Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 65(2) 2011 163-167

Kernel Density Estimation for Size-biased Sample under Multiplicative Censoring

Mohammad Abbaszadeh[*] **and Hassan Doosti**

Department of Statistics Ferdowsi, University of Mashhad, Mashhad, Iran

Received 22 Setpember 2010; Revised 24 Setpember 2010; Accepted 24 September 2010

SUMMARY

Here we propose a kernel based density estimator for the weighted univariate data under multiplicative censoring that may be useful in the context of inference for certain agricultural data. Asymptotic formulae for the MSE and MISE of the new estimator are derived and it is shown that the optimal rate of convergence of MISE of the new estimator is slower than that in the case of i.i.d. data as may be expected.

Keywords : Length-biased data, Multiplicative censoring, Kernel density, Size-biased data.

1. INTRODUCTION

Consider a probability space (Ω, $\mathcal{J}$, P) and a random variable (*rv*) X defined on $\Omega \to H$, where $H = (a, b)$ is an interval of the real line. For the continuous case, let $f_X(.)$ denote the probability density function (pdf) of X and $g(.)$ a nonnegative function satisfying $\mu = Eg(X) < \infty$, then the random variable Y with pdf f_Y

$$f_Y(y) = \frac{g(y) f_X(y)}{\mu} \tag{1.1}$$

is said to have size-biased or weighted distribution, corresponding to the distribution of X.

The examples of such distributions can be found in many applied fields including agriculture, ecology and forestry [see Rao (1965), Patil and Ord (1975), Patil and Rao (1977) and Rao (1977) and Patil and Rao (1978)]. The reader may also refer to Ricker (1969) that exemplifies the concerns of not incorporating the weighting function in the context of a fisheries study that may hold in other contexts also. Gupta and Kirmani (1990) have further discussed the role of weighted distributions in stochastic modeling whereas Navarro *et al*. (2001), Nair and Sunoj (2003) and Sunoj (2004) consider characterizing such distributions from various considerations. Recently, Ramirez and Vidakovic (2010) consider wavelet based density estimation and Chaubey *et al.* (2010) focus on nonparametric density estimation using a histogram smoother based on weights generated by an appropriate Poisson distribution.

Here we consider another practical situation, where the observations Y_i may be further damaged according to multiplicative censoring [see Vardi (1989)]. This results in observing $Z_i = U_i Y_i$, $i = 1, 2, ..., n$, where $\{U_1, ..., U_n\}$ is a random sample from the uniform distribution on (0, 1). In the agricultural context Y may represent actual crop production without damage. We are interested in estimation of the density f_X from the observations $Z_1, ..., Z_n$. Vardi (1989) showed how this model can be useful in statistical problems, including non-parametric inference for renewal processes, certain non-parametric deconvolution

[*] *Corresponding author* : Mohammad Abbaszadeh
E-mail address : abbaszadehmo@yahoo.com

problems, and estimation of decreasing densities. It is straight-forward to derive the density f_Z can be expressed by f_Y in the following manner:

$$f_Z(z) = \int_z^\infty \frac{f_Y(y)}{y} dy. \tag{1.2}$$

The problem is indirect since we observe a multiplicative censoring version of Y and want to estimate the density of an unobserved X. Andersen and Hansen (2001) have considered density estimation using series expansion. The method proposed here is based on kernel smoothing that may be considered as extensions of the proposals by Bhattacharyya *et al.* (1988) and Jones (1991) for multiplicative censoring coupled with weighted data.

The organization of the rest of the paper is as follows. In Section 2, we motivate the form of the estimator based the kernel density estimator of f_Z and in Section 3, we obtain its MSE and MISE. Here the form of the optimal smoothing parameter is also derived that can be used adaptively in computations.

2. KERNEL DENSITY ESTIMATOR UNDER MULTIPLICATIVE CENSORING

Suppose we have a random sample $Z_1,..., Z_n$ taken from a continuous, univariate density f_Z, its kernel estimator [see Wand and Jones (1995)] is given by

$$\hat{f}_Z(z) = \frac{1}{n}\sum_{i=1}^{n} K_h(z - Z_i), \tag{2.1}$$

where

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right),$$

$K(.)$ being the kernel function that has the following properties:

$$K(z) \geq 0, \int K(z)dz = 1, \int zK(z)dz = 0,$$
$$0 < \int z^2K(z)dz < \infty. \tag{2.2}$$

A natural estimator of the r-th derivative $\hat{f}_Z^{(r)}(z)$, is given by taking the r-derivative of $\hat{f}_Z(z)$, namely

$$\hat{f}_Z^{(r)}(z) = \frac{1}{nh^{r+1}}\sum_{i=1}^{n} K^{(r)}\left(\frac{z - Z_i}{h}\right), \tag{2.3}$$

assuming that K is r-times differentiable.

It is clear that from Eq. (1.2) that $f_Y(z) = -zf'_Z(z)$, hence using Eq. (1.1), the pdf f_X can be written in terms of f_Z as

$$f_X(z) = \frac{\mu f_Y(z)}{g(z)} = \frac{-\mu_Z f'_Z(z)}{g(z)} \tag{2.4}$$

and therefore a natural estimator of $\hat{f}_X(z)$ is given by

$$\hat{f}_X(z) = \frac{-\hat{\mu}z\hat{f}'_Z(z)}{g(z)}. \tag{2.5}$$

An estimator $\hat{\mu}$ of parameter μ required in the above density estimator is given below that has the property that $E(1/\hat{\mu}) = 1/\mu$ [see Lemma 2.1 below] similar to the estimator in the length-biased case [see Cox (1969)]:

$$\hat{\mu} = \begin{cases} \left(\frac{1}{n}\sum_{i=1}^{n}\frac{g(Z_i) - g'(Z_i)Z_i}{g^2_{(Z_i)}}\right)^{-1} & \text{for} \quad g(z) \neq z \\ \left(\frac{1}{2n}\sum_{i=1}^{n}\frac{1}{Z_i}\right)^{-1} & \text{for} \quad g(z) = z. \end{cases} \tag{2.6}$$

Lemma 2.1 : Let $\{Z_1,..., Z_n\}$ be a random sample from density f_Z then $1/\hat{\mu}$ is unbiased for $1/\mu$.

Proof : First, it is easily verified that for $g(y) \neq y$, $\int_0^y \frac{g(z) - g'(z)z}{g^2(z)}dz = \frac{y}{g(y)}$. Hence, for $g(y) \neq y$,

$$\begin{aligned} E\left(\frac{1}{\hat{\mu}}\right) &= E\left(\frac{g(Z) - g'(Z)Z}{g^2(Z)}\right) \\ &= \int_0^\infty \frac{g(z) - g'(z)z}{g^2(z)} f_Z(z)dz \\ &= \int_0^\infty \frac{g(z) - g'(z)z}{g^2(z)} \int_z^\infty \frac{f_Y(y)}{y}dydz \\ &= \int_0^\infty \left(\int_0^y \frac{g(z) - g'(z)z}{g^2(z)}dz\right)\frac{g(y)f_X(y)}{\mu y}dy \\ &= \frac{1}{\mu}\int_0^\infty f_X(y)dy = \frac{1}{\mu}. \end{aligned}$$

For the case when $g(y) = y$, it is clear from (1.1) that $E(1/Y) = 1/\mu$. Further, from the representation $Z = UY$, we have

$$E\left(\frac{1}{Y}\right) = E(U)E\left(\frac{1}{Z}\right) = \frac{1}{2}E\left(\frac{1}{Z}\right).$$

Hence, $(1/2n)\sum_{i=1}^{n}(1/Z_i)$ is unbiased for $E(1/Y) = 1/\mu$ and the proof of the lemma is completed.

3. ASYMPTOTIC PROPERTIES OF THE ESTIMATORS

A common way to measure the performance of the estimator $\hat{f}_X(z)$ is the MISE given by

$$\text{MISE}(\hat{f}_X) \equiv R = \int R_z dz \tag{3.1}$$

where

$$R_z \equiv MSE(\hat{f}_X(z)) = E(\hat{f}_X(z) - f_X(z))^2 \tag{3.2}$$

Theorem 3.1. Assume that f_Z''' is absolutely continuous and that $\int (f_Z''')^2(z)\ dz < \infty$. Also, assume that K satisfies (2.2). Then MSE $\hat{f}_X(z))$ is given by

$$R_z = \frac{\mu^2 z^2}{g^2(z)}\left(\frac{1}{nh^3}R(K')f_Z(z) + \frac{h^4}{4}M_2^2(K)(f_Z''')^2(z)\right) + O\left(\frac{1}{nh^3} + h^4\right) \tag{3.3}$$

and the MISE$(\hat{f}_X)$ is given by

$$R = \frac{\mu^2}{nh^3}R(K')\int\left(\frac{z}{g(z)}\right)^2 f_Z(z)dz + \frac{\mu^2 h^4 M_2^2(K)}{4}\int\left(\frac{z}{g(z)}f_Z'''(z)\right)^2 dz + O\left(\frac{1}{nh^3} + h^4\right). \tag{3.4}$$

where

$$R(K') = \int K'^2(u)\ du$$

and

$$M_2(K) = \int u^2 K(u)\ du.$$

Proof : First we compute the expectation and variance of $\hat{f}_Z'(z)$. We have

$$\begin{aligned}
E(\hat{f}_Z'(z)) &= \frac{1}{h}E(K_h'(z - Z_i)) \\
&= \frac{1}{h}(K_h' * f_Z)(z) \\
&= \frac{1}{h}\frac{d}{dz}(K_h * f_Z)(z) \\
&= \frac{d}{dz}\left[\int \frac{1}{h}K\left(\frac{z-t}{h}\right)f_Z(t)\,dt\right] \\
&= \frac{d}{dz}\left[\int K(u) f_Z(z - hu)\,du\right] \\
&= \frac{d}{dz}\left[\int K(u)\left(f_Z(z) - huf_Z'(z) + \frac{h^2u^2}{2}f_Z''(z) + \ldots\right)du\right] \\
&= f_Z'(z) + \frac{h^2}{2}M_2(K)\,f_Z'''(z) + O(h^2),
\end{aligned}$$

$$\begin{aligned}
V(\hat{f}_Z'(z)) &= E(\hat{f}_Z'(z))^2 - (E(\hat{f}_Z'(z)))^2 \\
&= \frac{1}{n^2h^2}\ E\left(\sum_{i=1}^{n}K_h'(z - Z_i)\right)^2 - \frac{1}{h^2}(K_h' * f_Z)^2(z) \\
&= \frac{1}{nh^2}(K_h'^2 * f_Z)(z) - \frac{1}{n^2h^2}\ E\left(\sum_{i \neq j}K_h'(z - Z_i)\,K_h'(z - Z_j)\right) - \frac{1}{h^2}(K_h' * f_Z)^2(z) \\
&= \frac{1}{nh^2}(K_h'^2 * f_Z)(z) - \left(\frac{n(n-1)}{n^2} - 1\right)\frac{1}{h^2}(K_h' * f_Z)^2(z) \\
&= \frac{1}{nh^2}(K_h'^2 * f_Z)(z) - \frac{1}{nh^2}(K_h' * f_Z)^2(z) \\
&= \frac{1}{nh^2}\left[(K_h'^2 * f_Z)(z) - (K_h' * f_Z)^2(z)\right] \\
&= \frac{1}{nh^2}\left[\int K_h'^2(z - t)\,f_Z(t)dt - \int K_h'(z - t)\,f_Z(t)dt\right] \\
&= \frac{1}{nh^4}\left[\int K'^2\left(\frac{z-t}{h}\right)f_Z(t)dt - \frac{1}{nh^3}\int K'\left(\frac{z-t}{h}\right)f_Z(t)dt\right] \\
&= \frac{1}{nh^3}\left[\int K'^2(u)\,f_Z(z - uh)du - \frac{1}{nh^2}\int K'(u)\,f_Z(z - uh)du\right]
\end{aligned}$$

$$= \frac{1}{nh^3}\int K'^2(u)\left(f_Z(z)+O(1)\right)du$$

$$-\frac{1}{nh^2}\int K'(u)\left(f_Z(z)+O(1)\right)du$$

$$= \frac{1}{nh^3}R(K')f_Z(z)+O\left(\frac{1}{nh^3}\right).$$

Hence

$$MSE(\hat{f}'_Z(z)) = \frac{1}{nh^3}R(K')f_Z(z)$$

$$+\frac{h^4}{4}M_2^2(K)\,(f_Z''')^2(z)+O\left(\frac{1}{nh^3}+h^4\right)$$

and

$$MISE(\hat{f}'_Z) = \frac{1}{nh^3}R(K')$$

$$+\frac{h^4}{4}M_2^2(K)\int(f_Z''')^2(z)dz+O\left(\frac{1}{nh^3}+h^4\right).$$

Next we note that the estimate $1/\hat{\mu}$ in Lemma 2.1 is unbiased and strongly consistent. Hence,

$$MISE(\hat{f}'_X(z)) = \frac{\mu^2 z^2}{g^2(z)}MSE(\hat{f}'_Z(z))$$

$$= \frac{\mu^2 z^2}{g^2(z)}\left(\frac{1}{nh^3}R(K')f_Z(z)\right.$$

$$\left.+\frac{h^4}{4}M_2^2(K)\,(f_Z''')^2(z)\right)+O\left(\frac{1}{nh^3}+h^4\right)$$

and

$$MISE(\hat{f}_X) = \frac{\mu^2}{nh^3}R(K')\int\left(\frac{z}{g(z)}\right)^2 f_Z(z)dz$$

$$+\frac{\mu^2 h^4 M_2^2(K)}{4}\int\left(\frac{z}{g(z)}f''^{Z}(z)\right)^2 dz$$

$$+O\left(\frac{1}{nh^3}+h^4\right).$$

This completes the proof of the theorem.

The asymptotic optimal bandwidth that minimizes $MISE(\hat{f}_X)$ is seen to be given by

$$h_* = \left(\frac{3R(K')\int\left(\frac{z}{g(z)}\right)^2 f_Z(z)dz}{nM_2^2(K)\int\left(\frac{z}{g(z)}f_Z'''(z))^2 dz\right)}\right)^{\frac{1}{7}} \tag{3.5}$$

Thus we find that the best bandwidth decreases at rate $n^{-\frac{1}{7}}$, and plugging h_* into (3.4) shows that if the optimal bandwidth is used then $R = O(n^{-\frac{4}{7}})$. Comparing this with the optimal rate of order $n^{-\frac{4}{5}}$ attainable in the case of kernel density estimation for the *i.i.d.* data, we conclude that the optimal rate in the present case is somewhat slower than that for the *i.i.d.* data.

Remark : Chaubey and Srivastava (1991) have considered characterizing the distribution of random variables subject to multiplicative censoring where the censoring random variable U is not necessarily uniform. For example U may have a general $Beta(p, q)$ distribution. The proposed estimator can be easily generalized to this case.

REFERENCES

Andersen, K. and Hansen, M. (2001). Multiuplicative censoring: Density estimation by a series expansion approach. *Jour. Statist. Plann. Inf.*, **98**, 137-155.

Bhattacharyya, B.B., Franklin, L.A. and Richardson, G.D. (1988). A comparison of nonparametric unweighted and length-biased density estimation of fibres. *Comm. Statist.-Theory Methods*, **A17**, 3629-3644.

Chaubey, Y.P., Sen, P.K. and Li, Jun (2010). Smooth density estimation for length-biased data. *J. Ind. Soc. Agril. Statist.*, **64(2)**, 145-155.

Chaubey, Y.P. and Srivastava, T.N. (1991). On a multiplicative damage model and characterization of some distributions useful in growth models. *Quaderni di Statistica e Matematica*, **13**, 23-32.

Cox, D.R. (1969). Some sampling problems in technology. In : *New Developments in Survey Sampling*. N.L. Johnson and H. Smith (eds.). John Wiley, New York, 506-527.

Guillamon, A., Navarro, J. and Ruiz, J.M. (1998). Kernel density estimation using weighted data. *Comm. Statist. - Theory Methods*, **27(9)**, 2123-2135.

Gupta, R.C. and Kirmani, S.N.U.A. (1990). The role of weighted distributions in stochastic modeling. *Comm. Statist.-Theory Methods*, **19(9)**, 3147-3162.

Jones, M.C. (1991). Kernel density estimation for length biased data. *Biometrika*, **78**, 511-519.

Nair, N.U. and Sunoj, S.M. (2003). Form-invariant bivariate weighted models. *Statistics*, **37(3)**, 259-269.

Navarro, J., Del Aguila, Y. and Ruiz, J.M. (2001). Characterizations through reliability measures from weighted distributions. *Statist*. *Papers,* **42**, 395-402.

Patil, G.P. and Ord, J.K. (1975). On size-biased sampling and related form-invariant weighted distributions. *Sankhyā* , **B38**, 48-61.

Patil, G.P. and Rao, C.R. (1977). Weighted distributions: A survey of their applications. In : *Applications of Statistics*. P.R. Krishnaiah, (ed.), North Holland Publishing Company, 383-405.

Patil, G.P. and Rao, C.R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**, 179-189.

Ramirez, P. and Vidakovic, B. (2010). Wavelet density estimation for stratied size-biased sample. *J. Statist. Plann. Inf.*, **140**, 419-432.

Rao, C.R. (1965). On discrete distributions arising out of methods of ascertainment. In : *Classical and Contagious Discrete Distribution*. G. P. Patil (ed.), Pergamon Press and Statistical Publishing Society, Calcutta, 320-332.

Rao, C.R. (1977). A natural example of a weighted binomial distribution. *Amer. Stat.*, **31**, 24-26.

Sunoj, S.M. (2004). Characterizations of some continuous distributions using partial moments. *Metron*, **LXII(3)**, 353-362.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, **76**, 751-761.

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