Nonparametric density estimation using partially rank-ordered set samples with application in estimating the distribution of wheat yield

Sahar Nazari

Department of Statistics, Shiraz University, Iran

Mohammad Jafari Jozani[*]

University of Manitoba, Department of Statistics
Winnipeg, MB, CANADA, R3T 2N2
e-mail: m_jafari_jozani@umanitoba.ca

and

Mahmood Kharrati-Kopaei

Department of Statistics, Shiraz University, Iran

Abstract: We study nonparametric estimation of an unknown density function f based on the ranked-based observations obtained from a partially rank-ordered set (PROS) sampling design. PROS sampling design has many applications in environmental, ecological and medical studies where the exact measurement of the variable of interest is costly but a small number of sampling units can be ordered with respect to the variable of interest by any means other than actual measurements and this can be done at low cost. PROS observations involve independent order statistics which are not identically distributed and most of the commonly used nonparametric techniques are not directly applicable to them. We first develop a kernel density estimator of f based on an imperfect PROS sampling procedure and study its theoretical properties. Then, we consider the problem when the underlying distribution is assumed to be symmetric and introduce some plug-in kernel density estimators of f. We use an EM type algorithm to estimate misplacement probabilities associated with an imperfect PROS design. Finally, we expand on various numerical illustrations of our results via several simulation studies and a case study to estimate the distribution of wheat yield using the total acreage of land which is planted in wheat as an easily obtained auxiliary information. Our results show that the PROS density estimate performs better than its SRS and RSS counterparts.

Keywords and phrases: Imperfect subsetting, kernel function, mean integrated square error, nonparametric procedure, optimal bandwidth, ranked set sampling.

Received January 2014.

---

[*]Corresponding author. Tel: +1 204 272 1563; fax: +1 204 474 7621.

1. Introduction

Nonparametric density estimation techniques are widely used to construct an estimate of a density function and to provide valuable information about several features of the underlying population (e.g., skewness, multimodality, etc.) without imposing any parametric assumptions. These methods are very popular in practice and their advantages over histograms or parametric techniques are greatly appreciated. For example, they are widely used for both descriptive and analytic purposes in economics to examine the distribution of income, wages and poverty (e.g., [19]); in environmental and ecological studies (e.g., [8]); or in medical research (e.g., [23]), among others.

Most of these density estimation techniques are based on simple random sampling (SRS) design which involve independent and identically distributed (i.i.d.) samples from the underlying population. There are only a few results available when the sampling design is different (e.g., [5, 7, 13, 18, 2, 3, 4] and [20]). The properties of nonparametric kernel density estimation based on i.i.d. samples are well known and extensively studied in the literature (e.g., [25] as well as [24]). In many applications, however, the data sets are often generated using more complex sampling designs and they do not meet the i.i.d. assumption. Examples include the rank-based sampling techniques which are typically used when a small number of sampling units can be ordered fairly accurately with respect to a variable of interest without actual measurements on them and this can also be done at low cost. This is a useful property since, quite often, exact measurements of these units can be very tedious and/or expensive. For example, for environmental risks such as radiation (soil contamination and disease clusters) or pollution (water contamination and root disease of crops), exact measurements require substantial scientific processing of materials and a high cost as a result, while the variable of interest from a small number of experimental (sampling) units may easily be ranked. These rank-based sampling designs provide a collection of techniques to obtain more representative samples from the underlying population with the help of the available auxiliary information. The samples obtained from such rank-based sampling designs often involve independent observations based on order statistics which are not identically distributed. So, it is important to develop kernel density estimators of the underlying population using such data sets and study their optimal properties.

In this paper, we study the problem of kernel density estimation based on a partially rank-ordered set (PROS) sampling design. [22] introduced PROS sampling procedure as a generalization of the ranked set sampling (RSS) design. To obtain a ranked set sample of size n one can proceed as follows. A set of n units is drawn from the underlying population. The units are ranked via some mechanism rather than the actual measurements of the variable of interest. Then, only the unit ranked as the smallest is selected for full measurement. Another set of n units is drawn and ranked and only the unit ranked as the second smallest is selected for full measurement. This process is repeated n times until the unit ranked the maximum is selected for the final measurement. See [6, 27, 26] and references therein for more details.

In RSS, rankers are forced to assign unique ranks to each observation even if they are not sure about the ranks. PROS design is aimed at reducing the impact of ranking error and the burden on rankers by not requiring them to provide a full ranking of all units in a set. Under PROS sampling technique, rankers have more flexibility by being able to divide the sampling units into subsets of pre-specified sizes based on their partial ranks. [22] obtained unbiased estimators for the population mean and variance using PROS samples. He also showed that PROS sampling has some advantages over RSS. [15] and [16] showed that the Fisher information of PROS samples is larger than the Fisher information of RSS samples of the same size. Since 2011, PROS sampling design has been the subject of many studies. Among others, see [11] for a two-sample distribution-free inference; [21] for quantile estimation; [10] for nonparametric estimation of the population mean; [1] for parametric inference in a location-scale family of distributions, and [14] for finite mixture model analysis based on PROS samples with a fishery application.

The outline of this paper is as follows. In Section 2, we introduce some pre-liminary results and present a general theory that can be used to obtain non-parametric estimates of some functionals of the underlying distribution based on imperfect PROS samples. In Section 3, we present a nonparametric kernel density estimate of the underlying distribution based on an imperfect PROS sampling scheme and study its properties. We also consider the problem of density estimation when the distribution of population is symmetric. In Section 4, we consider the problem of estimating the misplacement probabilities. To this end, we propose a modified EM-algorithm to estimate the probabilities of subsetting errors. This algorithm is fairly simple to implement and simulation results show that its performance is satisfactory. In Section 5, we compare our PROS density estimate with its RSS and SRS counterparts using simulation studies. In Section 6, we illustrate our proposed method by a real example. Finally, in Section 7, we provide some concluding remarks and directions for future works.

2. Necessary backgrounds and preliminary results

In this section, we first give a formal introduction to PROS sampling design and then present some notations and preliminary results. Also, a general theory is obtained to provide nonparametric estimates of some functionals of the underlying distribution based on PROS samples.

2.1. PROS sampling design

To obtain a PROS sample of size $N = nL$, we choose a set size s and a design parameter $D = \{d_1, \ldots, d_n\}$ that partitions the set $\{1, \ldots, s\}$ into n mutually exclusive subsets, where $d_j = \{(j-1)m+1, \ldots, jm\}$ and $m = s/n$. First s units are randomly selected from the underlying population and they are assigned into subsets $d_j, j = 1, \ldots, n$, without actual measurement of the variable of interest and only based on visual inspection or judgment, etc. These subsets

TABLE 1
An example of PROS design

cycle	set	Subsets	Observation
1	S_1	$D_1 = \{\boldsymbol{d_1}, d_2\} = \{\boldsymbol{\{1,2\}}, \{3,4\}\}$	$X_{[d_1]1}$
	S_2	$D_2 = \{d_1, \boldsymbol{d_2}\} = \{\{1,2\}, \boldsymbol{\{3,4\}}\}$	$X_{[d_2]1}$
2	S_1	$D_1 = \{\boldsymbol{d_1}, d_2\} = \{\boldsymbol{\{1,2\}}, \{3,4\}\}$	$X_{[d_1]2}$
	S_2	$D_2 = \{d_1, \boldsymbol{d_2}\} = \{\{1,2\}, \boldsymbol{\{3,4\}}\}$	$X_{[d_2]2}$

are partially judgment ordered, i.e. all units in subset d_j judged to have smaller ranks than all units in $d_{j'}$, where $j < j'$. Then a unit is selected at random for measurement from the subset d_1 and it is denoted by $X_{[d_1]1}$. Selecting another s units assigning them into subsets, a unit is randomly drawn from subset d_2 and then it is quantified and denoted by $X_{[d_2]1}$. This process is repeated until we randomly draw a unit form d_n resulting in $X_{[d_n]1}$. This constitutes one *cycle* of PROS sampling technique. The cycle is then repeated L times to generate a PROS sample of size $N = nL$, i.e. $\mathbf{X}_{\mathrm{PROS}} = \{X_{[d_j]i}; j = 1, \ldots, n; i = 1, \ldots, L\}$. Table 1 shows the construction of a PROS sample with $s = 4, n = 2, L = 2$ and the design parameter $D = \{d_1, d_2\} = \{\{1,2\}, \{3,4\}\}$. Each set includes four units assigned into two partially ordered subsets such that units in d_1 have smaller ranks than units in d_2. In this subsetting process we do not assign any ranks to units within each subset so that they are equally likely to take any place in the subset. One unit, in each set from the bold faced subset, is randomly drawn and is quantified. The fully measured units are denoted by $X_{[d_j]i}$, $j = 1, 2$; $i = 1, 2$.

Note that if all units in d_j have actually smaller ranks than all units in $d_{j'}, j < j'$, then there is no subsetting error and the PROS sample is perfect. Otherwise, we have subsetting error and this PROS sample is called imperfect. To model an imperfect PROS sampling design, following [1, 15] and [16], let $\boldsymbol{\alpha}$ be a double stochastic misplacement probability matrix,

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{d_1,d_1} & \cdots & \alpha_{d_1,d_n} \\ \vdots & \ddots & \vdots \\ \alpha_{d_n,d_1} & \cdots & \alpha_{d_n,d_n} \end{bmatrix}, \tag{1}$$

where α_{d_j,d_h} is the misplacement probability of a unit from subset d_h into subset d_j with $\sum_{h=1}^{n} \alpha_{d_j,d_h} = \sum_{j=1}^{n} \alpha_{d_j,d_h} = 1$. Throughout the paper, we use $\mathrm{PROS}_{\boldsymbol{\alpha}}(n, L, s, D)$ to denote an imperfect PROS sampling design with subsetting error probability matrix $\boldsymbol{\alpha}$, the number of subsets n, the number of cycles L, the set size s, and the design parameter $D = \{d_j, j = 1, \ldots, n\}$ where $d_j = \{(j-1)m+1, \ldots, jm\}$, in which $m = s/n$ is the number of unranked observations in each subset. We note that SRS and imperfect RSS (with ranking error probability matrix $\boldsymbol{\alpha}$) can be expressed as special cases of the $\mathrm{PROS}_{\boldsymbol{\alpha}}(n, L, s, D)$ design when $s = 1$ and $s = n$, respectively. For a perfect PROS design, since $\alpha_{d_j,d_j} = 1$ for $j = 1, \ldots, n$ and $\alpha_{d_j,d_h} = 0$ for $h \neq j$, we use $\mathrm{PROS}_{\mathbf{I}}(n, L, s, D)$, where $\mathbf{I}$ is the identity matrix.

2.2. Some notations and preliminary results

In what follows, the probability density function (pdf) and cumulative distribution function (cdf) of the variable of interest are denoted by f and F, respectively. The pdf and cdf of $X_{[d_j]i}$, for $i = 1, \ldots, L$, are denoted by $f_{[d_j]}$ and $F_{[d_j]}$, respectively, and the pdf of the r-th order statistic from a SRS of size s is denoted by $f_{(r:s)}$. We also use $i_k(g)$ to denote $\int x^k g(x)dx$ and work with a second-order kernel density function $K(\cdot)$ that is symmetric and satisfies the following conditions

$$i_0(K) = \int K(x)dx = 1, \quad i_0(K^2) = \int K^2(x)dx < \infty,$$

and

$$i_2(K) = \int x^2 K(x)dx < \infty.$$

The SRS, RSS and PROS density estimates of f are denoted by $\hat{f}_{\text{SRS}}$, $\hat{f}_{\text{RSS}}$ and $\hat{f}_{\text{PROS}}$, respectively. Now, we present a useful lemma to show the connection between $f_{[d_j]}$ and f.

Lemma 1. *Let* $\mathbf{X}_{PROS}$ *denote a* $PROS_\alpha(n, L, s, D)$ *sample of size N from a population with pdf f and cdf F, respectively. Then*

$$f_{[d_j]}(x) = nf(x) \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} \binom{s-1}{u-1} F(x)^{u-1}\overline{F}(x)^{s-u}$$

$$= \frac{1}{m} \sum_{h=1}^{n} \sum_{u \in d_h} \alpha_{d_j,d_h} f_{(u:s)}(x), \tag{2}$$

where $\overline{F}(x) = 1 - F(x)$, *and consequently*

$$f(x) = \frac{1}{n} \sum_{j=1}^{n} f_{[d_j]}(x) \quad and \quad F(x) = \frac{1}{n} \sum_{j=1}^{n} F_{[d_j]}(x).$$

Remark 1. For a $PROS_{\mathbf{I}}(n, L, s, D)$ design, we have

$$f_{[d_j]}(x) \;=\; nf(x) \sum_{u \in d_j} \binom{s-1}{u-1} F(x)^{u-1}\overline{F}(x)^{s-u}$$

$$=\; \frac{1}{m} \sum_{u \in d_j} f_{(u:s)}(x) = \frac{1}{m} \sum_{r=(j-1)m+1}^{jm} f_{(r:s)}(x).$$

Therefore, a perfect PROS sample is a special case of an imperfect PROS sample and hence the results for a perfect PROS sampling can be obtained as a special case.

Let $h(x)$ be a function of x with $\mu_h = \mathbb{E}[h(X)]$. We study the method of moments estimate of μ_h by using an imperfect PROS sampling procedure, assuming that the required moments of $h(X)$ exist. Note that different choices of $h(x)$ lead to different types of estimators. For example, $h(x) = x^l$ for $l = 1, 2, \ldots$, corresponds to the estimation of population moments; $h(x) = \frac{1}{\lambda} K(\frac{t-x}{\lambda})$ where $K(\cdot)$ is a kernel function and λ is a given constant, corresponds to the kernel estimate of pdf and $h(x) = I(x \leq c)$, where $I(A)$ is the indicator function of A, corresponds to the estimate of cdf at point c.

The method of moments estimate of μ_h based on an imperfect PROS sample of size N is given by

$$\hat{\mu}_{h.\text{PROS}} = \frac{1}{N} \sum_{i=1}^{L} \sum_{j=1}^{n} h(X_{[d_j]i}). \tag{3}$$

The properties of $\hat{\mu}_{h.\text{PROS}}$ is discussed in the following theorem.

Theorem 1. *Let* $\mathbf{X}_{PROS}$ *be a* $PROS_{\boldsymbol{\alpha}}(n, L, s, D)$ *sample of size* N *from a population with pdf* f, *and let the method of moments estimator of* μ_h *be defined as in* (3). *Then*

(i) $\hat{\mu}_{h.PROS}$ *is an unbiased estimator of* μ_h, *i.e.* $\mathbb{E}(\hat{\mu}_{h.PROS}) = \mu_h$.
(ii) $var(\hat{\mu}_{h.PROS}) \leq var(\hat{\mu}_{h.SRS})$ *where* $\hat{\mu}_{h.SRS}$ *is the method of moments estimator of* μ_h *based on a SRS sample of comparable size.*
(iii) $\hat{\mu}_{h.PROS}$ *is asymptotically distributed as a normal distribution with mean* μ_h *and variance* $var(\hat{\mu}_{h.PROS})$ *as* $L \to \infty$.
(iv) $\hat{\mu}_{h.PROS}$ *is a strong consistent estimator of* $\mu_{h.PROS}$ *as* $L \to \infty$.

Proof. The proof is essentially the same as the one given by [22] for $h(x) = x$ which we present here for the sake of completeness. Part (i) is an immediate consequence of Lemma 1. For part (ii), using [22], we have

$$
\begin{aligned}
var(\hat{\mu}_{h.PROS}) &= var(\hat{\mu}_{h.SRS}) - \frac{1}{n^2 L} \sum_{j=1}^{n} (\mu_{h[d_j]} - \mu_h)^2 \\
&\leq var(\hat{\mu}_{h.SRS}),
\end{aligned}
$$

where $\mu_{h[d_j]} = \mathbb{E}[h(X_{[d_j]})]$. Note that the equality holds if and only if $\mu_{h[d_j]} = \mu_h$ for all $j = 1, \ldots, n$; i.e. the subsetting process is purely random. For part (iii), note that $\hat{\mu}_{h.\text{PROS}} = \frac{1}{n} \sum_{j=1}^{n} \overline{h}[d_j]$, where $\overline{h}[d_j] = \frac{1}{L} \sum_{i=1}^{L} h(X_{[d_j]i})$. However, for a fixed j, $\overline{h}[d_j]$ converges asymptotically to a normal distribution with mean $\mathbb{E}[h(X_{[d_j]})]$ and variance $var[h(X_{[d_j]})]/L$ as $L \to \infty$ by the Central Limit Theorem. Therefore, the result holds for $\hat{\mu}_{h.PROS}$. Finally, part (iv) follows from the Strong Law of Large Numbers. □

3. Kernel density estimate based on PROS samples

In this section, we present a kernel density estimator of $f(x)$ based on an imperfect PROS sample of size N and study some theoretical properties of our

proposed estimator. Also, we consider the problem of density estimation when $f(\cdot)$ is assumed to be symmetric.

3.1. Main results

To obtain a PROS kernel density estimator of $f(\cdot)$ we first note that from Lemma 1 we have $f(x) = \frac{1}{n} \sum_{j=1}^{n} f_{[d_j]}(x)$. For a fixed j, the sub-sample $X_{[d_j]i}, i = 1, \ldots, L$ can be considered as a simple random sample of size L from $f_{[d_j]}(\cdot)$. Hence, $f_{[d_j]}(x)$ can be estimated by the usual kernel method as follows

$$\hat{f}_{[d_j]}(x) = \frac{1}{Lh} \sum_{i=1}^{L} K\left(\frac{x - X_{[d_j]i}}{h}\right), \tag{4}$$

where h is the bandwidth to be determined. We propose a kernel estimate of $f(x)$ as

$$\hat{f}_{\mathrm{PROS}}(x) = \frac{1}{n} \sum_{j=1}^{n} \hat{f}_{[d_j]}(x) = \frac{1}{nLh} \sum_{i=1}^{L} \sum_{j=1}^{n} K\left(\frac{x - X_{[d_j]i}}{h}\right). \tag{5}$$

Now, we establish some theoretical properties of $\hat{f}_{\mathrm{PROS}}(x)$. To this end, let $\hat{f}_{\mathrm{SRS}}(x) = \frac{1}{Nh} \sum_{j=1}^{N} K(\frac{x - X_j}{h})$ be a kernel density estimator of $f(x)$ based on a SRS of size N and note that (e.g., [24])

$$\mathbb{E}[\hat{f}_{\mathrm{SRS}}(x)] = f(x) + O(h^2),$$

and

$$var(\hat{f}_{\mathrm{SRS}}(x)) = \frac{1}{Nh} f(x) i_0(K^2) - \frac{1}{N} f^2(x) + O(\frac{h^2}{N}). \tag{6}$$

Theorem 2. *Suppose that $\hat{f}_{PROS}(x)$ is a kernel density estimator of $f(x)$ based on a $PROS_{\boldsymbol{\alpha}}(n, L, s, D)$ sample of size $N = nL$ and let $\hat{f}_{SRS}(x)$ denote its corresponding SRS kernel estimator based on a simple random sample of the same size. Then,*

(i) $\mathbb{E}[\hat{f}_{PROS}(x)] = \mathbb{E}[\hat{f}_{SRS}(x)]$,
(ii) $var(\hat{f}_{PROS}(x)) = var(\hat{f}_{SRS}(x)) - \frac{1}{Nn} \sum_{j=1}^{n} (\mu_{K[d_j]} - \mu_K)^2$, *where*

$$\mu_{K[d_j]} = \mathbb{E}\left[\frac{1}{h} K\left(\frac{x - X_{[d_j]}}{h}\right)\right] \ \ and \ \ \mu_K = \mathbb{E}\left[\frac{1}{h} K\left(\frac{x - X}{h}\right)\right],$$

in which $X_{[d_j]}$ is an observation obtained from a $PROS_{\boldsymbol{\alpha}}(n, L, s, D)$ design.
(iii) $\hat{f}_{PROS}(x)$ *at a fixed point x is distributed asymptotically as a normal distribution with mean $\mathbb{E}[\hat{f}_{PROS}(x)]$ and variance $var(\hat{f}_{PROS}(x))$ for large L.*

Proof. The results hold immediately from Theorem 1 by letting $h(t) = \frac{1}{h} K(\frac{x-t}{h})$. $\square$

Theorem 2 shows that $\hat{f}_{\text{PROS}}(x)$ has the same expectation as $\hat{f}_{\text{SRS}}(x)$ and a smaller variance than $\hat{f}_{\text{SRS}}(x)$. This implies that $\hat{f}_{\text{PROS}}(x)$ has a smaller mean integrated square error (MISE) than $\hat{f}_{\text{SRS}}(x)$, that is

$$
\begin{aligned}
\text{MISE}(\hat{f}_{\text{PROS}}) &= \int \mathbb{E}\left(\hat{f}_{\text{PROS}}(x) - f(x)\right)^2 dx \\
&\leq \int \mathbb{E}\left(\hat{f}_{\text{SRS}}(x) - f(x)\right)^2 dx \\
&= \text{MISE}(\hat{f}_{\text{SRS}}).
\end{aligned}
$$

In addition, by using part (iii) of Theorem 2, one can construct an asymptotic pointwise $100(1 - \nu)\%$ confidence interval for $f(x)$ as follows

$$
\hat{f}_{\text{PROS}}(x) \pm z_{\nu/2}\sqrt{\widehat{var}_{\text{PROS}}(x)},
$$

where $z_{\nu/2}$ is the $100(1 - \frac{\nu}{2})$-th quantile of the standard normal distribution and

$$
\widehat{var}_{\text{PROS}}(x) = \frac{1}{Nh}\hat{f}_{\text{PROS}}(x)i_0(K^2) - \frac{1}{Nn}\sum_{j=1}^{n}\hat{f}_{[d_j]}^2(x),
$$

where $\hat{f}_{[d_j]}(x)$ is a consistent estimators of $f_{[d_j]}(x)$ given by (4). If the lower bound of the proposed confidence interval is negative, one can truncate it at zero.

Note that our estimate $\hat{f}_{\text{PROS}}(x)$ depends on a bandwidth h which should be determined in practice. We present an asymptotic optimal bandwidth by minimizing the asymptotic expansion of $\text{MISE}(\hat{f}_{\text{PROS}})$. In this regard, we first present a lemma which is useful for obtaining the asymptotic expansion of $\text{MISE}(\hat{f}_{\text{PROS}})$.

Lemma 2. *Assuming that the underlying density $f(\cdot)$ is sufficiently smooth with desired derivatives and $K(\cdot)$ is a second-order kernel function, for a fixed n, as $h \to 0$, we have*

$$
\mathbb{E}^2\left[\frac{1}{h}K\left(\frac{x - X}{h}\right)\right] - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}^2\left[\frac{1}{h}K\left(\frac{x - X_{[d_j]}}{h}\right)\right] = f^2(x) - \frac{1}{n}\sum_{j=1}^{n}f_{[d_j]}^2(x) + O(h^2).
$$

Proof. Using (2) and by changing the variable $v = \frac{x-t}{h}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{h}K\left(\frac{x - X_{[d_j]}}{h}\right)\right] &= \int \frac{1}{h}K\left(\frac{x - t}{h}\right)f_{[d_j]}(t)dt \\
&= \frac{1}{m}\sum_{l=1}^{n}\sum_{u \in d_l}\alpha_{d_j,d_l}\int K(v)f_{(u:s)}(hv + x)dv.
\end{aligned}
$$

Replacing $f_{(u:s)}(hv + x)$ by its Taylor expansion

$$
f_{(u:s)}(x) + hvf'_{(u:s)}(x) + (hv)^2 f''_{(u:s)}(x) + O(h^2),
$$

and using the properties of the kernel function $K(\cdot)$, one can easily get

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{h}K\left(\frac{x - X_{[d_j]}}{h}\right)\right] &= \frac{1}{m}\sum_{l=1}^{n}\sum_{u\in d_l}\alpha_{d_j,d_l}\left(f_{(u:s)}(x) + O(h^2)\right) \\
&= f_{[d_j]}(x) + O(h^2).
\end{aligned}
$$

Consequently,

$$
\mathbb{E}^2\left[\frac{1}{h}K\left(\frac{x - X_{[d_j]}}{h}\right)\right] = f_{[d_j]}^2(x) + O(h^2),
$$

and it is similarly verified that

$$
\mathbb{E}^2\left[\frac{1}{h}K\left(\frac{x - X}{h}\right)\right] = f^2(x) + O(h^2),
$$

which completes the proof. □

Theorem 3. *Suppose that the same bandwidth is used in both $\hat{f}_{SRS}$ and $\hat{f}_{PROS}$. Then, for large N,*

$$
MISE(\hat{f}_{PROS}) = MISE(\hat{f}_{SRS}) - \frac{1}{N}\Delta(f,n) + O\left(\frac{h^2}{N}\right),
$$

where $\Delta(f,n) = \int[\frac{1}{n}\sum_{j=1}^{n}f_{[d_j]}^2(x) - f^2(x)]dx$.

Proof. Note that $bias(\hat{f}_{PROS}(x)) = \mathbb{E}[\hat{f}_{PROS}(x) - f(x)] = bias(\hat{f}_{SRS}(x))$. Therefore,

$$
\begin{aligned}
\text{MISE}(\hat{f}_{PROS}) &= \int\left[var(\hat{f}_{PROS}(x)) + bias^2(\hat{f}_{PROS}(x))\right]dx \\
&= \int\left[var(\hat{f}_{PROS}(x)) + bias^2(\hat{f}_{SRS}(x))\right]dx.
\end{aligned}
$$

Now, from Lemma 2

$$
var(\hat{f}_{PROS}(x)) = var(\hat{f}_{SRS}(x)) - \frac{1}{N}\left[\frac{1}{n}\sum_{j=1}^{n}f_{[d_j]}^2(x) - f^2(x)\right] + O\left(\frac{h^2}{N}\right). \quad (7)
$$

Therefore, the result holds. □

Theorem 3 shows that the optimal bandwidth which minimizes $\text{MISE}(\hat{f}_{SRS})$, asymptotically minimizes $\text{MISE}(\hat{f}_{PROS})$ up to order $O(N^{-1})$. That is, one can use the following optimal bandwidth which is obtained by minimizing asymptotical expansion of $\text{MISE}(\hat{f}_{SRS})$

$$
h_{opt.SRS} = i_2(K)^{-2/5}\left[\frac{i_0(K^2)}{i_0(f''^2)}\right]^{1/5}N^{-1/5},
$$

(see [7]). The optimal bandwidth $h_{opt.SRS}$ depends on $f(\cdot)$ which is unknown and, in practice, a nonparametric version of it can be used (see [24]). Theorem 3 also shows that the PROS estimate reduces the MISE of SRS estimate at order $O(N^{-1})$, and the amount of this reduction asymptotically is $\frac{1}{N}\Delta(f,n)$ (note that $\Delta(f,n)$ is non-negative). Unfortunately, the value of $\Delta(f,n)$ depends on $f(\cdot)$; however, as we show below, one can characterize asymptotic rate of this reduction in a perfect PROS sampling procedure as an upper bound for $\Delta(f,n)$.

Lemma 3. *Under a $PROS_{\mathbf{I}}(n,L,s,D)$ sampling design,*

$$\frac{1}{n}\sum_{j=1}^{n}f_{[d_j]}^2(x) = nf^2(x)\,\mathbb{P}\{Y = Z\},$$

where Y and Z are i.i.d. binomial random variables with parameters $s-1$ and $F(x)$, i.e. $Y, Z \sim B(s-1, F(x))$.

Proof. Using Remark 1, one can easily verify that

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}f_{[d_j]}^2(x) &= nf^2(x)\sum_{j=1}^{n}\left[\sum_{r=(j-1)m+1}^{jm}\binom{s-1}{r-1}F(x)^{r-1}\overline{F}(x)^{s-r}\right]^2 \\
&= nf^2(x)\sum_{j=1}^{n}\mathbb{P}^2\{(j-1)m \le Y \le jm-1\},
\end{aligned}
$$

where $Y \sim B(s-1, F(x))$. Let $A_j = \{(j-1)m, \dots, jm-1\}$ for $j = 1, \dots, n$. Suppose Z is also distributed as a $B(s-1, F(x))$ distribution and it is independent of Y, then

$$
\begin{aligned}
\sum_{j=1}^{n}\mathbb{P}^2\{(j-1)m \le Y \le jm-1\} &= \sum_{j=1}^{n}\mathbb{P}\{Y = Z, Z \in A_j\} \\
&= \mathbb{P}\{Y = Z, \bigcup_{j=1}^{n}(Z \in A_j)\} \\
&= \mathbb{P}\{Y = Z\},
\end{aligned}
$$

since A_js constitute a disjoint partition of the set $\{0, \dots, mn-1\}$ and this completes the proof. $\qquad\square$

By Lemma 3, we can derive an asymptotic result which provides more insight into the rate of reduction in MISE in a perfect PROS sampling procedure.

Theorem 4. *Under a $PROS_{\mathbf{I}}(n,L,s,D)$ sampling design, we have*

$$MISE(\hat{f}_{PROS}) = MISE(\hat{f}_{SRS}) - \frac{1}{N}\left[\sqrt{\frac{n}{m}}\delta(f^2) - i_0(f^2)\right] - o(\frac{1}{Nm}) + O(\frac{h^2}{N}),$$

where $\delta(f^2) = \int \frac{f^2(x)}{\sqrt{4\pi F(x)(1-F(x))}}dx$.

Proof. Note that by the Edgeworth expansion of $\mathbb{P}\{Y = Z\}$ in Lemma 3, we get

$$\mathbb{P}\{Y = Z\} = \frac{1}{\sqrt{4s\pi F(x)(1 - F(x))}} + o(\frac{1}{s}).$$

Consequently, we can write

$$\frac{1}{N}\Delta(f, n) = \frac{1}{N}\left[\sqrt{\frac{n}{m}}\delta(f^2) - i_0(f^2)\right] - o(\frac{1}{Nm}),$$

and this completes the proof. □

Theorem 4 shows that a perfect PROS density estimate reduces the MISE of $\hat{f}_{\text{SRS}}$ at order $O(N^{-1})$ and this reduction is increased by $\sqrt{n/m}$ linearly whenever $\sqrt{\frac{n}{m}}\delta(f^2) - i_0(f^2)$ is non-negative (a sufficient condition is $n \geq m\pi$). When $m = 1$, the result is reduced to the result for perfect RSS density estimate given by [7]. In Section 5, we compare $\hat{f}_{\text{PROS}}$ with $\hat{f}_{\text{SRS}}$ and $\hat{f}_{\text{RSS}}$ in a more general case where the sampling procedure can be either perfect or imperfect.

3.2. Density estimation under symmetry assumption

In this section, we consider the problem of kernel density estimation based on an imperfect PROS sample of size $N = nL$ under the assumption that $f(\cdot)$ is symmetric. To this end, suppose that $f(x)$ is symmetric about μ, that is $f(x) = f(2\mu - x)$ for all x. One can easily verify that $f_{[d_j]}(x) = f_{[d_{n-j+1}]}(2\mu - x)$ provided $\alpha_{d_j,d_h} = \alpha_{d_{n-j+1},d_{n-h+1}}$ for all $j, h = 1, \ldots, n$. Therefore, based on the sub-sample $X_{[d_j]i}, i = 1, \ldots, L$, it is reasonable to estimate $f_{[d_j]}(x)$ by

$$\hat{f}^*_{[d_j]}(x, \mu) = \frac{1}{2}\left(\hat{f}_{[d_j]}(x) + \hat{f}_{[d_{n-j+1}]}(2\mu - x)\right),$$

where $\hat{f}_{[d_j]}(x)$ is given in (4). Consequently, the estimate of $f(x)$ under the symmetry assumption can be defined by

$$\begin{aligned} \hat{f}^*_{\text{PROS}}(x, \mu) &= \frac{1}{n}\sum_{j=1}^{n}\hat{f}^*_{[d_j]}(x, \mu) \\ &= \frac{1}{2}\left(\hat{f}_{\text{PROS}}(x) + \hat{f}_{\text{PROS}}(2\mu - x)\right). \end{aligned}$$

Now, we consider the mean and the variance of $\hat{f}^*_{\text{PROS}}(x, \mu)$ in the following theorem.

Theorem 5. *Suppose that $f(x)$ is symmetric about μ and $\alpha_{d_j,d_h} = \alpha_{d_{n-j+1},d_{n-h+1}}$ for all $j, h = 1, \ldots, n$. Then, based on an imperfect $PROS_\alpha(n, L, s, D)$ sample of size $N = nL$, we have*

(i) $\mathbb{E}[\hat{f}^*_{PROS}(x, \mu)] = \mathbb{E}[\hat{f}_{PROS}(x)]$,
(ii) $var(\hat{f}^*_{PROS}(x, \mu)) \leq var(\hat{f}_{PROS}(x))$.

Proof. Part (i) is easily proved by the fact that $\mathbb{E}[\hat{f}_{\text{PROS}}(x)] = \mathbb{E}[\hat{f}_{\text{PROS}}(2\mu - x)]$ under the symmetry assumption and Theorem 2. For part (ii), note that for all x, we have

$$var(\hat{f}_{\text{PROS}}(x)) = var(\hat{f}_{\text{PROS}}(2\mu - x)),$$

and consequently

$$var\left(\hat{f}_{\text{PROS}}^*(x, \mu)\right) = \frac{1}{2}var\left(\hat{f}_{\text{PROS}}(x)\right) + \frac{1}{2}cov\left(\hat{f}_{\text{PROS}}(x), \hat{f}_{\text{PROS}}(2\mu - x)\right).$$

The result holds by using the Cauchy-Schwartz inequality. □

Theorem 5 shows that $\hat{f}_{\text{PROS}}^*(x, \mu)$ has the same bias as $\hat{f}_{\text{PROS}}(x)$; however, it has smaller variance. Therefore, under the symmetry assumption $\hat{f}_{\text{PROS}}^*(x, \mu)$ has smaller MISE and it dominates $\hat{f}_{\text{PROS}}(x)$. Note also that if the symmetry point μ is unknown, it can be estimated by the PROS sample to obtain a plug-in estimator as $\hat{f}_{\text{PROS}}^*(x, \hat{\mu})$. Based on a PROS sample of size N, several non-parametric estimators of μ can be defined as follows

$$
\begin{aligned}
\hat{\mu}_1 &= \frac{1}{N}\sum_{i=1}^{L}\sum_{j=1}^{n}X_{[d_j]i}, \\
\hat{\mu}_2 &= median\{X_{[d_j]i}, \quad i = 1, \ldots, L; j = 1, \ldots, n\}, \\
\hat{\mu}_3 &= median\left\{\frac{X_{[d_j]i} + X_{[d_k]l}}{2}, \quad i, l = 1, \ldots, L; j, k = 1, \ldots, n\right\}, \\
\hat{\mu}_4 &= \frac{1}{L}\sum_{i=1}^{L}median\{X_{[d_j]i}, \quad j = 1 \ldots, n\}.
\end{aligned}
\tag{8}
$$

Among these estimators $\hat{\mu}_1$ is the PROS sample mean which is not robust against outliers, while $\hat{\mu}_2$, $\hat{\mu}_3$, and $\hat{\mu}_4$ are robust estimators of μ. Note that $\hat{\mu}_3$ is a Hodges-Lehmann type estimator of the location parameter. In Section 5, we consider the effect of these estimators on the MISE of $\hat{f}_{\text{PROS}}^*(x, \hat{\mu})$.

4. Estimating the misplacement probabilities

So far we assumed that the misplacement probability matrix $\boldsymbol{\alpha}$ defined in (1) is given. In practice, the misplacement probabilities α_{d_j, d_h} are unknown and they should always be estimated. This is a very important problem as the performance of our kernel density estimator depends on the estimated values of α_{d_j, d_h}. In this section, we use a modification of the EM algorithm of [1] to estimate α_{d_j, d_h}'s. We present the result for a symmetric misplacement probability matrix $\boldsymbol{\alpha}$ with $\alpha_{d_j, d_h} = \alpha_{d_h, d_j}$. However, results for more general $\boldsymbol{\alpha}$ can be obtained by slight modifications of our results. Let

$$\pi_{[d_j, d_h]i} = \frac{\alpha_{d_j, d_h}\bar{\beta}_h(F(X_{[d_j]i}))}{\sum_{h=1}^{n}\alpha_{d_j, d_h}\bar{\beta}_h(F(X_{[d_j]i}))},$$

in which

$$\bar{\beta}_h(F(X_{[d_j]i})) = \frac{1}{m} \sum_{u \in d_h} \beta_{u,s-u+1}(F(X_{[d_j]i})),$$

and $\beta_{a,b}(\cdot)$ denotes the pdf of a beta distribution with parameters a and b. Following [1] we estimate the misplacement probability matrix $\boldsymbol{\alpha}$ through an iterative method. To this end, we start with an initial estimate of $\boldsymbol{\alpha}$ say $\boldsymbol{\alpha}^{(0)}$ which can be chosen to be a matrix associated with random subsetting with $\alpha_{d_j,d_h} = \frac{1}{n}$. Then, we use the following iterative method:

(i) For a given $\boldsymbol{\alpha}^{(t)}$ at step t of the iterative process, calculate

$$w_{h',h}^{(t)} = \sum_{i=1}^{L} \pi_{[d_{h'},d_h]i}^{(t)}.$$

(ii) Calculate

$$Q^{(t)}(\boldsymbol{\alpha}) = \sum_{h=1}^{n} \sum_{h'=1}^{n} w_{h',h}^{(t)} \log(\alpha_{d_{h'},d_h}).$$

(iii) Maximize $Q^{(t)}(\boldsymbol{\alpha})$ under the restrictions that the misplacement probabilities are symmetric and doubly stochastic and obtain the new $\boldsymbol{\alpha}$ and call it $\boldsymbol{\alpha}^{(t+1)}$. This can be done via a Lagrange multipliers method to enforce the constraints as follows

$$\mathcal{L}^{(t)}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_{h=1}^{n} \left\{ \sum_{h'=1}^{h-1} w_{h,h'}^{(t)} \log(\alpha_{d_{h'},d_h}) + \sum_{h'=h}^{n} w_{h,h'}^{(t)} \log(\alpha_{d_h,d_{h'}}) \right\}$$
$$+ \sum_{h=1}^{n} n\lambda_h \left\{ \sum_{h'=1}^{h-1} \alpha_{d_{h'},d_h} + \sum_{h'=h}^{n} \alpha_{d_h,d_{h'}} - 1 \right\},$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. The details of this process are given in [21] as well as [1].

(iv) Repeat Steps (i)–(iii) till the sum of absolute error (SAE) of $\boldsymbol{\alpha}^{(t)}$ and $\boldsymbol{\alpha}^{(t+1)}$ is less than a predetermined value, say δ, that is

$$\mathrm{SAE}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\alpha}^{(t+1)}) = \sum_{i=1}^{n(n+1)/2} |\alpha_i^{(t)} - \alpha_i^{(t+1)}| \leq \delta.$$

In practice, to calculate $\pi_{[d_j,d_h]i}^{(t)}$, one can replace $F(\cdot)$ by an estimate of F such as the empirical distribution function, i.e.

$$\hat{F}_{\mathrm{PROS}}(x) = \frac{1}{nL} \sum_{i=1}^{L} \sum_{j=1}^{n} I(X_{[d_j]i} \leq x).$$

To investigate the accuracy of our method, we perform a small simulation study when $n = m = 3$, and $L = 4, 10$. Following [1], we consider three misplacement

TABLE 2
The averages and standard deviations (in parentheses) of estimated misplacement
probabilities when $n = m = 3$, $L = 4, 10$ and the underlying distributions are the standard
Normal and Exponential distributions

Distributions	$\boldsymbol{\alpha}$	L	$\hat{\alpha}_{d_1,d_1}$	$\hat{\alpha}_{d_1,d_2}$	$\hat{\alpha}_{d_1,d_3}$	$\hat{\alpha}_{d_2,d_2}$	$\hat{\alpha}_{d_2,d_3}$
	$\boldsymbol{\alpha}_1$	4	0.9640(0.099)	0.0355(0.097)	0.0006(0.005)	0.9179(0.155)	0.0466(0.109)
		10	0.9775(0.051)	0.0204(0.051)	0.0021(0.012)	0.9565(0.070)	0.0232(0.054)
	$\boldsymbol{\alpha}_2$	4	0.8934(0.156)	0.0717(0.140)	0.0348(0.087)	0.8247(0.198)	0.1036(0.143)
Normal		10	0.8961(0.106)	0.0759(0.096)	0.0280(0.044)	0.8356(0.153)	0.0886(0.102)
	$\boldsymbol{\alpha}_3$	4	0.7578(0.207)	0.1535(0.188)	0.0888(0.116)	0.6583(0.274)	0.1882(0.217)
		10	0.7365(0.140)	0.1605(0.137)	0.1030(0.082)	0.6902(0.174)	0.1493(0.125)
	$\boldsymbol{\alpha}_1$	4	0.9384(0.134)	0.0612(0.134)	0.0004(0.004)	0.8751(0.185)	0.0637(0.139)
		10	0.9730(0.052)	0.0265(0.052)	0.0005(0.005)	0.9501(0.068)	0.0234(0.048)
	$\boldsymbol{\alpha}_2$	4	0.8918(0.167)	0.0833(0.164)	0.0249(0.059)	0.7880(0.225)	0.1287(0.158)
Exponential		10	0.8802(0.107)	0.0891(0.104)	0.0307(0.042)	0.8014(0.161)	0.1095(0.112)
	$\boldsymbol{\alpha}_3$	4	0.7486(0.232)	0.1475(0.187)	0.1039(0.136)	0.6535(0.254)	0.1990(0.204)
		10	0.7515(0.132)	0.1513(0.125)	0.0972(0.090)	0.6893(0.173)	0.1595(0.128)

probability matrices $\boldsymbol{\alpha}_1$, $\boldsymbol{\alpha}_2$, and $\boldsymbol{\alpha}_3$, where

$$\boldsymbol{\alpha}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \boldsymbol{\alpha}_2 = \begin{bmatrix} 0.900 & 0.075 & 0.025 \\ 0.075 & 0.850 & 0.075 \\ 0.025 & 0.075 & 0.900 \end{bmatrix},$$

and

$$\boldsymbol{\alpha}_3 = \begin{bmatrix} 0.75 & 0.15 & 0.10 \\ 0.15 & 0.70 & 0.15 \\ 0.10 & 0.15 & 0.75 \end{bmatrix}.$$

We generate PROS samples when the underlying population distributions are the standard Normal and Exponential distributions. For each distribution, the misplacement probabilities are estimated by using our proposed iterative method with the help of the package "Rsolnp" ([12] and [28]) in R with $\delta = 10^{-4}$. This process is repeated 100 times and the average of these estimates are used as the estimates of the misplacement probabilities. The values of the estimates and their corresponding standard deviations (given in parentheses) are shown in Table 2. Note that following the properties of $\boldsymbol{\alpha}$ we present the results for α_{d_1,d_1}, α_{d_1,d_2}, α_{d_1,d_3}, α_{d_2,d_2}, and α_{d_2,d_3}. We observe that the estimates are close to the true values and they have satisfactory biases given the fact that our proposed method is a fully nonparametric procedure and the sample size is very small. We observe that our proposed procedure slightly underestimates α_{d_j,d_j}, especially for $\boldsymbol{\alpha}_1$. This is because the perfect ranking model is at the boundary of the parameter space and as noted by [1] the estimates are truncated whenever they exceed 1 due to the constraints on misplacement probabilities. However, the biases and standard deviations get smaller as the cycle size increases.

Remark 2. We note that one can easily modify the Lagrangian multipliers method described above to consider a more general case dealing with asymmetric misplacement probabilities. Under the symmetry assumption the estimates are

more reliable since the number of estimated parameters is less than the case with asymmetric misplacement probabilities. Results of a numerical study (not presented here) show that one can get a better performance for the PROS density estimate under asymmetric misplacement probabilities especially for the tails of the distribution. However, for other regions, PROS estimator under the symmetry assumption might perform better.

5.   Simulation study

In this section, we compare the performance of $\hat{f}_{\text{PROS}}$ with its SRS and RSS counterparts. We first discuss the asymptotic reduction rate in the variance (RRV) of $\hat{f}_{\text{SRS}}$ and $\hat{f}_{\text{RSS}}$ by using a PROS density estimate $\hat{f}_{\text{PROS}}$. We then compare the $\text{MISE}(\hat{f}_{\text{PROS}})$ with $\text{MISE}(\hat{f}_{\text{SRS}})$ and $\text{MISE}(\hat{f}_{\text{RSS}})$. Finally, we consider the effect of estimating the symmetry point on the MISE of $\hat{f}_{\text{PROS}}^{*}(x, \hat{\mu})$ when we assume that the underlying distribution is symmetric.

5.1.   Comparing the reduction in variances

Using (7), the RRV of $\hat{f}_{\text{PROS}}$ over $\hat{f}_{\text{SRS}}$ that measures at what rate $\hat{f}_{\text{PROS}}$ reduces the asymptotic variance of $\hat{f}_{\text{SRS}}$ can be defined as

$$
\text{RRV}(\hat{f}_{\text{PROS}}, \hat{f}_{\text{SRS}}) = \frac{\frac{1}{n}\sum_{j=1}^{n} f_{[d_j]}^2(x) - f^2(x)}{\frac{1}{n}\sum_{j=1}^{n} f_{[d_j]}^2(x)}
$$
$$
= 1 - \left[n\sum_{j=1}^{n} \left\{ \sum_{r=1}^{n} \sum_{u\in d_r} \alpha_{d_j,d_r} \binom{s-1}{u-1} p^{u-1}(1-p)^{s-u} \right\}^2 \right]^{-1},
$$

where $p = F(x)$. It is clear that $\text{RRV}(\hat{f}_{\text{PROS}}, \hat{f}_{\text{SRS}})$ is a nonparametric measure which does not depend on the underlying distribution function. Note that if $\text{RRV}(\hat{f}_{\text{PROS}}, \hat{f}_{\text{SRS}}) = 0$ at certain percentiles p, then $\hat{f}_{\text{PROS}}$ and $\hat{f}_{\text{SRS}}$ have equal variances at these percentiles asymptotically. However, if $\text{RRV}(\hat{f}_{\text{PROS}}, \hat{f}_{\text{SRS}}) = \beta > 0$, then $\hat{f}_{\text{PROS}}$ reduces the variance of $\hat{f}_{\text{SRS}}$ at order $O(N^{-1})$ and this reduction increases linearly at rate β. The values of $\text{RRV}(\hat{f}_{\text{PROS}}, \hat{f}_{\text{SRS}})$ can be easily calculated when n, m and the misplacement probabilities α_{d_i,d_j} are given. For $m = 3$ and $n = 2, \ldots, 7$ and misplacement probabilities $\alpha_{d_i,d_i} = \alpha_0$ and $\alpha_{d_i,d_j} = (1-\alpha_0)/(n-1)$ for $i \neq j$, the values of $\text{RRV}(\hat{f}_{\text{PROS}}, \hat{f}_{\text{SRS}})$ are presented in Figure 1 when $\alpha_0 = 0, 0.3, 0.7, 1$.

We observe that for all values of α_0 the amount of RRV increases symmetrically as p gets away from 0.5 to 0 and 1. This shows that the best performance of the PROS density estimate over its SRS counterpart happens at the tail of the distribution. When $n = 2$, the PROS and SRS estimates have equal precision at $p = 0.5$; otherwise, the PROS estimate reduces the variance of SRS estimate. When $\alpha_0 = 0$, the value of RRV decreases when n increases. This suggests using a small sample size when the misplacement probabilities (ranking errors) are large. We also note that RRV increases as both α_0 and n increase. The best
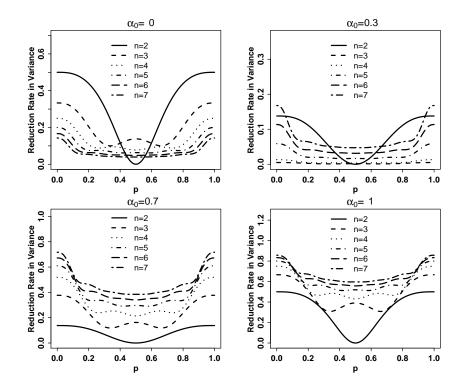
FIG 1. $RRV(\hat{f}_{PROS}, \hat{f}_{SRS})$ for $n = 2, \ldots, 7$ when $m = 3$ and $\alpha_0 = 0, 0.3, 0.7, 1$.

performance of PROS design over SRS design happens when the subsetting is either perfect or it is moderately good, that is when $\alpha_0 = 1$ or $\alpha_0 = 0.7$, respectively. Similar results are observed when $m = 5$ which we do not present here.

To obtain the RRV of $\hat{f}_{\text{PROS}}$ over $\hat{f}_{\text{RSS}}$ we first note that (see [7])

$$var(\hat{f}_{\text{RSS}}(x)) = var(\hat{f}_{\text{SRS}}(x)) - \frac{1}{N}\left[\frac{1}{n}\sum_{r=1}^{n} f_{[r]}^2(x) - f^2(x)\right] + O(\frac{h^2}{N}).$$

Now, using (7), the RRV of $\hat{f}_{\text{PROS}}$ over $\hat{f}_{RSS}$ is defined as

$$\text{RRV}(\hat{f}_{\text{PROS}}, \hat{f}_{RSS}) = \frac{\frac{1}{n}\sum_{j=1}^{n} f_{[d_j]}^2(x) - \frac{1}{n}\sum_{r=1}^{n} f_{[r]}^2(x)}{\frac{1}{n}\sum_{j=1}^{n} f_{[d_j]}^2(x)},$$

where

$$\frac{1}{n}\sum_{r=1}^{n} f_{[r]}^2(x) = nf^2(x)\sum_{r=1}^{n}\left[\sum_{k=1}^{n} p_{rk}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}\right]^2,$$

in which $p = F(x)$ and p_{rk} for $r, k = 1, \ldots, n$ are the ranking error probabilities in an imperfect RSS procedure.
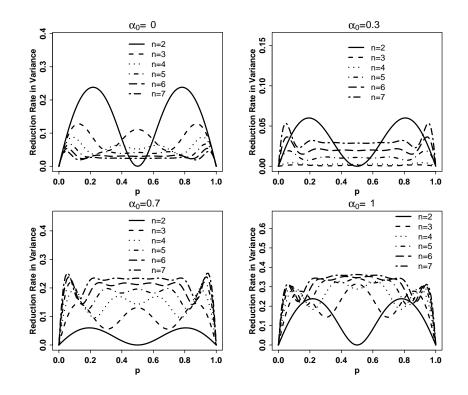
FIG 2. $RRV(\hat{f}_{PROS}, \hat{f}_{RSS})$ for $n = 2, \ldots, 7$ when $m = 3$ and $\alpha_0 = 0, 0.3, 0.7, 1$.

For $m = 3$, $n = 2, \ldots, 7$ and the ranking error probabilities equal to the misplacement error probabilities in its corresponding imperfect PROS design, the values of $RRV(\hat{f}_{PROS}, \hat{f}_{RSS})$ are presented in Figure 2. It is seen that the values of RRV are symmetric about $p = 0.5$. When $\alpha_0 = 0$, the RRV decreases as n increases, and by increasing α_0 RRV of $\hat{f}_{RSS}$ increases as n increases. The RRV of RSS is zero when $p = 0$ and 1 (when $n = 2$, the value of RRV is also zero at $p = 0.5$). This means that the PROS and RSS estimates have the same precision at these percentiles. The maximum value of RRV is more than 35 percent when the sampling procedure is perfect and $n = 7$. Similar results are obtained when $m = 5$ which are not presented here.

5.2. Comparing MISE's of $\hat{f}_{PROS}$, $\hat{f}_{RSS}$ and $\hat{f}_{SRS}$

In order to compare $\mathrm{MISE}(\hat{f}_{PROS})$ with $\mathrm{MISE}(\hat{f}_{RSS})$ and $\mathrm{MISE}(\hat{f}_{SRS})$, following [7], we consider (a) the standard Normal distribution, (b) the Gamma distribution with shape parameter 3 and scale parameter 1, and (c) the standard Gumbel distribution. We use the Epanechnikov kernel in all estimates and the

TABLE 3
The efficiency of PROS density estimate with respect to RSS (RP) and SRS (SP) for different values of n, L, α_0 and Normal, Gamma, and Gumbel distributions when $m = 3$

			α_0				
			0	0.3	0.5	0.7	1
Distributions	n	L	(RP, SP)	(RP, SP)	(RP, SP)	(RP, SP)	(RP, SP)
Normal	6	4	(1.012,1.030)	(1.000,0.993)	(1.022,1.084)	(1.100,1.281)	(1.399,2.151)
	6	8	(1.026,1.058)	(1.015,1.015)	(1.041,1.093)	(1.099,1.268)	(1.309,1.960)
	8	3	(0.997,0.991)	(1.008,1.008)	(1.044,1.128)	(1.113,1.337)	(1.408,2.453)
	8	6	(0.998,1.014)	(1.006,1.018)	(1.042,1.119)	(1.106,1.310)	(1.398,2.265)
Gamma	6	4	(1.021,1.022)	(1.004,1.009)	(1.010,1.069)	(1.060,1.190)	(1.233,1.650)
	6	8	(1.005,1.007)	(1.013,1.014)	(1.033,1.061)	(1.059,1.159)	(1.160,1.486)
	8	3	(1.018,1.054)	(0.992,1.010)	(1.025,1.113)	(1.077,1.278)	(1.224,1.811)
	8	6	(1.009,0.998)	(1.002,0.995)	(1.022,1.063)	(1.071,1.195)	(1.173,1.546)
Gumbel	6	4	(0.985,1.040)	(0.984,1.023)	(1.013,1.095)	(1.102,1.284)	(1.283,1.866)
	6	8	(1.007,0.999)	(1.016,0.990)	(1.058,1.064)	(1.074,1.172)	(1.239,1.619)
	8	3	(1.022,1.002)	(0.992,1.010)	(1.031,1.099)	(1.070,1.265)	(1.269,1.936)
	8	6	(0.996,0.984)	(1.009,1.007)	(1.037,1.075)	(1.092,1.208)	(1.237,1.738)

bandwidth h is determined by

$$h = (4/3)^{1/5} A N^{-1/5},$$

where $A = \min\{$ standard devision of the sample, interquartile range of the sample $/1.34\}$; see [24]. For given n, m, L and different misplacement probabilities, we use the following procedure to estimate the values of $\mathrm{MISE}(\hat{f}_{\mathrm{PROS}})$, $\mathrm{MISE}(\hat{f}_{\mathrm{RSS}})$, and $\mathrm{MISE}(\hat{f}_{\mathrm{SRS}})$. For each estimator, the integrated square error (ISE) $\int(\hat{f}(x) - f(x))^2 dx$ is calculated based on the corresponding SRS, imperfect RSS and imperfect PROS samples. Then, the ISE of 5,000 PROS, RSS, and SRS estimates is obtained. For each procedure, the average of these 5,000 ISEs is used as an estimate of the corresponding MISEs. The ratios

$$\mathrm{RP} = \mathrm{MISE}(\hat{f}_{\mathrm{RSS}})/\mathrm{MISE}(\hat{f}_{\mathrm{PROS}}) \quad \text{and} \quad \mathrm{SP} = \mathrm{MISE}(\hat{f}_{\mathrm{SRS}})/\mathrm{MISE}(\hat{f}_{\mathrm{PROS}})$$

are obtained as the efficiency of $\hat{f}_{\mathrm{PROS}}$ with respect to $\hat{f}_{\mathrm{RSS}}$ and $\hat{f}_{\mathrm{SRS}}$, respectively. Table 3 shows the values of RP and SP for these distributions with different values of n, L, $\alpha_0 = 0, 0.3, 0.5, 0.7, 1$, and $m = 3$.

Form Table 3, it is seen that as the misplacement probabilities decrease the efficiency of PROS with respect to RSS and SRS increases and, as we expect, the efficiency with respect to SRS is more than RSS procedure. When the misplacement probabilities are large, $\alpha_0 < 0.5$, the three estimators have efficiency near one. The efficiency of PROS with respect to RSS and SRS increases slightly as n increases (this increment is faster when $m = 4$, results in which are not presented here); however, the efficiency decreases as L increases. The amount of efficiency for the Normal distribution is higher than the Gamma and Gumbel distributions. For example, when $n = 8$, $L = 3$, and $\alpha_0 = 1$ the efficiencies of PROS with respect to SRS for the Normal, Gamma, and Gumbel distributions are 145%, 81%, and 94%, respectively. We observe that the main parameter that controls the efficiency is the misplacement probability matrix $\boldsymbol{\alpha}$ or equivalently

TABLE 4

The efficiencies of $\hat{f}^*_{PROS}(x, \hat{\mu}_i)$, $i = 1, \ldots, 4$ with respect to $\hat{f}_{PROS}(x)$ and the efficiency of $\hat{f}^*_{PROS}(x, \mu = 0)$ with respect to $\hat{f}_{PROS}(x)$ for $m = 3$, $n = 6, 8$, and $L = 3, 4$ when underlying distributions are the standard Normal, Logistic, Laplace, and t-student with 2 degrees of freedom

| Distributions | n | L | Estimators | | | | |
			$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\mu = 0$
	6	3	1.176	1.062	1.212	1.155	1.364
	6	4	1.185	1.057	1.225	1.165	1.353
Normal	8	3	1.161	1.031	1.205	1.125	1.313
	8	4	1.173	1.054	1.219	1.146	1.309
	6	3	1.103	1.098	1.198	1.172	1.339
	6	4	1.102	1.104	1.207	1.185	1.339
Logistic	8	3	1.108	1.095	1.205	1.159	1.307
	8	4	1.108	1.092	1.205	1.162	1.293
	6	3	0.675	1.150	1.191	1.201	1.299
	6	4	0.645	1.133	1.182	1.190	1.284
t(2)	8	3	0.621	1.133	1.186	1.176	1.268
	8	4	0.615	1.129	1.185	1.176	1.257
	6	3	0.998	1.130	1.151	1.158	1.234
	6	4	1.002	1.119	1.141	1.146	1.217
Laplace	8	3	1.007	1.112	1.139	1.135	1.193
	8	4	1.011	1.108	1.131	1.129	1.172

the ranking error. When the ranking errors are high, there is no substantial difference between $\hat{f}_{\mathrm{PROS}}$, $\hat{f}_{\mathrm{RSS}}$, and $\hat{f}_{\mathrm{SRS}}$. However, as the ranking errors decrease, our simulation results show that the PROS density estimate performs better than RSS and SRS density estimates in terms of MISE.

5.3. Results under symmetry assumption

To investigate the effect of estimating the symmetry point μ on the MISE of $\hat{f}^*_{\mathrm{PROS}}$, we consider four distributions (a) the standard Normal and (b) Logistic distributions as light tail distributions, (c) t-student with 2 degrees of freedom, and (d) the standard Laplace distributions as heavy tail distributions. For each distribution, a perfect PROS sample of size $N = nL$ with subset size m are generated and the four symmetry point estimators given in (8) were calculated. Then, $\mathrm{MISE}(\hat{f}_{\mathrm{PROS}})$ and $\mathrm{MISE}(\hat{f}^*_{\mathrm{PROS}})$ for $\hat{\mu}_i$, $i = 1, \ldots, 4$, were calculated and their ratios are obtained as the efficiency of $\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_i)$'s with respect to $\hat{f}_{\mathrm{PROS}}(x)$. The results for $m = 3$, $n = 6, 8$, and $L = 3, 4$ are shown in Table 4. The last column shows the efficiency of $\hat{f}^*_{\mathrm{PROS}}$ with respect to $\hat{f}_{\mathrm{PROS}}$ when the symmetry point is known.

We observe that $\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_3)$ performs the best in all cases which suggests using the Hodges-Lehmann type estimator, $\hat{\mu}_3$, for estimating the symmetry point. For Normal distribution, the efficiencies of $\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_1)$ and $\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_3)$ with respect to $\hat{f}_{\mathrm{PROS}}(x)$ are competitive. However, for $t(2)$ distribution, which is a heavy tail distribution, it does not hold. On the other hand, the efficiencies of

$\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_4)$ and $\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_3)$ with respect to $\hat{f}_{\mathrm{PROS}}(x)$ are very close especially in heavy tail distributions. Generally, we recommend using $\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_3)$ when it is assumed that the underlying population distribution is symmetric. Comparing the efficiencies of $\hat{f}^*_{\mathrm{PROS}}(x, \mu = 0)$ and $\hat{f}^*_{\mathrm{PROS}}(x, \hat{\mu}_3)$ with respect to $\hat{f}_{\mathrm{PROS}}(x)$ indicates how much the efficiency reduces when the symmetry point is estimated. This reduction is larger when $n = 6$ in comparison with $n = 8$ (the maximum value of reduction is about 11% when $n = 6$ and $L = 3$ in Normal distribution and the minimum value is 3% when $n = 8$ and $L = 4$ in Laplace distribution).

6. Real data application

In this section, we illustrate our method with a real data set collected by the Iranian Ministry of Jihade-Agricultural (IMJA) in 2005. [17] used this data set in a different context to examine the accuracy of several ratio estimators of the population mean based on RSS design. The data set contains the information of the wheat yield and the total acreage of land which is planted in wheat for 304 cities in 31 provinces of Iran in 2005. Wheat yield estimation is important for advanced planning and implementation of policies related to food distribution, import-export decision, etc. We provide kernel density estimates of the distribution of $Y =$ wheat yield (in ton) as the variable of interest by using $X =$ total acreage of the planted land in wheat (in acre) as the auxiliary variable which can be used for the ranking purpose. The correlation coefficient between X and Y is 0.786. For ease of computations, we divided the values of Y by 100,000,000.

In order to estimate the density function of wheat yield, we regarded this data set as a population and extracted PROS, RSS and SRS with replacement samples of size $N = nL$ from the population. For each design, the density estimates are obtained and the asymptotic variance estimates are calculated. Then, this process is repeated M times and the average of density estimates at a fixed point are considered as the density estimates. In addition, for each design, the average of asymptotic variance estimates are also calculated for constructing asymptotic pointwise confidence bounds. We take $(n, m, L, M) = (3, 4, 6, 20)$ and consider a general case where we do not make any symmetry assumption about the misplacement probabilities. The histogram of 304 records of Y is shown in Figure 3. The PROS density estimate and its 95 percent asymptotic pointwise confidence bounds are shown in third column of Figure 3. The SRS and RSS density estimates and their corresponding 95 percent pointwise confidence bounds are also shown in the 1st and 2nd columns of Figure 3. In all cases, we used Epanechnikov kernel and the bandwidth was determined as in Section 5.2. To obtain the probabilities of subsetting errors, we used a simple modification of the proposed algorithm in Section 4 that accounts for asymmetric misplacement probabilities. We estimated the probabilities of subsetting errors for each 20 samples with SAE=0.001. The average of these estimates are given below

$$\hat{\alpha} = \left[\begin{array}{ccc} 0.860(0.14) & 0.140(0.14) & 6.33 \times 10^{-6}(2.80 \times 10^{-5}) \\ 0.104(0.14) & 0.724(0.22) & 0.172(0.12) \\ 0.036(0.08) & 0.136(0.13) & 0.828(0.12) \end{array} \right],$$
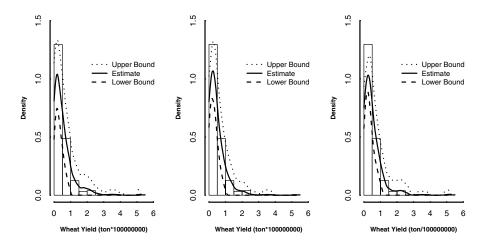
FIG 3. *The histogram of wheat yield (in ton × 100,000,000), the SRS, RSS and PROS kernel density estimates and their corresponding asymptotic 95 percent truncated pointwise confidence bounds.*

where the standard deviations of the estimates are reported in parentheses. Our estimates show that the probabilities of correct subsetting are much higher than the probabilities of incorrect subsetting and this is due to the fact that X and Y are highly correlated. We observe that the density estimates look similar. However, the truncated confidence bounds for PROS design are much narrower than their SRS and RSS counterparts, where for each design the truncated confidence bounds for $f(x)$ based on an estimator $\hat{f}(x)$ is obtained as follows:

$$\left(\max\{0, \hat{f}(x) - z_{0.975}\sqrt{\hat{V}(\hat{f}(x))}\}, \hat{f}(x) + z_{0.975}\sqrt{\hat{V}(\hat{f}(x))} \right),$$

where $\hat{V}(\hat{f}(x))$ is the estimated approximate variance of $\hat{f}(x)$.

7. Concluding remarks

In this paper, we have considered the problem of nonparametric density estimation of the pdf $f(x)$ of a continuous random variable based on partially ordered ranked set samples. First, we developed kernel density estimates of f based on an imperfect PROS sampling design and studied its theoretical properties. Then, we considered the problem when the underlying distribution is assumed to be symmetric. We produced some plug-in kernel density estimates of f under the symmetry assumption. We have shown that, under imperfect PROS design with symmetric misplacement probabilities, PROS estimates of f perform better than their SRS and RSS counterparts. It would naturally be of interest to expand our results to a more general PROS design where design parameters are more flexible and the misplacement probabilities are not necessarily symmetric. For example, one may be interested is studying how well the PROS density es-

timates perform if the symmetry assumption on the misplacement probabilities is not imposed. In such a case, we observe that the bias of $\hat{f}_{PROS}(x)$ does not change as $\mathbb{E}[\hat{f}_{PROS}(x)] = \mathbb{E}[\hat{f}_{SRS}(x)]$. However, the behaviour of the variance of $\hat{f}_{PROS}(x)$ as a function of misplacement probabilities will need further numerical and theoretical investigations. Also, it would be natural to study whether the symmetry assumption results in better estimates for the misplacement probabilities. Another problem of interest is to use other models for subsetting errors in order to improve the PROS density estimates presented in this paper. For example, as mentioned by a referee, one may want to study the possibility of extending the model proposed by [9] for ranking errors in RSS to the subsetting errors in PROS design to improve our PROS density estimates. Another interesting problem which worth further consideration is to develop some diagnostic tools to test the symmetry of misplacement probabilities. One possible way to have a prior information about the values of the misplacement probabilities is to take a preliminary sample (or rely on the results of previous studies) and use a bootstrap technique to estimate these probabilities. Some of these problems are currently under investigation and the results will be presented elsewhere.

Acknowledgments

References

[1] ARSLAN, G. AND OZTURK, O. (2013). *Parametric inference based on partially rank ordered set samples.* Journal of the Indian Statistical Association, **51**, 1–24.

[2] BARABESI, L. AND FATTORINI, L. (2002). *Kernel estimation of probability density functions by ranked set sampling.* Communication in Statistics: Theory and Methods, **31**, 597–610. MR1902313

[3] BREUNIG, R.V. (2001). *Density estimation for clustered data.* Econometric Reviews, **20**, 353–367. MR1863870

[4] BREUNIG, R.V. (2008). *Nonparametric density estimation for stratified samples.* Statistics and Probability Letters, **78**, 2194–2200. MR2528636

[5] BUSKIRK, T.D. (1998). *Nonparametric density estimation using complex survey data.* In ASA Proceedings of the Section on Survey Research Methods, 799–801. American Statistical Association.

[6] CHEN, Z., BAI, Z. AND SINHA, B.K. (2004). *Ranked Set Sampling: Theory and Applications.* Springer-Verlag, New York. MR2151099

[7] CHEN, Z. (1999). *Density estimation using ranked set sampling data.* Environmental and Ecological Statistics, **6**, 135–146.

[8] Fieberg, J. (2007). *Kernel density estimators of home range: Smoothing and the autocorrelation red herring.* Ecology, **88**, 1059–1066.

[9] Fligner, M.A. and MacEachern, S.N. (2006). *Nonparametric two–sample methods for ranked set sample data.* Journal of the American Statistical Association, **101**, 1107–1118. MR2324150

[10] Frey, J. (2012). *Nonparametric mean estimation using partially ordered sets.* Environmental and Ecological Statistics, **6**, 309–326. MR2972686

[11] Gao, J.L. and Ozturk, O. (2012). *Two sample distribution-free inference based on partially rank-ordered set samples.* Statistics and Probability Letters, **82**, 876–884. MR2910033

[12] Ghalanos, A. and Theussl, S. (2012). *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method.* R package version 1.14.

[13] Gulati, S. (2004). *Smooth non-parametric estimation of the distribution function from balanced ranked set samples.* Environmetric, **15**, 529–539.

[14] Hatefi, A., Jafari Jozani, M. and Ozturk, M. (2014). *Mixture model analysis of partially rank ordered set samples: Estimating the age-groups of fish from length-frequency data.* Submitted.

[15] Hatefi, A. and Jafari Jozani, M. (2013a). *Information content of partially rank ordered set samples.* Submitted.

[16] Hatefi, A. and Jafari Jozani, M. (2013b). *Fisher information in different types of perfect and imperfect ranked set samples from finite mixture models.* Journal of Multivariate Analysis, **119**, 16–31. MR3061412

[17] Jafari Jozani, M., Majidi, S. and Perron, F. (2012). *Unbiased and almost unbiased ratio estimators of the population mean in ranked set sampling.* Statistical Papers, **53**, 719–737. MR2958750

[18] Lam, K.F., Yu, P.L.H. and Lee, C.F. (2002). *Kernel method for the estimation of the distribution function and the mean with auxiliary information in ranked set sampling.* Environmetrics, **13**, 397–406.

[19] Minoiu, C. and Reddy, S.G. (2012). *Kernel density estimation on grouped data: The case of poverty assessment.* The Journal of Economic Inequality. To appear.

[20] Opsomer, J.D. and Miller, C.P. (2005). *Selecting the amount of smoothing in nonparametric regression estimation for complex surveys.* Journal of Nonparametric Statistics, **17**, 593–611. MR2141364

[21] Ozturk, O. (2012). *Quantile inference based on partially rank ordered set samples.* Journal of Statistical Planning and Inference, **142**, 2116–2127. MR2903416

[22] Ozturk, O. (2011). *Sampling from partially rank-ordered sets.* Environmental Ecological Statistics, **18**, 757–779. MR2860114

[23] Miladinovic, B., Kumar, A. and Djulbegovic, B. (2013). *Kernel density estimation for random-effects meta-analysis.* International Journal of Mathematical Sciences in Medicine, **1**, 1–5.

[24] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London, New York. MR0848134

[25] WAND, M.P. AND JONES, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London. MR1319818

[26] WOLFE, D.A. (2012). *Rank set sampling: Its relevance and impact on statistical inference.* ISRN Probability and Statistics, doi 10.5402/2012/568385.

[27] WOLFE, D.A. (2004). *Ranked set sampling: An approach to more efficient data collection.* Statistical Science, **19**, 636–643. MR2185585

[28] YE, Y. (1987). *Interior algorithms for linear, quadratic, and linearly constrained non-linear programming.* Ph.D. Thesis, Department of EES, Stanford University.