(y+n^{-1/4})||/(2n^{-1/4}) \\
&\quad + \lim_{n\to\infty} ||\tilde{\pi}_n(y-n^{-1/4}) - \pi_0(y-n^{-1/4})||/(2n^{-1/4}) = 0
\end{aligned}$$

with probability 1. The last equality is a result of the uniform convergence rate of $o_p(n^{-1/4})$ from Proposition 1. □

Proof of Proposition 2. From Proposition 1, we obtain $||\tilde{\pi} - \pi_0|| = o_p(1)$ and $||\tilde{\pi}' - \pi_0'|| = o_p(1)$. In the following, we show that, when the true nuisance function π_0 is replaced by its estimated version $\tilde{\pi}$, the resulting estimated efficient score still yields an efficient estimator, given that the estimation of the nuisance parameter is performed on separate observations. For notational simplicity, we treat β as if it were a scalar; the general vector case is similar.

Denote the efficient score function by $S_{\text{eff}}(X, \beta_0, \pi_0)$ and let $||\tilde{\pi} - \pi_0|| = o_p(n^{-1/4})$. Assume that $\hat{\beta}$ is a solution of the efficient estimating equation. Then we have

$$\begin{aligned}
0 &= \frac{1}{m^{1/2}} \sum_{i=1}^{m} S_{\text{eff}}(X_i, \hat{\beta}, \tilde{\pi}) \\
&= \frac{1}{m^{1/2}} \sum_{i=1}^{m} S_{\text{eff}}(X_i, \beta_0, \pi_0) + \frac{1}{m} \sum_{i=1}^{m} \frac{\partial S_{\text{eff}}}{\partial \beta}(X_i, \beta^*, \tilde{\pi})\{m^{1/2}(\hat{\beta} - \beta_0)\} \\
&\quad + \frac{1}{m^{1/2}} \sum_{i=1}^{m} \{S_{\text{eff}}(X_i, \beta_0, \tilde{\pi}) - S_{\text{eff}}(X_i, \beta_0, \pi_0)\} \\
&= \frac{1}{m^{1/2}} \sum_{i=1}^{m} S_{\text{eff}}(X_i, \beta_0, \pi_0) + \frac{1}{m} \sum_{i=1}^{m} \left\{ \frac{\partial S_{\text{eff}}}{\partial \beta}(X_i, \beta_0, \pi_0) + o_p(1) \right\} \{m^{1/2}(\hat{\beta} - \beta_0)\} \\
&\quad + \frac{1}{m^{1/2}} \sum_{i=1}^{m} \{S_{\text{eff}}(X_i, \beta_0, \tilde{\pi}) - S_{\text{eff}}(X_i, \beta_0, \pi_0)\},
\end{aligned}$$

where β^* is between β_0 and $\hat{\beta}$. Since $||\tilde{\pi} - \pi_0|| = o_p(1)$, each term in the last summation is $o_p(1)$, and thus

$$\begin{aligned}
&\frac{1}{m^{1/2}} \sum_{i=1}^{m} \{S_{\text{eff}}(X_i, \beta_0, \tilde{\pi}) - S_{\text{eff}}(X_i, \beta_0, \pi_0)\} \\
&= m^{1/2} E\{S_{\text{eff}}(X_i, \beta_0, \tilde{\pi}) - S_{\text{eff}}(X_i, \beta_0, \pi_0)\} + o_p(1).
\end{aligned}$$

If we define $\eta(t) = t\tilde{\pi} + (1-t)\pi_0$, it is easily checked that, by replacing π_0 by η in the density in (3), we obtain a parametric submodel of the original semiparametric model, with $\eta(0) = \pi_0$ and $\eta(1) = \tilde{\pi}$. Using a Taylor expansion around $t = 0$, we obtain

$$\begin{aligned}
&m^{1/2} E\{S_{\text{eff}}(X_i, \beta_0, \tilde{\pi}) - S_{\text{eff}}(X_i, \beta_0, \pi_0)\} \\
&= m^{1/2} E\left[\frac{\partial S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}}{\partial t}\Big|_{t=0} (1-0) + \frac{1}{2} \frac{\partial^2 S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}}{\partial t^2}\Big|_{t=t*} (1-0)^2 \right]. \qquad \text{(A4)}
\end{aligned}$$

The first term in the above display is zero because of the orthogonality between S_{eff} and the nuisance tangent space:

$$E\left[\frac{\partial S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}}{\partial t}\Big|_{t=0}\right] = E[S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}S_t\big|_{t=0}] = 0,$$

where S_t represents the partial derivative of the log likelihood with respect to t. Since $d\eta(t)/dt = \tilde{\pi} - \pi_0$, the second term, $\partial^2 S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}/\partial t^2$, is generally of the same order as $(\tilde{\pi} - \pi_0)^2 = o_p(n^{-1/2})$, provided that the form of S_{eff} allows us to calculate $\partial S_{\text{eff}}(X_i, \beta_0, \eta)/\partial\eta$. This then leads to the conclusion that (A4) is of order $o_p(1)$, which is needed for the rest of the proof.

In our special setting, it is not obvious that the operation $\partial S_{\text{eff}}(X_i, \beta_0, \eta)/\partial\eta$ is meaningful, and hence we investigate $\partial^2 S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}/\partial t^2$ itself. The form of S_{eff} shows that, as a function of t, $S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}$ is in fact linear; the second component of S_{eff} is even independent of the nuisance parameter. Hence, $\partial^2 S_{\text{eff}}\{X_i, \beta_0, \eta(t)\}/\partial t^2 \equiv 0$, and we automatically obtain the desired $o_p(1)$ rate of (A4).

What we have so far is that $\sqrt{m}(\hat{\beta} - \beta_0) \to N(0, V)$, where V is the inverse of the efficient Fisher information. Since $m = n - n^{\epsilon}$, we have

$$\begin{aligned}&\sqrt{n}(\hat{\beta} - \beta_0) - \sqrt{m}(\hat{\beta} - \beta_0) = (\hat{\beta} - \beta_0)(n - m)/(\sqrt{n} + \sqrt{m})\\ &= \sqrt{m}(\hat{\beta} - \beta_0)(n - m)/\{\sqrt{(nm)} + m\},\end{aligned}$$

which converges in distribution to 0, and hence $\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, V)$ in distribution.

□

REFERENCES

ARNOLD, B. C. & BEAVER, R. J. (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. *Scand. J. Statist.* **12**, 171–8.

BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353–60.

EGUCHI, S. & COPAS, J. (1998). A class of local likelihood methods and near-parametric asymptotics. *J. R. Statist. Soc.* B **60**, 709–24.

GENTON, M. G. (2004). *Skew-Elliptical Distributions and their Applications: A Journey Beyond Normality*. Boca Raton, FL: Chapman and Hall/CRC.

GENTON, M. G. & LOPERFIDO, N. (2005). Generalized skew-elliptical distributions and their quadratic forms. *Ann. Inst. Statist. Math.* **57**, 389–401.

HJORT, N. L. & JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24**, 1619–47.

LOADER, C. R. (1996). Local likelihood density estimation. *Ann. Statist.* **24**, 1602–18.

MA, Y. & GENTON, M. G. (2004). A flexible class of skew-symmetric distributions. *Scand. J. Statist.* **31**, 459–68.

MA, Y., GENTON, M. G. & TSIATIS, A. A. (2005). Locally efficient semiparametric estimators for generalized skew-elliptical distributions. *J. Am. Statist. Assoc.* **100**, 980–89.

MURPHY, S. A. & VAN DER VAART, A. W. (2000). On profile likelihood. *J. Am. Statist. Assoc.* **95**, 449–65.

NEWEY, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Economet.* **5**, 99–135.

RAO, C. R. (1985). Weighted distributions arising out of methods of ascertainment: what populations does a sample represent? In *A Celebration of Statistics: The ISI Centenary Volume*, Ed. A. C. Atkinson & S. E. Fienberg, pp. 543–69, New York: Springer-Verlag.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

SEVERINI, T. A. & WONG, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768–802.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

WANG, J., BOYER, J. & GENTON, M. G. (2004). A skew-symmetric representation of multivariate distributions. *Statist. Sinica* **14**, 1259–70.

[*Received June 2005. Revised August 2006*]