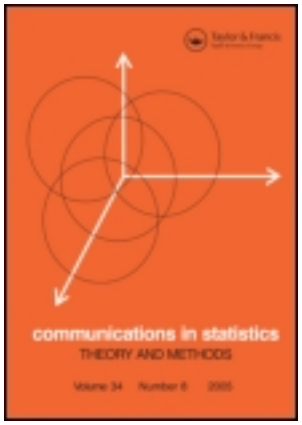
This article was downloaded by: [University of Maryland Baltimore County]
On: 25 April 2014, At: 06:27
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Kernel Density Estimator From Ranked Set Samples

Johan Lim[a], Min Chen[b], Sangun Park[c], Xinlei Wang[d] & Lynne Stokes[d]
[a] Department of Statistics, Seoul National University, Seoul, Korea
[b] Department of Mathematical Sciences, University of Texas at Dallas, Dallas, Texas, USA
[c] Department of Applied Statistics, Yonsei University, Seoul, Korea
[d] Department of Statistical Science, Southern Methodist University, Dallas, Texas, USA
Published online: 23 Apr 2014.

To cite this article: Johan Lim, Min Chen, Sangun Park, Xinlei Wang & Lynne Stokes (2014) Kernel Density Estimator From Ranked Set Samples, Communications in Statistics - Theory and Methods, 43:10-12, 2156-2168

To link to this article: http://dx.doi.org/10.1080/03610926.2013.791372

PLEASE SCROLL DOWN FOR ARTICLE

Taylor & Francis makes every effort to ensure the accuracy of all the information (the "Content") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content.

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden. Terms &

Conditions of access and use can be found at http://www.tandfonline.com/page/terms-and-conditions

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

Communications in Statistics—Theory and Methods, 43: 2156–2168, 2014
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print / 1532-415X online
DOI: 10.1080/03610926.2013.791372

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

Order Statistics and Ordered Data Analysis

Kernel Density Estimator From Ranked Set Samples

JOHAN LIM,[1] MIN CHEN,[2] SANGUN PARK,[3] XINLEI WANG,[4] AND LYNNE STOKES[4]

[1]Department of Statistics, Seoul National University, Seoul, Korea
[2]Department of Mathematical Sciences, University of Texas at Dallas, Dallas, Texas, USA
[3]Department of Applied Statistics, Yonsei University, Seoul, Korea
[4]Department of Statistical Science, Southern Methodist University, Dallas, Texas, USA

We study kernel density estimator from the ranked set samples (RSS). In the kernel density estimator, the selection of the bandwidth gives strong influence on the resulting estimate. In this article, we consider several different choices of the bandwidth and compare their asymptotic mean integrated square errors (MISE). We also propose a plug-in estimator of the bandwidth to minimize the asymptotic MISE. We numerically compare the MISE of the proposed kernel estimator (having the plug-in bandwidth estimator) to its simple random sampling counterpart. We further propose two estimators for a symmetric distribution, and show that they outperform in MISE all other estimators not considering symmetry. We finally apply the methods in this article to analyzing the tree height data from Platt et al. (1988) and Chen et al. (2003).

Keywords Kernel density estimator; Optimal bandwidth; Ranked set sampling.

Mathematics Subject Clssification 62G30; 62G07.

1. Introduction

Ranked set sampling (RSS) is a data collection scheme that usually yields more efficient estimators than simple random sampling. It was originally proposed by McIntyre (1952) for situations where actual measurements of sampling units are difficult or expensive to obtain, but ranking a set of subjects is relatively easy or less costly. Ever since then, RSS has been successfully applied to many fields such as agriculture, forestry, biology, ecology, environmental sciences, and medical studies. We refer readers to Chen et al. (2003) and the references therein for more details of the RSS.

Received August 29, 2012; Accepted March 27, 2013.
Address Correspondences to Johan Lim, Department of Statistics, Seoul National University, Seoul, Korea. E-mail: johanlim@snu.ac.kr
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lsta.

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

The density function or the cumulative distribution function is the first step to understand the stochastic nature of the underlying population. For the RSS, several estimators have been proposed to estimate the cumulative distribution function (cdf) in the literature. Stokes and Sager (1988) estimated the cdf by averaging the empirical cdf of each rank stratum. Kvam and Samaniego (1994) studied the nonparametric maximum likelihood estimator (MLE). Oztúrk (2007) proposed an isotonized estimator and numerically showed that it is robust to ranking error. He further extended the method of Kvam and Samaniego (1994) to the NPMLE with ranking errors (Oztúrk 2009). However, unlike the cdf problem, less effort was made to estimate the density function, although it might be as important as the cumulative distribution to understand the population.

Suppose we observe the RSS data:

$$X_{(1)1}, \ldots, X_{(1)n_1}, X_{(2)1}, \ldots, X_{(2)n_2}, \ldots, X_{(H)1}, \ldots, X_{(H)n_H},$$

where $X_{(h)j}$ is the $h-$th smallest observation in a set of H independent samples in cycle j. We assume the samples are from the distribution with density function $f(x)$, which has a Hölder continuous and square-integrable fourth derivative. Chen (1999) studied a kernel density estimator for the balanced RSS:

$$\widehat{f}_{\mathrm{RSS}}(x;h) = \frac{1}{H}\sum_{r=1}^{H}\frac{1}{mh}\sum_{j=1}^{m}K\left(\frac{x - X_{(r)j}}{h}\right).$$

He showed that, when the optimal bandwidth for the simple random sample (SRS), denoted by h_{SRS}, is used, it has smaller mean integrated square error (MISE) than the kernel density estimator based on the SRS. The MISE of $\widehat{f}(x;h)$ is

$$\mathrm{MISE}(\widehat{f}(h)) \equiv \int (\widehat{f}(x;h) - f(x))^2 dx,$$

and Chen (1999) showed analytically that

$$\mathrm{MISE}(\widehat{f}_{\mathrm{RSS}}(h_{\mathrm{SRS}})) \le \mathrm{MISE}(\widehat{f}_{\mathrm{SRS}}(h_{\mathrm{SRS}})) = \min_h \mathrm{MISE}(\widehat{f}_{\mathrm{SRS}}(h)). \tag{1}$$

In Chen's estimator, the bandwidth $h = h_{\mathrm{SRS}}$ depends on the unknown $f(x)$. Thus, $\widehat{f}_{\mathrm{RSS}}(x;h_{\mathrm{SRS}})$ is not directly computable from the data.

Barabesi and Fattorini (2002) studied the kernel density estimator for the unbalanced RSS, that is

$$\widehat{f}_{\mathrm{RSS}}(x;h) = \frac{1}{H}\sum_{r=1}^{H}\frac{1}{n_r h}\sum_{i=1}^{n_r}\mathrm{K}\left(\frac{x - X_{(r)i}}{h}\right). \tag{2}$$

They computed its MISE and asymptotic MISE, and show that asymptotically the balanced RSS is equally efficient in MISE to the unbalanced RSS with the optimal sample allocation. They also studied the properties of the kernel estimator for the balanced RSS, and considered a few choices of bandwidth, which were originally suggested for the SRS.

In this article, we study the kernel density estimator (2), where the kernel function $K(t)$ is symmetric about 0 and

$$\int K(t)dt = 1, \quad \int t^2 K(t)dt \neq 0, \tag{3}$$

and $|t|^3 K(t) \to 0$ as t increases. Note that the kernel function $K(t)$ is the same as that in Chen (1999). We show that the optimal bandwidth for the balanced RSS is same with that for the SRS, and can be estimated by treating the RSS data as the SRS data. We propose a plug-in estimator of the bandwidth for the unbalanced RSS data.

This article is organized as follows. In Sec. 2, we review the asymptotic MISE of the estimator $\widehat{f}_{\mathrm{RSS}}$ in (2) and find the optimal bandwidth, denoted by h_{RSS}, that minimizes the asymptotic MISE. We show that h_{SRS}, the optimal bandwidth for SRS used in Chen (1999), is also optimal for the balanced RSS, i.e., it minimizes the MISE of (2) when the RSS is balanced. Thus, in (1),

$$\min_h \mathrm{MISE}(\widehat{f}_{\mathrm{RSS}}(h)) = \mathrm{MISE}(\widehat{f}_{\mathrm{RSS}}(h_{\mathrm{SRS}})). \tag{4}$$

The optimal bandwidth h_{RSS} also depends on the unknown density function $f(x)$ and is not directly computable from data. In this section, we propose to estimate h_{SRS}, the optimal bandwidth for SRS, by treating RSS as SRS, and then estimate h_{RSS} using the asymptotic relationship between h_{RSS} and h_{SRS}. We write the proposed estimator of h_{SRS} and h_{RSS} as $h_{\mathrm{p.SRS}}$ and $h_{\mathrm{p.RSS}}$, respectively. In Sec. 3, we consider the case when the distribution is symmetric. The symmetry in the density function allows a better kernel estimator. We introduce two new estimators to take advantage of the symmetry. If the density $f(x)$ is symmetric with respect to μ, *i.e.*, $f(x+\mu) = f(-x+\mu)$, then we have $f_{(r)}(\mu - x) = f_{(H-r+1)}(\mu + x)$ for $r = 1, 2, \ldots, H$. We define new estimators by merging the samples of rank strata r and $H - r + 1$ for $r = 1, 2, \ldots, H$. We numerically show that this estimator outperforms its asymmetric (without symmetry) counterpart. In Sec. 4, we numerically compare the MISEs of the proposed $\widehat{f}_{\mathrm{RSS}}(x; h_{\mathrm{p.RSS}})$ with its SRS counterpart. We also show that the symmetrized kernel estimator has smaller MISE than the asymmetric estimator if the true density $f(x)$ is symmetric. In Sec. 5, we analyze the tree height data from Chen et al. (2003) and show that the unbalanced RSS may provide a better density function estimator than the balanced RSS. In Sec. 6, we conclude the article with a brief summary.

2. Asymptotic MISE and Optimal Bandwidth

The bandwidth selection is an important step in estimating the density using the kernel method. In this section, we derive the optimal bandwidth that minimizes the (MISE) of $\widehat{f}_{\mathrm{RSS}}$. We assume that the population density $f(x)$ is twice differentiable at every x.

2.1. MISE

We first compute the MISE of $\widehat{f}_{\mathrm{RSS}}$. For each rank stratum r, let

$$\widehat{f}_{(r)}(x; h) = \frac{1}{n_r h} \sum_{i=1}^{n_r} \mathrm{K}\left(\frac{x - X_{(r)i}}{h}\right),$$

which is the kernel density estimator of the density function $f_{(r)}(x)$ of the $r-$th stratum. Using the results on the RSS (see Wand and Jones, 1995), for each r, the bias and the variance of $\widehat{f}_{(r)}(x; h)$ are, respectively,

$$\mathrm{bias}(\widehat{f}_{(r)}(x; h)) = \frac{1}{2} i_2(\mathrm{K}) f^{(2)}_{(r)}(x) h^2 + o(h^2) \tag{5}$$

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

and

$$\operatorname{var}(\widehat{f}_{(r)}(x;h)) = \frac{1}{n_r h} i_0(\mathrm{K}^2) f_{(r)}(x) + o\left(\frac{1}{n_r h}\right), \tag{6}$$

where $i_l(g) = \int x^l g(x)dx$. Since

$$\operatorname{bias}(\widehat{f}_{\mathrm{RSS}}(x;h)) = \frac{1}{H}\sum_{r=1}^{H} \operatorname{bias}(\widehat{f}_{(r)}(x;h))$$

and

$$\operatorname{var}(\widehat{f}_{\mathrm{RSS}}(x;h)) = \frac{1}{H^2}\sum_{r=1}^{H} \operatorname{var}(\widehat{f}_{(r)}(x;h)),$$

we have from (5) and (6) that

$$\begin{aligned}
\operatorname{bias}(\widehat{f}_{\mathrm{RSS}}(x;h))^2 &= \frac{1}{4} i_2(\mathrm{K})^2 \left\{\frac{1}{H}\sum_{r=1}^{H} f_{(r)}^{(2)}(x)\right\}^2 h^4 + o(h^4) \\
&= \frac{1}{4} i_2(\mathrm{K})^2 \{f^{(2)}(x)\}^2 h^4 + o(h^4) \\
\operatorname{var}(\widehat{f}_{\mathrm{RSS}}(x)) &= \frac{1}{H^2}\sum_{r=1}^{H} \frac{1}{n_r h} i_0(\mathrm{K}^2) f_{(r)}(x) + o\left(\frac{1}{h \min_r(n_r)}\right).
\end{aligned}$$

Thus, the MISE of $\widehat{f}_{\mathrm{RSS}}(x;h)$ becomes

$$\begin{aligned}
\operatorname{MISE}(\widehat{f}_{\mathrm{RSS}}(x;h)) = {} & \frac{1}{H^2 h}\sum_{r=1}^{H} \frac{1}{n_r} i_0(\mathrm{K}^2) \\
& + \frac{1}{4} i_2(\mathrm{K})^2 i_0\{(f^{(2)})^2\} h^4 + o\left(\max\left(\frac{1}{h \min(n_r)}, h^4\right)\right).
\end{aligned} \tag{7}$$

2.2. *Optimal Bandwidth*

Simple algebra shows that the optimal bandwidth h to minimize the MISE in (7) is

$$h_{\mathrm{RSS}} = \left\{\frac{\sum_{r=1}^{H}(1/(n_r H^2)) i_0(\mathrm{K}^2)}{i_2(\mathrm{K})^2 i_0((f^{(2)})^2)}\right\}^{1/5}. \tag{8}$$

This brings two interesting findings. First, in the balanced RSS with $n_r = m$ for every r, h_{RSS} equals to h_{SRS}, which is the optimal bandwidth for SRS with a size $N = \sum_{r=1}^{H} n_r$:

$$h_{\mathrm{RSS}} = h_{\mathrm{SRS}} = i_2(\mathrm{K})^{-2/5} \left\{\frac{i_0(\mathrm{K}^2)}{i_0((f^{(2)})^2)}\right\}^{1/5} N^{-1/5}. \tag{9}$$

This implies that Chen (1999)'s estimator $\widehat{f}_{\mathrm{RSS}}(h_{\mathrm{SRS}})$ is optimal to minimize the MISE, and

$$\begin{aligned}
\min_h \operatorname{MISE}(\widehat{f}_{\mathrm{RSS}}(h)) &= \operatorname{MISE}(\widehat{f}_{\mathrm{RSS}}(h_{\mathrm{SRS}})) \\
&\le \operatorname{MISE}(\widehat{f}_{\mathrm{SRS}}(h_{\mathrm{SRS}})) = \min_h \operatorname{MISE}(\widehat{f}_{\mathrm{SRS}}(h)).
\end{aligned}$$

Second, in unbalanced RSS, h_{RSS} and h_{SRS} are asymptotically proportional to each other as

$$\frac{h_{\mathrm{RSS}}}{h_{\mathrm{SRS}}} \approx \left\{ \frac{\sum_{r=1}^{H}(1/(n_r H^2))}{1/\sum_{r=1}^{H} n_r} \right\}^{1/5} = \left\{ \frac{N}{H^2} \sum_{r=1}^{H}(1/n_r) \right\}^{1/5}, \tag{10}$$

where the proportionality constant depends only on sample sizes.

2.3. *Estimation of Bandwidth*

The optimal bandwidth h_{RSS} and h_{SRS} are functions of the unknown density function $f(x)$, because in (8),

$$i_0((f^{(2)})^2) = \int \left\{ f^{(2)}(x) \right\}^2 dx = \int f^{(4)}(x) f(x) dx$$

is a function of unknown $f(x)$ and here we aim to estimate this quantity.

For SRS, several procedures have been proposed in the literature. Suppose $X_1, \ldots, X_n$ form a SRS from $f(x)$. Park and Marron (1990) estimated $\widehat{i_0}\left((f^{(2)})^2\right)$ with

$$\frac{1}{n(n-1)} \frac{1}{\alpha^5} \sum_{i \neq j} \mathrm{L}^{(4)}\left(\frac{X_i - X_j}{\alpha} \right) \tag{11}$$

where L is the kernel function satisfying the conditions in (3). Hall and Marron (1987) used the optimal bandwidth α to estimate $i_0((f^{(2)})^2)$ as

$$\alpha_1 = \left\{ \frac{18 \cdot i_0(\mathrm{L}^{(4)} * \mathrm{L})}{i_2(\mathrm{L} * \mathrm{L})} \cdot \frac{i_0(f^2)}{i_0\{(f^{(3)})^2\}} \right\}^{1/13} n^{-2/13} \tag{12}$$

with $\mathrm{L} * \mathrm{L}(x) = \int \mathrm{L}(x-t)\mathrm{L}(t)dt$. The bandwidth α_1 again depends on derivatives of the unknown density $f(x)$. Park and Marron (1990) proposed to estimate α_1 by replacing f in (12) with a density of the scale family; they used the normal distribution with mean 0 and estimated variance. On the other hand, Sheather and Jones (1991) used the estimator

$$\widehat{i_0}\left\{ \left(f^{(2)} \right)^2 \right\} = \frac{1}{n^2} \frac{1}{\alpha^5} \sum_{i,j} \mathrm{L}^{(4)}\left(\frac{X_i - X_j}{\alpha} \right) \tag{13}$$

and

$$\alpha_2 = \left\{ \frac{2\mathrm{L}^{(4)}(0)}{i_2(\mathrm{L}) i_0(f^{(3)})} \right\}^{1/7} n^{-1/7}.$$

They showed that (13) produces a bandwidth estimator which performs better than that based on α_1 in both theory and computation.

In balanced RSS, $i_0((f^{(2)})^2)$ can be estimated using (11) or (13) by treating the RSS as SRS. The kernel estimator for the balanced RSS is

$$\widehat{f}_{\mathrm{RSS}}(x; h) = \sum_{r=1}^{H} \sum_{i=1}^{m} w_{(r)j} \frac{1}{h} \mathrm{L}\left(\frac{x - X_{(r)i}}{h} \right)$$

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

and its fourth derivative is

$$\widehat{f}^{(4)}_{\mathrm{RSS}}(x;h) = \sum_{r=1}^{H}\sum_{i=1}^{m} w_{(r)j}\frac{1}{h^5}\mathrm{L}^{(4)}\left(\frac{x - X_{(r)i}}{h}\right)$$

with $w_{(r)j} = 1/(H \cdot m)$. We then estimate

$$i_0\{(f^{(2)})^2\} = \int f^{(4)}(x)f(x)dx$$

unbiasedly as

$$\sum_{r_1=1}^{H}\sum_{j_1=1}^{m} w_{(r_1)j_1}\left\{\sum_{r_2=1}^{H}\sum_{j_2=1}^{m} w_{(r_2)j_2}\frac{1}{h^5}\mathrm{L}^{(4)}\left(\frac{X_{(r_1)j_1} - X_{(r_2)j_2}}{h}\right)\right\}. \tag{14}$$

The resulting bandwidth equals to the bandwidth of Sheather and Jones (1991) for SRS and, simply, can be evaluated by treating RSS as SRS.

For unbalanced RSS, we propose to estimate h_{SRS} by treating the RSS as a SRS as in balanced RSS. We denote the resulting estimate as $h_{\mathrm{p.SRS}}$. Then, motivated by the identity (10), we estimate h_{RSS} as

$$h_{\mathrm{p.RSS}} = \left\{\frac{N}{H^2}\sum_{r=1}^{H}(1/n_r)\right\}^{1/5} h_{\mathrm{p.SRS}}. \tag{15}$$

In balanced RSS, $h_{\mathrm{p.RSS}}$ equals to $h_{\mathrm{p.SRS}}$. The bandwidth estimators $h_{\mathrm{p.SRS}}$ and $h_{\mathrm{p.RSS}}$ lead to new kernel estimators $\widehat{f}_{\mathrm{RSS}}(x;h_{\mathrm{p.SRS}})$ and $\widehat{f}_{\mathrm{RSS}}(x;h_{\mathrm{p.RSS}})$, respectively.

3. Symmetric Distribution

We show that, when the population distribution is symmetric with respect to μ, $f_{(r)}(\mu - x) = f_{(H-r+1)}(\mu + x)$ for $r = 1, \ldots, H$. Suppose that $X_1, \ldots, X_H$ are identically and independently distributed (IID) from a symmetric distribution with a density function $f(x)$, satisfying $f(\mu - x) = f(\mu + x)$. Let $F(x)$ and $F_{(r)}(x)$ be the population cumulative distribution function (cdf) and the cdf of the rth order statistic, respectively. Since we know

$$F(\mu - x) + F(\mu + x) = 1$$

from the symmetry,

$$\begin{aligned}
F_{(r)}(\mu - x) &= \frac{1}{\mathrm{Beta}(r, H - r + 1)}\int_0^{F(\mu-x)} t^{r-1}(1-t)^{H-r}dt \\
&= \frac{1}{\mathrm{Beta}(r, H - r + 1)}\int_{1-F(\mu-x)}^{1} (1-s)^{r-1}s^{H-r}ds \\
&= 1 - \frac{1}{\mathrm{Beta}(r, H - r + 1)}\int_0^{F(\mu+x)} (1-s)^{r-1}s^{H-r}ds \\
&= 1 - F_{(H-r+1)}(\mu + x).
\end{aligned}$$

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

Thus, we have

$$f_{(r)}(\mu - x) = f_{(H-r+1)}(\mu + x). \tag{16}$$

We now propose a new estimator $\widehat{f}^{\text{sym}}(x)$ for the symmetric density function using (16):

$$\begin{aligned}
\widehat{f}^{\text{sym}}(x) &= \frac{1}{2}\{\widehat{f}(x) + \widehat{f}(2\mu - x)\} \\
&= \frac{1}{H}\sum_{r=1}^{H}\frac{1}{2}\{\widehat{f}_{(r)}(x) + \widehat{f}_{(r)}(2\mu - x)\} \\
&= \frac{1}{H}\sum_{r=1}^{H}\frac{1}{2}\{\widehat{f}_{(r)}(x) + \widehat{f}_{(H-r+1)}(2\mu - x)\}
\end{aligned}$$

In the above, μ is unknown in practice and is replaced with $\widehat{\mu}_{\text{RSS}}$.

We show that the new estimator has smaller MISE than the asymmetric estimator $\widehat{f}$. First, the bias of $\widehat{f}^{\text{sym}}$ is:

$$\begin{aligned}
\text{bias}(\widehat{f}^{\text{sym}}(x)) &= \frac{1}{2}\{\text{bias}(\widehat{f}(x)) + \text{bias}(\widehat{f}(2\mu - x))\} \\
&= \frac{1}{2H}\sum_{r=1}^{H}\left\{(1/2)i_2(K)f_{(r)}^{(2)}(x)h^2 + (1/2)i_2(K)f_{(r)}^{(2)}(2\mu - x)h^2 + o(h^2)\right\} \\
&= \frac{1}{4}i_2(K)h^2\left\{\frac{1}{H}\sum_{r=1}^{H}\left(f_{(r)}^{(2)}(x) + f_{(r)}^{(2)}(2\mu - x)\right)\right\} + o(h^2) \\
&= \frac{1}{4}i_2(K)h^2\left\{\frac{1}{H}\sum_{r=1}^{H}\left(f_{(r)}^{(2)}(x) + f_{(H-r+1)}^{(2)}(2\mu - x)\right)\right\} + o(h^2) \\
&= \frac{1}{2}i_2(K)h^2 f^{(2)}(x) + o(h^2),
\end{aligned}$$

which is equal to that of $\widehat{f}(x)$.

Second, the variance equals to

$$\frac{1}{4}\{\text{var}(\widehat{f}(x)) + \text{var}(\widehat{f}(2\mu - x)) + 2\text{cov}(\widehat{f}(x), \widehat{f}(2\mu - x))\},$$

where

$$\text{cov}(\widehat{f}(x), \widehat{f}(2\mu - x)) \le \{\text{var}(\widehat{f}(x))\}^{1/2}\{\text{var}(\widehat{f}(2\mu - x))\}^{1/2}.$$

In addition, we find that

$$\begin{aligned}
\text{var}(\widehat{f}(2\mu - x)) &= \frac{1}{H^2}\sum_{r=1}^{H}\frac{1}{n_r h}i_0(\mathrm{K}^2)f_{(r)}(2\mu - x) + o\left(\frac{1}{h\,\min_r(n_r)}\right) \\
&= \frac{1}{H^2}\sum_{r=1}^{H}\frac{1}{n_r h}i_0(\mathrm{K}^2)f_{(r)}(x) + o\left(\frac{1}{h\,\min_r(n_r)}\right) \\
&= \text{var}(\widehat{f}(x)) + o\left(\frac{1}{h\,\min_r(n_r)}\right).
\end{aligned} \tag{17}$$

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

Thus, $\text{var}(\widehat{f}^{\text{sym}}(x)) \leq \text{var}(\widehat{f}(x)) + o(1/(h \min_r(n_r)))$. Asymptotically, $\widehat{f}^{\text{sym}}$ has smaller MISE (or variance) than $\widehat{f}$.

We propose another estimator which would be more efficient than $\widehat{f}^{\text{sym}}$. The currently proposed estimator $\widehat{f}^{\text{sym}}(x)$ estimates $f_{(r)}(x)$ by

$$\widehat{f}^{\text{sym}}_{(r)}(x) = \frac{1}{2}\{\widehat{f}_{(r)}(x) + \widehat{f}_{(H-r+1)}(2\mu - x)\} \tag{18}$$

using the fact $f_{(r)}(\mu - x) = f_{(H-r+1)}(\mu + x)$. We could instead consider

$$\tilde{f}^{\text{sym}}_{(r)}(x) = \frac{n_r}{n_r + n_{H-r+1}}\widehat{f}_{(r)}(x) + \frac{n_{H-r+1}}{n_r + n_{H-r+1}}\widehat{f}_{(H-r+1)}(2\mu - x). \tag{19}$$

This estimator can be computed by estimating $f_{(r)}(x)$ using both $X_{(r)i}$ and $2\mu - X_{(H-r+1)j}$ for $i = 1, 2, \ldots, n_r$ and $j = 1, 2, \ldots, n_{H-r+1}$. To be specific, suppose we refer the combined samples as $y_{(r)i}$'s for $i = 1, 2, \ldots, n'_r$ with $n'_r = n_r + n_{H-r+1}$. We estimate $f_{(r)}(x)$ as follows.

- If $H = 2m + 1$, for $r = 1, \ldots, m$,

$$\tilde{f}_{(r)}(x; h) = \frac{1}{n'_r}\sum_{i=1}^{n'_r}\frac{1}{h}K\left(\frac{x - y_{(r)i}}{h}\right),$$

  and for $r = m + 1$,

$$\tilde{f}_{(m+1)}(x; h) = \frac{1}{n_{m+1}}\sum_{i=1}^{n_{m+1}}\frac{1}{h}K\left(\frac{x - y_{(m+1)i}}{h}\right).$$

- If $H = 2m$, for $r = 1, \ldots, m$,

$$\tilde{f}_{(r)}(x; h) = \frac{1}{n'_r}\sum_{i=1}^{n'_r}\frac{1}{h}K\left(\frac{x - y_{(r)i}}{h}\right),$$

  .

- In both cases, for $[H/2] + 1 \geq r$,

$$\tilde{f}_{(r)}(x; h) = \tilde{f}_{(H-r+1)}(x; h).$$

It is worth noting that

$$\tilde{f}_{(r)}(x; h) = \frac{n_r}{n_r + n_{(H-r+1)}}\widehat{f}_{(r)}(x; h) + \frac{n_{(H-r+1)}}{n_r + n_{(H-r+1)}}\widehat{f}_{(H-r+1)}(2\mu - x; h).$$

and it provides the estimator:

$$\tilde{f}^{\text{sym}}(x; h) = \frac{1}{H}\sum_{r=1}^{H}\tilde{f}_{(r)}(x; h).$$

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

The unequally weighted estimator $\tilde{f}^{\text{sym}}(x;h)$ may have smaller MISE than $\widehat{f}^{\text{sym}}(x;h)$, since, for each r, $\tilde{f}_{(r)}(x;h)$ is more efficient (has smaller variance) than $\widehat{f}^{\text{sym}}_{(r)}(x;h)$. We numerically compare their MISEs in next section.

4. Simulation Studies to Compare Relative Efficiency

In this section, we numerically compare the MISEs of the kernel estimators in previous sections with respect to their SRS counterpart.

This study is designed as follows. We use four different distributions as the true underlying distributions: (i) standard normal distribution; (ii) gamma distribution with $\alpha = 1$ and $\beta = 3$; (iii) t-distribution with 3 degrees of freedom; and (iv) uniform distribution on [0, 1]. We consider $H = 2, 3, 4$, and $\mathbf{n}$ are vectors of 4 and 8. We use the grid approximation to compute the MISE, where the grids are from the 5th to 95th percentile of each distribution. In each case, we generate 10,000 data sets and compute the MISE of the estimates.

In the estimation, we use the normal kernel and estimate the bandwidth for SRS using the function "dpik" in the R package "kernsmooth." It implements the procedure by Sheater and Jones (1991). We consider six different kernel estimators in the study:

(i) $\widehat{f}_{\text{SRS}}$: the kernel estimator based on SRS and h_{SRS}. We estimate h_{SRS} using the procedure by Sheather and Jones (1991).
(ii) $\widehat{f}_{\text{oracle}}$: the kernel estimator based on RSS with bandwidth h_{RSS}. We compute h_{SRS} by plugging the true density into (8).
(iii) $\widehat{f}_{\text{p.SRS}}$: the kernel estimator based on RSS with bandwidth $h_{\text{p.SRS}}$.
(iv) $\widehat{f}_{\text{p.RSS}}$: the kernel estimator based on RSS with bandwidth $h_{\text{p.RSS}}$.
(v) $\widehat{f}^{\text{sym}}_{\text{p.RSS}}$: the equally weighted symmetric kernel estimator based on $\widehat{f}_{\text{p.RSS}}$.
(vi) $\tilde{f}^{\text{sym}}_{\text{p.RSS}}$: the unequally weighted symmetric kernel estimator based on $\widehat{f}_{\text{p.RSS}}$.

We compute the relative efficiencies (RE) of (ii)–(vi) against (i).

Table 1 reports the REs of (ii)–(vi) against (i). The RE of $\widehat{f}_1$ against $\widehat{f}_2$ is defined by

$$\text{RE}(\widehat{f}_1, \widehat{f}_2) = \frac{\text{MISE}(\widehat{f}_2, f)}{\text{MISE}(\widehat{f}_1, f)},$$

where, for $i = 1, 2$,

$$\text{MISE}(\widehat{f}_i, f) = \int \{\widehat{f}_i(x) - f(x)\}^2 dx.$$

The REs in the table are the average of MISEs of densities from 10,000 simulated data sets.

Table 1 shows that the proposed $\widehat{f}_{\text{p.RSS}}$ performs better than $\widehat{f}_{\text{SRS}}$ except one case (the case with unbalanced small size sample from the $t-$distribution). Two estimators $\widehat{f}_{\text{p.RSS}}$ and $\widehat{f}_{\text{SRS}}$ are equal to each other in balanced RSS, and are optimal in terms of MISE. In unbalanced RSS, $\widehat{f}_{\text{p.RSS}}$ performs better than $\widehat{f}_{\text{p.SRS}}$ in most of cases. Second, the table shows that the REs of $\widehat{f}_{\text{p.RSS}}$ are higher for the balanced RSS than the unbalanced RSS. For example, the RE of $\widehat{f}_{\text{p.RSS}}$ with $\mathbf{n} = (4, 4, 4)$ performs better than $\widehat{f}_{\text{p.RSS}}$ with $\mathbf{n} = (4, 4, 8)$ in all cases. This would not be surprising if we recall that $\widehat{f}_{\text{p.RSS}}$ is sub-optimal for unbalanced RSS. The third finding we make is that the efficiency-loss from estimating the bandwidth is significant; the REs of $\widehat{f}_{\text{p.RSS}}$ are smaller than $\widehat{f}_{\text{oracle}}$ in all cases. Finally, in the table, the RE tends to decrease as the number of replication increases. Please find the differences between (4, 4) and (8, 8), (4, 4, 4) and (8, 8, 8), and (4, 4, 4, 4) and (8, 8, 8, 8).

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

Table 1
Relative efficiencies of kernel estimators

Dist.	H	**n**	$\widehat{f}_{\text{oracle}}$	$\widehat{f}_{\text{p.SRS}}$	$\widehat{f}_{\text{p.RSS}}$	$\widehat{f}^{\text{sym}}_{\text{p.RSS}}$	$\tilde{f}^{\text{sym}}_{\text{p.RSS}}$
Normal	2	(4,4)	1.330	1.002	1.002	1.374	1.374
	2	(4,8)	1.296	1.100	1.127	1.565	1.609
	2	(8,8)	1.299	1.015	1.015	1.348	1.348
	3	(4,4,4)	1.507	1.247	1.247	1.646	1.646
	3	(4,4,8)	1.431	1.216	1.236	1.673	1.690
	3	(8,8,8)	1.389	1.136	1.136	1.489	1.489
	4	(4,4,4,4)	1.579	1.330	1.330	1.816	1.816
	4	(4,4,8,8)	1.491	1.208	1.233	1.677	1.691
	4	(8,8,8,8)	1.493	1.305	1.305	1.773	1.773
Gamma	2	(4,4)	1.283	1.146	1.146	−−	−−
	2	(4,8)	1.183	1.064	1.058	−−	−−
	2	(8,8)	1.140	1.045	1.045	−−	−−
	3	(4,4,4)	1.182	1.099	1.099	−−	−−
	3	(4,4,8)	1.200	1.071	1.061	−−	−−
	3	(8,8,8)	1.132	1.065	1.065	−−	−−
	4	(4,4,4,4)	1.183	1.134	1.134	−−	−−
	4	(4,4,8,8)	1.162	1.022	1.009	−−	−−
	4	(8,8,8,8)	1.127	1.097	1.097	−−	−−
t	2	(4,4)	1.395	1.193	1.193	1.480	1.480
	2	(4,8)	1.166	0.962	0.983	1.232	1.256
	2	(8,8)	1.294	1.090	1.090	1.255	1.255
	3	(4,4,4)	1.378	1.253	1.253	1.474	1.474
	3	(4,4,8)	1.260	1.151	1.169	1.413	1.415
	3	(8,8,8)	1.380	1.166	1.166	1.365	1.365
	4	(4,4,4,4)	1.560	1.319	1.319	1.556	1.556
	4	(4,4,8,8)	1.338	1.149	1.169	1.379	1.412
	4	(8,8,8,8)	1.403	1.248	1.248	1.472	1.472
Uniform	2	(4,4)	1.409	1.129	1.129	1.398	1.398
	2	(4,8)	1.227	1.010	1.032	1.289	1.319
	2	(8,8)	1.253	1.108	1.108	1.336	1.336
	3	(4,4,4)	1.494	1.331	1.331	1.639	1.639
	3	(4,4,8)	1.309	1.129	1.147	1.438	1.431
	3	(8,8,8)	1.290	1.180	1.180	1.394	1.394
	4	(4,4,4,4)	1.506	1.401	1.401	1.692	1.692
	4	(4,4,8,8)	1.313	1.196	1.213	1.487	1.509
	4	(8,8,8,8)	1.299	1.203	1.203	1.419	1.419

This is simply because the MISE of $\widehat{f}_{\text{RSS}}$ decreases as the number of replication increases, and tells that the RSS is more effective than the SRS when the sample size is small.

The symmetrization notably improves the REs of the estimators. Two symmetrized estimators even outperform $\widehat{f}_{\text{oracle}}$ in several cases. Between the two, the unequally weighted $\tilde{f}^{\text{sym}}_{\text{p.RSS}}$ performs better than the equally weighted $\widehat{f}_{\text{p.RSS}}$.

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

5. An Empirical Study

In this section, we apply the proposed kernel estimators to analyzing the tree height data from Platt et al. (1988) and Chen et al. (2003). We treat the reported tree height data as true population. We apply unbalanced designs along with their balanced counterpart to illustrate the usefulness of the unbalanced designs. The results show that a well-selected unbalanced design can provide a better density estimate than balanced one.

The tree height data has a right-skewed density on the positive line as shown in Fig. 1. In the RSS experiment with $H = 2$, we expect that the observations of $r = 2$ are more volatile (have higher variance) than $r = 1$ (because those are bounded below by 0). Thus, the estimation of $f_{(2)}(x)$ needs more observations than that of $f_{(1)}(x)$ to get a given level of efficiency (or MISE). To have this in mind, we apply three designs with $H = 2$ to estimate the density of tree height; (i) $(n_1, n_2) = (5, 5)$; (ii) $(n_1, n_2) = (3, 7)$; (iii) $(n_1, n_2) = (2, 8)$.

As in the numerical study, we generate 1,000 RSS data sets and estimate their density functions for each of the case. In Fig. 2, we plot the bias, the variance, and the MSE of kernel estimators based on the three designs.

The results show that both unbalanced designs $(n_1, n_2) = (3, 7)$ and $(n_1, n_2) = (2, 8)$ perform better than their balanced counterpart. This would not be surprising if we recall that the underlying density of tree height data is skewed to the right.

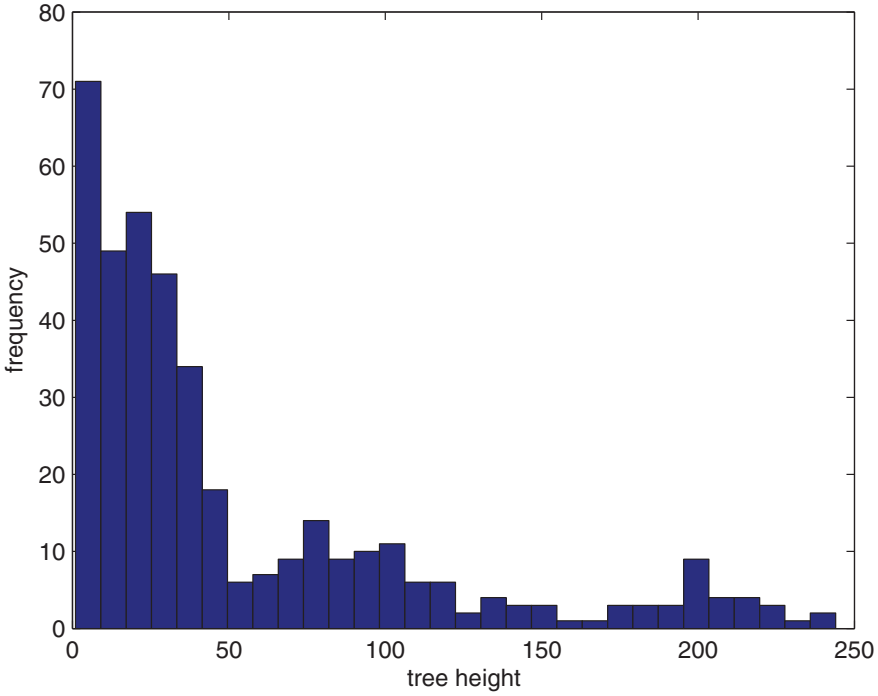


Figure 1. Histogram of tree height data.

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

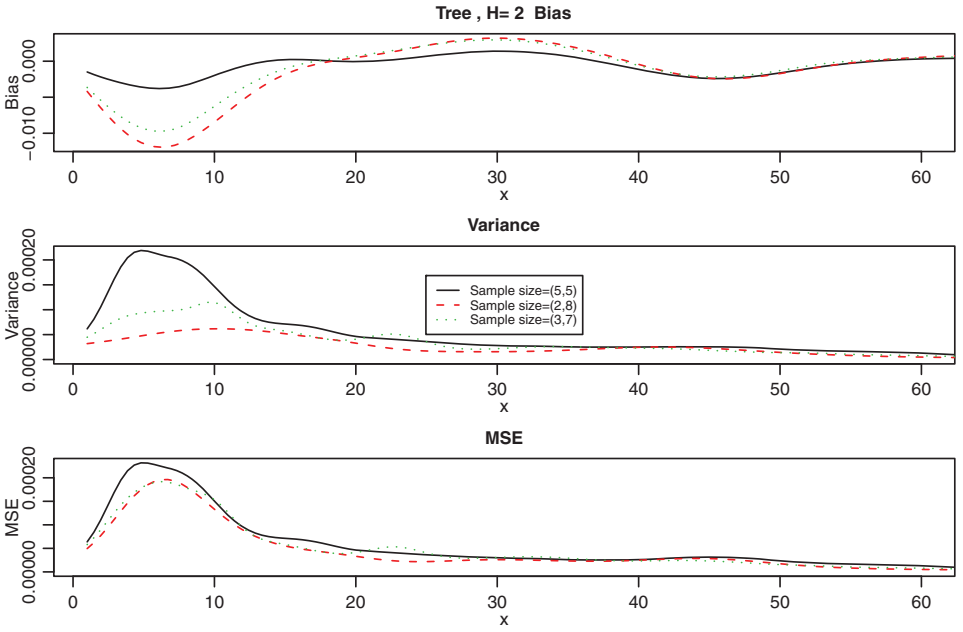


Figure 2. The bias, variance, and MSE of kernel estimators

6. Conclusion

We finish this article with a brief summary. Here, we study the kernel density estimator based on RSS. First, we compute asymptotic MISE for RSS and the asymptotic optimal bandwidth to minimize it. We show that the optimal bandwidth for balanced RSS is equal to that for SRS. Second, we propose a plug-in estimator of the asymptotic optimal bandwidth and the kernel estimator using it. The numerical study shows that the proposed estimator performs better than its SRS counterpart. Third, we propose two estimators for a symmetric distribution, which outperforms their asymmetric counterpart. Finally, the analysis of tree height data illustrates that a well designed unbalanced RSS can provide a better density estimator than the balanced RSS.

Funding

This work was supported in part by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (2008-0061196).

References

Baraneso, L., Fattorini, L. (2002). Kernel estimators for probability density functions by ranked-set sampling. *Communi. Statist. Theor. Meth.* 31(4):597–610.

Chen, Z. (1999). Density estimation using ranked set sampling data. *Enviro. Ecolog. Statist.* 6:135–146.

Chen, Z., Bai, Z., Sinha, B. K. (2003). *Ranked Set Sampling: Theory and Applications*. New York: Springer.

Hall, P., Marron, J. S. (1987). Estimation of integrated squared density derivatives. *Statis. Probab. Lett.* 6:109–115.

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014

Kvam, P. H., Samaniego, F. J. (1994). Nonparametric maximum likelihood estimation based on ranked set samples. *J. Amer. Statist. Assoc*. 89:526–537.

Mode, N., Conquest, L., Marker, D. (1999). Ranked set sampling for ecological research: accounting for the total costs of sampling. *Austral. J. Agricult. Res* 3:385–390.

McIntyre, G. A. (1952). A method for unbiased selective sampling using ranked sets. *Environmetrics* 10:179–194.

Oztúrk, O. (2007). Minimum distance estimator of judgment class distributions in a ranked set sample. *J. Nonparametric Statist*. 19:131–144.

Oztúrk, O. (2009). Nonparametric maximum-likelihood estimation of within-set ranking errors in ranked set sampling. *J. Nonparametric Statist*. 22(7):823–840.

Park, B. U., Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc*. 85:409, 66–72.

Platt, W. J., Evans, G. M., Rathbun, S. L. (1988). The population dynamics of long-lived conifer (Pinus plaustris) (1988). *Amer. Natrualist* 131:491–525.

Sheather, S. J., Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* 53(3):683–690.

Stokes, S. L., Sager, T. W. (1988). Characterization of a ranked set sample with application to estimating distribution functions. *J. Amer. Statist. Associ*. 83:374–381.

Wand, M. P., Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman and Hall.

Downloaded by [University of Maryland Baltimore County] at 06:27 25 April 2014