Transformation–based density estimation for weighted distributions

Hammou El Barmi[1] and Jeffrey S. Simonoff[2]
[1]Department of Statistics, Kansas State University
[2]Department of Statistics and Operations Research, New York University

Abstract

In this paper we consider the estimation of a density f on the basis of random sample from a weighted distribution G with density g given by

$$g(x) = w(x)f(x)/\mu_w,$$

where $w(u) > 0$ for all u and

$$\mu_w = \int w(u)f(u)\,du < \infty.$$

A special case of this situation is that of length–biased sampling, where $w(x) = x$. In this paper we examine a simple transformation–based approach to estimating the density f. The approach is motivated by the form of the nonparametric estimator of f in the same context and under a monotonicity constraint. Since the method does not depend on the specific density estimate used (only the transformation), it can be used to construct both simple density estimates (histograms or frequency polygons) and more complex methods with favorable properties (e.g., local or penalized likelihood estimates). Monte Carlo simulations indicate that transformation–based density estimation can outperform the kernel–based estimator of Jones (1991) depending on the weight function w, and leads to much better estimation of monotone densities than the nonparametric maximum likelihood estimator.

AMS 1980 Subject Classification: Primary: 62G07, Secondary: 62G05, 62D05.

Some key words and phrases: Density estimation, isotonic regression, selection bias, weighted distributions.

1 Introduction

Weighted distributions are used in statistics to model selection biased sampling. They arise, for example, when observations do not have an equal chance of being recorded. Suppose that an observation viewed without bias has probability density function (pdf) f and let $w(x)$ be (proportional to) the probability of recording the observation with value x. Then, the pdf of the recorded observation is

$$g(x) = \frac{w(x)f(x)}{\mu_w}, \tag{1.1}$$

where $\mu_w = E_F(w(X))$ is the normalizing factor which makes the total probability equal to one. Length (size) bias occurs when $w(x) = x$ with f supported on the positive half–line $[0, \infty)$, and arises naturally in industrial and work sampling, in sampling from stochastic processes such as queues, telephone networks and renewal processes. An illustration of this is given in Vardi (1982) and concerns m independent and identically distributed stationary renewal processes with a common underlying distribution function F. Suppose that from each such process we sample the inter-arrival time that includes a fixed time T (the time point T is assumed to be independent of the process itself). This sampling scheme (which gives rise to the well known inspection paradox) results for large T in an approximate sample of size m from G where

$$G(x) = \frac{1}{\mu_F} \int_0^x x dF(x),$$

with μ_F being the mean of F. This setup holds also, for example, when sampling velocities of passing and passed cars from a car traveling at a fixed velocity μ_0. These velocities are, under some regularity conditions, selected from a weighted distribution, G, not the underlying distribution of velocities, F. To be specific, consider an infinite highway (the interval $(-\infty, +\infty)$), where the initial positions of the cars are determined according to a Poisson process with rate λ. Assume that each car's velocity is constant along the highway and that the velocities are independent and identically distributed and independent of the positions. Breiman (1962) showed that under these conditions the only time invariant point process for car positions is the Poisson process. Furthermore, it can be shown that the velocities sampled form a random sample from

$$G(y) = \frac{1}{E_F|x - \mu_0|} \int_0^y |x - \mu_0| dF(x).$$

See Smith and Parnes (1994) for a further discussion on sampling velocities by an observer.

Estimation based on sampling from weighted distributions has been considered before. Vardi (1982) and Vardi (1985) give, respectively, the nonparametric maximum likelihood estimator (NPMLE) of the underlying distribution function in the presence of length bias and based on several independent samples, each subject to a different selection bias. El Barmi and Rothmann (1998) consider the estimation of a distribution function in the presence of selection bias when it is known that there is a d–dimensional parameter $\boldsymbol{\theta}$ associated with the distribution function through a set of estimating equations. They also obtain a test of $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ based asymptotically on a χ^2_d distribution. This test is closely connected to the results obtained for empirical likelihoods in a selection biased model by Qin (1993).

The problem of estimating a density is of major importance in statistics and has been widely studied using many methods, including histograms, frequency polygons, kernel estimates, nearest neighbor estimates, local and penalized likelihood estimates, and so on. See Simonoff (1996, chapter 3) for discussion of many of these methods. Bandwidth selection (i.e., determination of the amount of smoothing) is a problem with all of these approaches and one way to avoid this difficulty is to place shape restrictions on the estimator such as unimodality or monotonicity. Grenander(1956) derived the maximum likelihood estimator under such constraints which turned out to be a histogram where the bandwidth depends on the data. Nonparametric maximum likelihood estimation of a probability density functions has been studied by several authors. Robertson (1965, 1967) considered the estimation of a unimodal density function when the mode is known and Wegman (1970) extended these results to the case where the mode is unknown. Prakasa Rao (1969) obtained the asymptotic distribution of the maximum likelihood estimator of a unimodal density with a known mode and the pointwise limiting distribution of an isotonic estimator was obtained by Brunk (1970) and Wright (1981). Recently, El Barmi and Nelson (1998) extended Prakasa Rao's theorem to the weighted distributions case. They show that the NPMLE of a monotone density f when sampling from g is a scaled Grenander type estimator after transforming the data (note that g might, but need not, be monotone and can take on a variety of shapes). To be specific, suppose $X_1, X_2, \ldots, X_n$, is a random sample of size n from G whose probability density function is given by g where g is defined by (1.1). Let $X_{(0)} = 0$ and define

$$a_{in} = \int_{X_{(i-1)}}^{X_{(i)}} w(u)du, \quad i = 1, 2, \ldots, n.$$

El Barmi and Nelson (1998) proved the following theorem.

Theorem 1.1 *The nonparametric maximum likelihood estimator $\hat{f}_n$ of f is given by*

$$\hat{f}_n(x) = \begin{cases} \dfrac{E_{\mathbf{a}_n}(1/n\mathbf{a}_n|\mathcal{K})_i}{\sum_{i=1}^{n} E_{\mathbf{a}_n}(1/n\mathbf{a}_n|\mathcal{K})_i (X_{(i)} - X_{(i-1)})}, & X_{(i-1)} < x \leq X_{(i)}, \quad i = 1, 2, \ldots, n, \\ 0, & \textit{otherwise}, \end{cases}$$

where $\mathbf{a}_n = (a_{1n}, a_{2n}, \ldots, a_{nn})$ and $E_{\mathbf{a}_n}(1/n\mathbf{a}_n|\mathcal{K})$ is the least square projection of $1/n\mathbf{a}_n$ onto $\mathcal{K} = \{(u_1, u_2, \ldots, u_n), \quad u_1 \geq u_2 \geq \cdots \geq u_n\}$ with weights $a_{1n}, a_{2n}, \ldots, a_{nn}$.

It turns out that $\hat{f}_n$ can also be expressed using the properties of isotonic regression as

$$\hat{f}_n(x) = \begin{cases} \dfrac{\min\limits_{s \leq i-1} \max\limits_{t \geq i} \dfrac{t/n - s/n}{W(X_{(t)}) - W(X_{(s)})}}{\sum_{i=1}^{n} \min\limits_{s \leq i-1} \max\limits_{t \geq i} \dfrac{t/n - s/n}{W(X_{(t)}) - W(X_{(s)})} (X_{(i)} - X_{(i-1)})}, & X_{(i-1)} < x \leq X_{(i)} \\ 0, & \text{otherwise}, \end{cases}$$

where $W(x) = \int_0^x w(u)du$.

Note that if $w(x) = 1$ for all x (i.e. the sample if from f), then $a_{in} = X_{(i)} - X_{(i-1)}, i = 1, 2, \ldots, n$, and it follows from the properties of isotonic regression that

$$\sum_{i=1}^{n} E_{\mathbf{a}_n}(\frac{1}{n\mathbf{a}_n}|\mathcal{K})_i (X_{(i)} - X_{(i-1)}) = \sum_{i=1}^{n} \frac{1}{na_{ni}} (X_{(i)} - X_{(i-1)}) = 1,$$

so that $\hat{f}_n$ is the Grenander estimator of f. We also note that the least squares projection in the previous theorem can be computed using one of the algorithms described in Robertson, Wright and Dykstra (1988) such as the Pool Adjacent Violators Algorithm (PAVA).

Let $m(.)$ be the density of $Y_1 = W(X_1)$. Careful inspection of $\hat{f}_n$ shows that the numerator in the previous formula is the NPMLE $\hat{m}_n(.)$ of $m(.)$ evaluated at $W(x)$ under the constraint that $m(.)$ is non-decreasing based on the random sample $Y_1, y_2, \ldots, Y_n$ where $Y_i = W(X_i), i = 1, 2, \ldots, n$. That is, $\hat{f}_n(x)$ is proportional to $\hat{m}_n(W(x))$.

Note that if $Y_i = W(X_i)$, where

$$W(x) = \int_0^x w(u)\, du,$$

then

$$M(y) \equiv P(Y \leq y) \quad = \quad P[X \leq W^{-1}(y)]$$

$$= \quad G[W^{-1}(y)].$$

Differentiating $M(y)$ gives the density of y:

$$\begin{aligned} m(y) &= \frac{g[W^{-1}(y)]}{w[W^{-1}(y)]} \\ &= \frac{w[W^{-1}(y)]f[W^{-1}(y)]/\mu_w}{w[W^{-1}(y)]} \\ &= \frac{f[W^{-1}(y)]}{\mu_w}. \end{aligned}$$

Thus,

$$f(x) = m[W(x)]\mu_w. \tag{1.2}$$

That is, the transformation to $y = W(x)$ results in a density estimation problem (estimating m) identical to the unweighted density estimation problem (estimating f) after back–transforming from the $y = W(x)$ argument to x and scaling to integrate to one. So, for example, length–biased sampling is the special case $W(x) = x^2/2$, so the estimation strategy is to estimate the density in the $y = x^2/2$ space and then back–transform:

$$\tilde{f}(x) = \hat{m}(x^2/2)\hat{\mu}_w.$$

In this paper we will examine several ways that the transformation approach can be used in density estimation for weighted distributions. Depending on how sophisticated one wants to be, and how much one wants to assume, this approach can be applied in different ways, which is the strength of this approach. A potential advantage for this approach is that since density estimation is done in the y space, all of the literature on density estimation can be applied directly (e.g., bandwidth selection to control the amount of smoothing). This statement is not precisely correct, as it focuses on producing good estimates of m, rather than f, but (1.2) implies that a good estimate of m should result in a good estimate of f. If μ_w can be estimated to $\sqrt{n}$–consistency, then the convergence rates of density estimates for m will carry over directly to estimation of f by (1.2); a simple way to accomplish this is to standardize the estimate to integrate to one after constructing $\hat{m}$.

2 Simple density estimates

The simplest density estimator is the histogram, which takes the form

$$\hat{u}(x) = \frac{(\#\{X_i \leq b_{j+1}\} - \#\{X_i \leq b_j\})/n}{b_{j+1} - b_j}, \qquad x \in (b_j, b_{j+1}],$$

where u is a density function and $(b_j, b_{j+1}]$ defines the boundaries of the jth bin. The histogram is the unique maximum likelihood density estimator over the set of estimates of f that are piecewise constant on the set of bins (de Montricher, Tapia, and Thompson, 1975). By the invariance of the maximum likelihood, a simple transformation–based density estimation scheme for weighted distributions, which is also maximum likelihood, is to estimate m using a histogram, and then back–transform.

This strategy is particularly natural if the data are continuous, but are presented in the form of rounded (binned) counts. In this situation the histogram could be based on the pre–binning, or bins could be combined to yield a smoother estimate.

A potentially more accurate density estimate that is as easy to construct as the histogram is the frequency polygon, which connects the midbin heights of the histogram bins with straight lines:

$$\begin{aligned} \hat{u}(x) &= \frac{1}{n(c_{j+1} - c_j)} \left[\frac{n_j c_{j+1}}{b_{j+1} - b_j} - \frac{n_{j+1} c_j}{b_{j+2} - b_{j+1}} \right. \\ &\quad \left. + \left(\frac{n_{j+1}}{b_{j+2} - b_{j+1}} - \frac{n_j}{b_{j+1} - b_j} \right) x \right], \qquad x \in [c_j, c_{j+1}], \end{aligned}$$

where $\{c_0, \ldots, c_{K+1}\}$ are the midpoints of the bin intervals. Besides producing an estimate that is more pleasing to the eye, if the bin widths are chosen appropriately, the resultant density estimator converges to the true density at a $n^{-2/5}$ rate, rather than the $n^{-1/3}$ rate of the histogram. Other frequency polygon–like estimators with desirable properties also can be constructed; see Jones, Samiuddin, Al–Harbey, and Maatouk (1998).

Figure 1 illustrates the use of the transformation–based frequency polygon estimator. The data are the number of years players in the National Basketball Association (NBA) have played in the league, for members of the 1998 rosters of the teams from Miami, New Jersey, New York, Orlando, Philadelphia, and Washington. The data were obtained from the NBA's official World Wide Web site. The data are given rounded up to the nearest integer. Figure 1(a) is a frequency polygon of the data using a bin width

of one year (the binning in the data as given). The frequency polygon is not a representation that can be used to estimate the probability that a randomly selected NBA player will stay in the league a given number of years, since the longer a player plays, the more likely he is to be observed. That is, the data are length–biased ($w(x) = x$).

Figure 1(b) is a transformation–based length–bias–corrected frequency polygon estimate using the same binning as in Figure 1(a). The estimate is constructed by transforming the x values to $y = x^2/2$, constructing a variable bin width frequency polygon for $m(y)$, and then back–transforming using (1.2) to get the estimate for f. Note that the bias towards larger values is corrected. Figure 1(c) is a transformation–based frequency polygon when binning into six bins. The shape of the estimate is similar to that in Figure 1(b), but is smoother.

3 Smoother density estimation

Given the assumption of a smooth true density f, it is reasonable to try to construct a density estimate that is smoother than the histogram or frequency polygon forms of the previous section. Jones (1991) proposed a weighted kernel estimator for this problem:

$$\hat{f}(x) = n^{-1}\hat{\mu}_w \sum_{i=1}^{n} w(X_i)^{-1} K\left(\frac{x - X_i}{h}\right),$$

where $\hat{\mu}_w = n\Big[\sum_i w(X_i)^{-1}\Big]^{-1}$ and K is the kernel function (typically a symmetric unimodal density with finite variance). This estimate suffers from boundary bias (as all kernel estimators do), but this can be corrected by using boundary kernel functions; see Simonoff (1996, sections 3.2.1 and 3.3.1) for discussion. Jones (1991) demonstrated good properties of the estimate compared to an earlier proposal of Bhattacharyya, Franklin and Richardson (1988). Further discussion of kernel–type estimation for weighted distributions can be found in Richardson, Kazempour, and Bhattacharyya (1991), Ahmad (1995), Wu and Mao (1996), Wu (1997a, b), and Guillamon, Navarro, and Ruiz (1998).

Smooth density estimates also can be constructed for weighted distributions based on transformation. The only question is then how to estimate m, and by implication f. The obvious first choice is a kernel estimator:

$$\tilde{f}(x) = \frac{\hat{\mu}_w}{nh} \sum_{i=1}^{n} K\left(\frac{W(x) - W(X_i)}{h}\right).$$

Unfortunately, the length–biased sampling problem illustrates a difficulty with the transformation–based approach. If the density m is difficult to estimate, the resultant estimate of f can be less accurate than when using a more direct approach. For example, for a unimodal x density (f) the density of $y = W(x) = x^2/2$ (m) has a sharp rise followed by a long right tail. This is a difficult density estimation problem, particularly for the kernel estimator; the estimate either is severely biased at low values, or has spurious bumps at high values (for discussion see Simonoff, 1996, section 3.2.2).

This suggests estimating m using a method better suited for data that might have long tails. One possibility is the local quadratic likelihood density estimator of Hjort and Jones (1996) and Loader (1996), which is the maximizer of

$$\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h}\right) \log[m(Y_i)] - n \int K\left(\frac{y-u}{h}\right) m(u)du$$

over the family $m(t, \theta) = \theta_0 \exp[\theta_1(t-y) + \theta_2(t-y)^2]$. This estimator has the advantage of automatically correcting for boundary bias, achieves faster convergence rates in the interior, and is more accurate in the tails than the usual kernel estimator. For discussion see Simonoff (1996) section 3.4.

Table 1 summarizes the results of a small simulation study comparing the Jones (1991) estimator with the transformation–based local quadratic likelihood estimator. Two weight functions, $w(x) = x$ and $w(x) = 1/x$, are used. The data are generated based on a χ^2_k random variable, which has the advantage that weighted distribution data are easy to generate (if $w(x) = x$ the weighted data are χ^2_{k+2}, while if $w(x) = 1/x$ the weighted data are χ^2_{k-2}). There were 500 Monte Carlo replications at two different settings of k for each weight function (reflecting an asymmetric and roughly symmetric true density) for a small ($n = 50$) and moderate ($n = 200$) sample size. The values reported are the average minimum integrated squared error ($ISE = \int [\hat{f}(x) - f(x)]^2 \, dx$) when the smoothing parameter is chosen in each replication to minimize ISE.

As would be expected, the more symmetric densities (larger k) are easier to estimate than the asymmetric ones. The dominant effect, however, comes from the weight function. While the kernel–based estimator is noticeably more accurate when $w(x) = 1/x$ compared with when $w(x) = x$, this effect is far stronger for the transformation–based estimator. As was noted earlier, when $w(x) = x$ the density is estimated in the $W(x) = x^2/2$ space, which is difficult; this results in comparatively poorer performance. On the other hand, when $w(x) = 1/x$ the density is estimated in the $W(x) = \log(x)$ space, which is a generally favorable

transformation. As a result, the transformation–based estimator is considerably more accurate than the kernel–based estimator.

Figure 2 illustrates an example with inverse size bias ($w(x) = 1/x$). An important consideration for university faculty, administrators, and students is class size. It is relatively easy to sample classes for class size, but this provides an estimate of typical class size at the class level. From the student's point of view, these data are inverse size biased, since the chances of being in a class increase with the size of the class (i.e., sampling classes leads to undersampling of students in larger classes).

The data used are the number of students responding to end–of–semester course evaluations for Spring 1998 MBA core courses at New York University's Leonard N. Stern School of Business. Figure 2(a) is the Jones kernel–based estimator. It can be seen that the kernel estimator has trouble with the long–tailed density, since the weighting leads to bumps at each of the large values. The transformation–based local quadratic estimate (Figure 2(b)) has a similar shape to the Jones estimator, including a bulge at around 30 students and a peak between 50 and 60 students, but the right tail is estimated smoothly, since estimation is done in the log scale.

The smoothing parameters for the two densities in Figure 2 were chosen by eye. One advantage of the transformation–based approach is that automatic selectors can be used in the transformed space; this should be reasonably effective, although it does not guarantee good performance in the original scale. Simonoff (1998) discussed an automatic selector that can be used for the local quadratic likelihood estimator, although for these data it leads to a somewhat under smoothed estimate.

4 Multiple Samples

In this section we consider estimating f on the basis of several samples each of which is subject to a different form of selection bias. Let $X_{i1}, X_{i2}, \ldots, X_{in_i}, i = 1, 2, \ldots, k$, be independent random samples and assume that X_{ij} has a probability density function given by

$$g_i(x) = \frac{w_i(x)f(x)}{\mu_{w_i}}, \qquad j = 1, 2, \ldots, n_i, \quad i = 1, 2, \ldots, k,$$

where $w_i(\cdot), i = 1, 2, \ldots, k$, are known functions, $w_i(.) > 0$ for some i and $\mu_{wi} = E_F(w_i(X)) < \infty$ for all i. The estimator we propose to use in this case is a convex combination of the estimators of f based on the

different samples and is given by

$$\tilde{f}(x) = \sum_{i=1}^{k} \alpha_i \tilde{f}_i(x),$$

where $\alpha_i = n_i / \sum_{j=1}^{k} n_j$ and $\tilde{f}_i$ is the estimator of f based on the ith sample. So, for example, if kernel estimators are used, then

$$\tilde{f}_i(x) = \frac{\hat{\mu}_{w_i}}{n_i h} \sum_{j=1}^{n_i} K\left(\frac{W_i(x) - W_i(X_{ij})}{h}\right)$$

where $W_i(x) = \int_0^x w_i(u)du, i = 1, 2, \ldots, k$.

If it is the case that $k = 2, w_1(.) \equiv 1$ and $w_2(x) = x$, for all x, then

$$\tilde{f}(x) = \frac{n_1}{n_1 + n_2} \frac{1}{n_1 h} \sum_{j=1}^{n_1} K\left(\frac{x - X_{1j}}{h}\right) + \frac{n_2}{n_1 + n_2} \frac{\hat{\mu}_{w_2}}{n_2 h} \sum_{j=1}^{n_2} K\left[\frac{(x^2 - X_{2j}^2)/2)}{h}\right].$$

Finally we note that because of independence, the properties of $\tilde{f}$ follow immediately from those of $\tilde{f}_i$, $i = 1, 2, \ldots, k$. For example $MSE(\tilde{f}(x)) = \sum_{i=1}^{k} \alpha_i^2 MSE(\tilde{f}_i(x))$.

5 Monotone densities

The nonparametric maximum likelihood estimate of El Barmi and Nelson (1998) assumes a monotone density, since otherwise the estimate cannot be constructed. It is possible that a data analyst might believe that the true underlying density is both smooth and monotone. The transformation–based approach makes it easy to construct a monotone density estimate as long as the transformation is one–to–one (one–to–one transformations include those for inverse size bias, and size bias if the data are nonnegative), since a monotone density in the transformed space will be monotone in the original space.

Density estimators can be constructed from regression estimators by binning the data and smoothing the observed frequencies in each bin (see Simonoff, 1998), so monotone regression estimators can be adapted to produce monotone density estimates. In this way the frequency polygon can be adapted to be monotone using the corresponding "pool adjacent violators" algorithm for regression function estimation. Smoother density estimates also can be constructed using smooth monotone regression functions; see, e.g., Ramsay (1988).

Table 2 summarizes the results of a small simulation study comparing the (monotone) nonparametric MLE with two transformation–based monotone frequency polygon estimator. Two weight functions, $w(x) = x$ and $w(x) = 1/x$, are used. The data are generated based on either a uniform distribution

on $(1, 2)$, or (if $w(x) = x$) an exponential density (resulting in a Gamma(2, 1) weighted density) or (if $w(x) = 1/x$) a Gamma(2, 1) density over $[1, \infty)$ rescaled to integrate to one (resulting in a shifted exponential weighted density). There were 500 Monte Carlo replications for each weight function for a small ($n = 50$) and moderate ($n = 200$) sample size. The values reported are the average ISE for the MLE, the optimal monotone frequency polygon (in the sense that the number of bins was chosen to minimize the ISE for that simulation run), and a Gaussian–based monotone frequency polygon. The latter estimator is based on choosing the bin width in the x space for each run to be $2.15sn^{-1/5}$ (where s is the sample standard deviation of the observed x values), the optimal choice if $g(x)$ is Gaussian. This data–based choice is clearly overly simplistic, but provides a crude way to choose the bin width in practice.

Since the true densities being estimated are, in fact, smooth, it would be expected that a frequency polygon can outperform the nonsmooth MLE, and this is, in fact, the case. The average ISE's for the optimal frequency polygon are 35–95% smaller than those of the MLE. The generally strong performance of the data–based frequency polygon shows that even a naive application of the frequency polygon beats the MLE; no doubt a better choice of bin width would achieve performance even closer to that of the optimal frequency polygon.

Figures 1(b) and 1(c) suggest that the probability function for the number of years in the NBA could be monotonic. Figure 3 gives a monotonic version of the frequency polygon in Figure 1(b), created using a pooled adjacent violators algorithm on the estimate of Figure 1(b). The original frequency polygon is also superimposed on the plot. The monotonic version smooths the frequency polygon further, removing the bumpiness in the original estimate. An even smoother version can be created using wider (and fewer) bins. For comparative purposes, the (nonsmooth) nonparametric MLE is also given on the plot.

6 Conclusion

In this paper a simple transformation–based technique is proposed for density estimation for weighted distributions. The approach has the advantage of being widely applicable, allowing simple density estimation using histograms and frequency polygons, smoother estimation using (for example) local likelihood estimates, and monotone density estimation using isotonic regression techniques. The general nature of the approach means that any methods used for ordinary samples also can be used for weighted distribution

data, including, for example, censored data, densities with sharp edges and jumps, and multivariate data.

Naturally, if the weight function w was under the control of the data analyst, the choice $w = 1$ corresponding to unbiased sampling would be best. As was noted in the introduction, different observation processes lead to different forms of w, typically out of the analyst's control. The simplicity of the transformation–based technique encourages its routine application, but the simulation results of section 3 do suggest that if the weight function w is such that the transformed data have a density that is difficult to estimate (because of long tails, for example), other estimates might be preferable.

References

1. Ahmad, I.A. (1995). On multivariate kernel estimation for samples from weighted distributions. *Statistics and Probability Letters*, **22**, 121–129.

2. Bhattacharyya, B.B., Franklin, L.A., and Richardson, G.D. (1988). A comparison of nonparametric unweighted and length–biased density–estimation of fibers. *Communications in Statsitics — Theory and Methods*, **17**, 3629–3644.

3. Breiman, L. (1962). On some probability distributions occurring in traffic flow. *Bulletin of the International Statistical Institute, Paris*, **33**, 155-161.

4. Brunk, H.D. (1970). Estimation of isotonic regression (with discussion). In M.L. Puri (ed). *Nonparametric Techniques in Statistical Inference*, Cambridge University Press, 177–197.

5. de Montricher, G.M., Tapia, R.A., and Thompson, J.R. (1975). Nonparametric maximum likelihood estimation of probability densities by penalty function methods. *Annals of Statistics*, **3**, 1329–1348.

6. El Barmi, H. and Nelson, P. (1998). Restricted density estimation in selection biased models. *Technical Report # 98–5*, Department of Statistics, Kansas State University.

7. El Barmi, H. and Rothmann, M. (1998). Nonparametric estimation in selection biased models in the presence of estimating equations. *Journal of Nonparametric Statistics*, **9**, 381–399.

8. Grenander, U. (1956). On the theory of mortality measurement, *Part II, Skand. Akt.*, **39**, 125–153.

9. Guillamon, A., Navarro, J., and Ruiz, J.M. (1998). Kernel density estimation using weighted data. *Communications in Statistics — Theory and Methods*, **27**, 2123–2135.

10. Hjort, N.L. and Jones, M.C. (1996). Locally parametric density estimation. *Annals of Statistics*, **24**, 1619–1647.

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

12. Jones, M.C., Samiuddin, M., Al–Harbey, A.H., and Maatouk, T.A.H. (1998). The edge frequency polygon. *Biometrika*, **85**, 235–239.

13. Loader, C.R. (1996). Local likelihood density estimation. *Annals of Statistics*, **24**, 1602–1618.

14. Prakasa Rao, B.L.S. (1969). Estimation of unimodal density. *Sankhya A*, **31**, 23–36.

15. Qin, J. (1993). Empirical likelihood in biased sample problems. *Annals of Statistics*, **21**, 1182–1196.

16. Ramsay, J.O. (1988). Monotone regression splines in action (with discussion). *Statistical Science*, **3**, 425–461.

17. Richardson, G.D., Kazempour, M.K., and Bhattacharyya, B.B. (1991). Length biased density estimation of fibres. *Journal of Nonparametric Statistics*, **1**, 127–141.

18. Robertson, T. (1965). A note on the reciprocal of the conditional expectation of a positive random variable. *Annals of Statistics*, **36**, 1302–1205.

19. Robertson, T. (1967). On estimating a density which is measurable with respect to a σ-lattice. *Annals of Statistics*, **38**, 482–493.

20. Simonoff, J.S. (1996), *Smoothing Methods in Statistics*, New York: Springer–Verlag.

21. Simonoff, J.S. (1998). Three sides of smoothing: categorical data smoothing, nonparametric regression, and density estimation. *International Statistical Review*, **66**, 137–156.

22. Smith, W. and Parnes, M. (1994). Mean streets: the median of a size-biased sample and the population mean. *American Statistician*, 106–110.

23. Vardi, Y. (1982). Nonparametric distribution in the presence of length bias. *Annals of Statistics*, **10**, 616–620.

24. Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics*, **13**, 178–203.

25. Wegman, E. (1970). Maximum likelihood estimation of a unimodal density function. *Annals of Statistics*, **2**, 457–471.

26. Wu, C.O. (1997a). A cross–validation bandwidth choice for kernel density estimates with selection biased data. *Journal of Multivariate Analysis*, **61**, 38–60.

27. Wu, C.O. (1997b). The effects of kernel choices in density estimation with biased data. *Statistics and Probability Letters*, **34**, 373–383.

28. Wu, C.O. and Mao, A.Q. (1996). Minimax kernels for density estimation with biased data. *Annals of the Institute of Statistical Mathematics*, **48**, 451–467.

Table 1. Average minimum ISE values from Monte Carlo investigation of Jones (1991) kernel–based estimator and transformation–based local quadratic likelihood estimator for different weight functions and sample sizes. Data were generated based on a true χ^2_k density.

$w(x) = x$				$w(x) = 1/x$			
k	n	*Kernel–based*	*Transformation–based*	k	n	*Kernel–based*	*Transformation–based*
2	50	.0312	.0309	3	50	.0155	.00568
	200	.0155	.0317		200	.00744	.00216
12	50	.002033	.00434	16	50	.00179	.00105
	200	.000891	.00184		200	.000691	.000293

Table 2. Average minimum ISE values from Monte Carlo investigation of nonparametric MLE and monotone frequency polygons for different weight functions and sample sizes. Data were generated based on a true uniform(1, 2) density, an exponential density ($w(x) = x$), or a truncated Gamma(2, 1) density ($w(x) = 1/x$).

$w(x) = x$

Density	n	*MLE*	*Optimal frequency polygon*	*Gaussian–based f.p.*
Uniform	50	.02209	.00571	.00824
	200	.00612	.00134	.00496
Exponential	50	.47191	.03326	.10587
	200	.29408	.01274	.04190

$w(x) = 1/x$

Density	n	*MLE*	*Optimal frequency polygon*	*Gaussian–based f.p.*
Uniform	50	.10458	.00595	.01366
	200	.02723	.00156	.00402
Truncated gamma	50	.03466	.02243	.04803
	200	.01085	.00441	.00859

Figure 1. Frequency polygons for years in NBA data. (a) Frequency polygon without correction for length bias. (b) Transformation–based length–bias–corrected frequency polygon. (c) Transformation–based length–bias–corrected frequency polygon based on six bins.