Wavelet density estimation for stratified size-biased sample

Pepa Ramírez[a,*], Brani Vidakovic[b]

[a] *GIPSA-lab, Dept. Images and Signals, 38402 Saint Martin d'Hères Cedex, France*

[b] *Georgia Institute of Technology, Atlanta, GA 30332, USA*

April 23, 2009

Abstract

In this paper we consider the estimation of a density function on the basis of a random stratified sample from weighted distributions. We propose a linear wavelet density estimator and prove its consistency. The behavior of the proposed estimator and its smoothed versions is eventually illustrated by simulated examples and a case study involving alcohol blood level in DUI cases.

Key words: Size-biased data; Wavelet density estimation; Daubechies-Lagarias algorithm.

1 Introduction

Size-biased data arise when the likelihood for an observation to appear in a sample depends on its magnitude. The density associated with a size-biased random variable Y, f^Y, is related to the underlying true density f^X by

$$f^Y(y) = \frac{g(y)f^X(y)}{\mu},$$

where g is the so-called weighting or biasing function and μ is defined as the expected value of $g(X)$, $\mu = \mathbb{E}[g(X)] < \infty$. In many cases a linear g is recommended, but in general the form of g should be studied via additional experiments.

*Corresponding author: Pepa Ramírez (*e-mail:* Pepa.Ramirezcobo@gipsa-lab.inpg.fr)

Several examples of this situation can be found in the literature. For instance, in the founding paper by Cox (1969) it is shown that in renewal theory, inter-event data constitute a biased sample if they are obtained by sampling lifetimes in progress at a randomly chosen time instants. Pretend, as in Efromovich (1999) that the distribution of the concentration of alcohol in the blood of intoxicated drivers is of interest. Since the drunken driver has a larger chance of being arrested, the collected data are size-biased. In the field of medicine some phenomena are related to size-biased data. If, in the context of cancer diagnostics, the cases detected by screening are compared to the cases detected by symptoms, it will appear that screened cases have a survival advantage. This is likely due to the larger proportion of patients with slow-growing tumors, detected by screening but not by symptoms. A different situation where size-biased data appear is in the estimation of length of stay in a healthcare facility. People who have been in the facility longer (i.e. have longer length of stay) are more likely to be sampled than people with a shorter one. This is analogous to the classical example of the "waiting-time" or "inspection paradox" (see Feller (1971) and Ross (1983)). The estimation of length of stay is not in itself usually of interest, as most healthcare facilities have records on patients' length of stay, but the bias towards longer lengths of stay is of importance when point-in-time surveys are used to estimate average hospital charges, for example. In this instance, since longer lengths of stay generally correspond to higher charges, the estimation of the distribution of healthcare expense will be biased due to the higher probability of sampling patients with longer lengths of stays.

Several approaches to estimate the unobserved density f^X can be found in the literature. Given the observed sample $y_1, \ldots, y_n$, the so-called naive estimator can be obtained by dividing an estimate of f^Y by a biasing function g. However, according to Efromovich (1999), this method does not lead to an optimal estimation according to asymptotic theory. Also, difficulties may arise for a set of points y where $g(y)$ is relatively small. In Vardi (1982), the nonparametric maximum likelihood estimate for f^X is derived. Jones (1991) discusses the mean squared error properties of a new kernel density estimation for size-biased data. In El Barmi and Simonoff (2000), a simple transformation-based approach to estimating the density f^X is examined. In Efromovich (2004a, 2004b), f^X is estimated via the Fourier methodology where in addition, asymptotic results on sharp minimax density estimators are given. An EM algorithm to approximate the nonparametric maximum likelihood estimate is derived in De la Uña and Rodíguez-Casal (2007).

The importance of wavelets in density estimation is well established. The local nature of wavelet functions promises superiority over projection estimators that use classical orthonormal bases, as Fourier, Hermite, etc. The

estimation procedures fall into the class of so-called projection estimators, or their non-linear modifications. The wavelet estimators are simple, well-localized in space/frequency, and share a variety of optimality properties.

It is interesting that some of the earliest contributions of wavelets in statistics were in density estimation. Doukhan (1988), and Doukhan and Léon (1990) first introduced linear wavelet density estimators and explored their mean-square errors. Antoniadis and Carmona (1991), Kerkyacharian and Picard (1992, 1993) and Walter (1992, 1994) considered the linear wavelet estimators in Sobolev and Besov spaces, while Donoho et al. (1995, 1996), Delyon and Juditsky (1993), among others, explored non-linear estimators and their minimax optimality in Besov spaces. Recently, Chacón and Rodríguez-Casal (2005) prove that under mild conditions on the family of wavelets, the wavelet density estimators are universally consistent in the $\mathcal{L}_1$ sense. For a critical discussion of the advantages and disadvantages of wavelet in density estimation, see Walter and Ghorai (1992). To the best of our knowledge the unpublished manuscript by Nikolaidou and Sapatinas (2006) constitutes the single reference where the wavelet approach is undertaken, in the case of weighted (or size-) biased data. For a detailed discussion of the performance of wavelets density estimators and some of their advantages over traditional methods, see Vidakovic (1999).

In this work we develop a wavelet-based estimator of f^X, in the context of stratified size-biased data, which appear when the biasing process is not homogeneous, and illustrate our methodology on both simulated and real data sets. The paper is organized as follows. In Section 2 a wavelet-based estimator of the underlying density f^X is derived, when only a single biasing function is considered. This estimator is extended to the case where the sample is stratified in Section 3. In addition, the consistency of the estimator, in the mean squared integrated error sense is proven. Section 4 illustrates our methodology for simulated data, where the Daubechies-Lagarias algorithm, described in the Appendix, is implemented. Section 5 deals with a real-life application of the proposed methodology: estimation of the density f^X when a sample of alcohol levels in fatal driving accidents is stratified. Finally, in Section 6 we provide conclusions and delineate possible directions for future research.

2 Wavelet density estimator

Assume that observations $y_1, \ldots, y_n$ are modeled by a random variable Y supported on $[0, 1]$ whose density function is given by

$$f^Y(y) = \frac{g(y)f^X(y)}{\mu},$$

where $g(y)$ is a positive function known as biasing function and $f^X(y) \in \mathcal{L}_2([0, 1])$ is a probability density of interest. The density $f^Y(y)$ is normalized by

$$\mu = \int_0^1 g(y)f^X(y)dy.$$

As $f^X(y)$ is unknown, the parameter μ is also unknown. The problem is indirect since one observes Y and wants to estimate the density of an unobserved X.

Since $f^X(y)$ is in $\mathcal{L}_2([0, 1])$, it allows wavelet representation

$$f(y) = \sum_{k\in\mathbb{Z}} c_{J,k}\phi_{J,k}(y) + \sum_{j=J}^{\infty}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(y),$$

where the wavelet basis functions

$$\{\phi_{J,k}(y) = 2^{J/2}\phi(2^J y-k),\ k \in \mathbb{Z}\}, \quad \{\psi_{j,k}(y) = 2^{j/2}\psi(2^j y-k),\ j \geq J, k \in \mathbb{Z}\}$$

are generated from the scaling function $\phi(x)$, and

$$c_{J,k} = \int f(y)\phi_{J,k}(y)dy, \quad d_{j,k} = \int f(y)\psi_{j,k}(y)dy.$$

Here, J indicates the coarsest scale or lowest resolution of analysis, and a larger j corresponds to higher resolutions (for a detailed wavelets theory, see Vidakovic 1999).

A wavelet-based estimator of f^X can be defined in terms of the projection of f^X on V_J, the multiresolution space spanned by the functions $\phi_{J,k}(y)$, as

$$f_J^X(y) = \sum_{k\in\mathbb{Z}} c_{J,k}\phi_{J,k}(y).$$

We first estimate the coefficients $c_{J,k}$ via the expectation

$$\mathbb{E}\left[\frac{\mu\phi_{J,k}(Y)}{g(Y)}\right],$$

with empirical counterpart

$$\hat{c}_{J,k} = \frac{\mu}{n} \sum_{i=1}^{n} \frac{\phi_{J,k}(y_i)}{g(y_i)}, \tag{1}$$

based on n i.i.d realizations $y_1, \ldots, y_n$ of Y. As the parameter μ is not known, it is estimated by

$$\hat{\mu} := \frac{1}{n^{-1} \sum_{i=1}^{n} g^{-1}(y_i)}.$$

It can be shown that the estimate $1/\hat{\mu}$ is unbiased for $1/\mu$:

$$\mathbb{E}(1/\hat{\mu}) = \mathbb{E}(g^{-1}(Y)) = \mu^{-1} \int_0^1 g(y) f^X(y) g^{-1}(y) dy = 1/\mu.$$

Then, a linear (or projection) estimator of f^X can be defined using (1) as

$$\widehat{f}_J^X(y) = \sum_{k \in \mathbb{Z}} \hat{c}_{J,k} \phi_{J,k}(y). \tag{2}$$

The estimator in (2) gains in performance if regularized. Regularization could be achieved by wavelet shrinkage. For given levels J_0 and J such that $J_0 < J$, scaling and wavelet coefficients $\hat{c}_{J_0,k}$, and $\hat{d}_{j,k}$, for $J_0 \leq j < J$ can be obtained by utilizing fast Mallat's cascade algorithm that starts with $\hat{c}_{J,k}$. Thus, the original estimator (2) can be represented as

$$\widehat{f}_J^X(y) = \sum_{k \in \mathbb{Z}} \hat{c}_{J_0,k} \phi_{J_0,k}(y) + \sum_{J_0 \leq j < J} \sum_{k \in \mathbb{Z}} \hat{d}_{j,k} \psi_{j,k}(y). \tag{3}$$

To regularize $\widehat{f}_J^X$, wavelet shrinkage to detail coefficients, $\hat{d}_{j,k}$, can be applied. Although our formulation is not a regression problem, we can think of noisy linear projection estimator as a smoothing problem and apply standard wavelet-based regression techniques. The most common thresholding policies is *hard* thresholding, for which the analytic expression is

$$\delta^h(\hat{d}_{j,k}, \lambda) = \hat{d}_{j,k} \ \mathbf{1}(|\hat{d}_{j,k}| > \lambda),$$

where the threshold $\lambda \geq 0$ depends on the distributional properties of the wavelet coefficients. It is our experience that the optimal minimax thresholds proposed by Donoho et al. (1996), and Delyon and Juditsky (1993), do not work well for small samples. In that context, the universal threshold is standardly chosen:

$$\lambda = \sqrt{2 \log n} \ \sigma,$$

where σ can be estimated from the wavelet decomposition by for example,

$$\hat{\sigma} = 1/0.6745 \cdot \text{MAD}\left(d_{(J-1),\cdot}\right),$$

where MAD stands for median absolute deviation from the median.

The shrunk coefficients will be denoted by $d^{\star}_{j,k}$, for $J_0 \leq j < J$. Thus, the new "smoothed" estimator of f^X is

$$\widetilde{f}^X_J(y) = \sum_{k\in\mathbb{Z}} \hat{c}_{J_0,k}\phi_{J_0,k}(y) + \sum_{J_0\leq j<J}\sum_{k\in\mathbb{Z}} d^{\star}_{j,k}\psi_{j,k}(y). \tag{4}$$

If the underlying density f^X is smoother than the decomposing wavelet or the sample size is not large, the wavelet shrinkage estimators may contain peaks and artifacts. An alternative in these cases may be another smoothing method, such us the local linear regression smoother (Fan, 1992), which is known to possess nice sampling properties and high minimax efficiency (Fan, 1993). In next section we will extend the projection estimate in (2) to the stratified case and prove its consistency.

3 Extension to stratified sized biased data

In the previous we assumed that the biasing function is common for the population, that is, all sampled observations are biased in the same way. It might be the case that the biasing function is controlled by a covariate thus introducing inhomogeneity in the biasing process. For instance, if in the case of hospital-length-stay we stratify the sample according to gender, age, or type of hospital, the biasing functions may differ.

We extend the estimate (2) of $f^X(y)$, to the case of a stratified sample with a stratum-dependent biasing function. We will assume M differently biased i.i.d random variables (or *strata*) Y_m, for $m = 1, \ldots, M$, with corresponding biasing functions $g_m(y)$, and common underlying density f^X. If we observe the stratified sample,

$$y_{11}, \ldots, y_{1n_1}; y_{21}, \ldots, y_{2n_2}; \ldots; y_{M1}, \ldots, y_{Mn_M},$$

then the observed densities are given by

$$f^Y_m(y) = \frac{g_m(y)f^X(y)}{\mu_m}, \quad \text{for } m = 1, \ldots, M.$$

If $N = n_1 + \ldots + n_M$ is the total sample size, then an estimator of f^X based on all observations, can be defined as

$$\widehat{f}^X_J(y) = \sum_{m=1}^{M} \alpha_m \widehat{f}^X_{m,J(m)}(y), \tag{5}$$

where $\alpha_m = n_m/N$, $J(m)$ is the projection level in the mth stratum, and J is defined as

$$J := \min\{J(1), \ldots, J(M)\}. \tag{6}$$

In (5), $\hat{f}^X_{m,J(m)}$ represents the projection estimate of f^X defined in (2), based on the mth stratum, depending on

$$\hat{c}^m_{J(m),k} = \frac{\hat{\mu}_m}{n_m} \sum_{i=1}^{n_m} \frac{\phi_{J(m),k}(Y_{mi})}{g_m(Y_{mi})},$$

the estimate of $c_{J(m),k}$ within stratum m.

The estimator $\widehat{f}^X_J(y)$ in (5) is consistent in the mean integrated squared error or *MISE* sense. The following result holds.

Theorem 3.1. *Assume that the density for X, $f^X(y)$ belongs to $\mathcal{L}_2([0,1])$ and $\mu_m f^X(y)/g_m(y) < B$, for all $y \in (0,1)$ and $m = 1, \ldots, M$. Let J be as in (6) and $k(n_m)$ be the number of coefficients $\hat{c}^m_{J(m),k}$ in (5). If*

$$k(n_m) \to \infty \quad \textit{and} \quad \frac{k(n_m)}{n_m} \to 0 \quad \textit{as} \quad n_m \to \infty,$$

for all m then, the wavelet-based estimator in (5) is $\mathcal{L}_2$-consistent, i.e.,

$$\lim_{J\to\infty} \mathbb{E}\left\{ \int_0^1 \left(f^X(y) - \widehat{f}^X_J(y) \right)^2 dy \right\} = 0.$$

Remark. If $f^X(y)$ is compactly supported, then $k(n_m)$ is finite, for a fixed value of J. Indeed, as $y \in [0,1]$, then $k(n_m) = 2^J$, for all m. The proof of Theorem 3.1 is given in the Appendix.

Again, as in the previous section, the estimators based on projection

$$\widehat{f}^X_{m,J(m)}(y) = \sum_{k\in\mathbb{Z}} \hat{c}^m_{J_0(m),k} \phi_{J_0(m),k}(y) + \sum_{J_0(m)\le j<J(m)} \sum_{k\in\mathbb{Z}} \hat{d}^m_{j,k} \psi_{j,k}(y),$$

can be regularized to gain in performance. By applying wavelet shrinkage to $\hat{d}^m_{j,k}$, we obtain the smoothed versions

$$\widetilde{f}^X_{m,J(m)}(y) = \sum_{k\in\mathbb{Z}} \hat{c}^m_{J_0(m),k} \phi_{J_0(m),k}(y) + \sum_{J_0(m)\le j<J(m)} \sum_{k\in\mathbb{Z}} d^{m\star}_{j,k} \psi_{j,k}(y).$$

In our analysis, we used the universal thresholding rule as described in the previous section. Finally, the smoothed estimator of f^X, given a stratified sample is

$$\widetilde{f}_J^X(y) = \sum_{m=1}^{M} \alpha_m \widetilde{f}_{m,J(m)}^X(y), \tag{7}$$

where J was defined in (6). The regularized estimator is consistent as well, as next result states.

Theorem 3.2. *Let J and $k(n_m)$ be as in Theorem 3.1 and let*

$$J_0 = \min\{J_0(1), \ldots, J_0(M)\}$$

be the minimum multiresolution level. The regularized wavelet-based estimator $\widetilde{f}_J^X(y)$ in (7) is $\mathcal{L}_2$-consistent if for all $m = 1, \ldots, M$,

$$J_0 \to \infty \quad \textit{and} \quad \frac{k(n_m)}{n_m} \to 0, \quad \textit{as} \quad n_m \to \infty.$$

The proof of Theorem 3.2 is given in the Appendix.

Denote by $\hat{f}_s^X(y)$ the smoothed version of the projection estimate (5) after applying local linear regression. Under the assumptions of Theorem 3.1 and smoothness constraint on the wavelet base, $\hat{f}_s^X(y)$ is a $\mathcal{L}_2$-consistent estimate of $f^X(y)$. This follows from Fan (1992) page 1000, with the assumption that wavelet has at least three vanishing moments.

4 Simulation results

In this section we illustrate the results from the previous sections on selected simulated data. We also consider the problem of estimating strata dependent biasing functions if the biasing for one single stratum is known. To implement the estimator in (7) one needs to compute the value of the scaling function at an arbitrary point y. Most of wavelet functions do not possess an analytic expression in finite form, so this evaluation is not a straightforward task.

To calculate values of scaling functions we will use the Daubechies-Lagarias algorithm, explained in detail in the Appendix. One can always tabulate the scaling function and evaluate its value at x using scaled and shifted table values, however Daubechies-Lagarias algorithm provides an elegant and economic way to evaluate the scaling function.

Next five examples illuminate various aspects of our methodology.

Example 1. In this example we assume a uniform underlying density, $X \sim U(0,1)$, and a single stratum with the biasing function is $g(y) = e^y$. Then,

$$f^Y(y) \propto e^y, \quad y \in [0,1].$$

We generated $n = 600$ data points from the observed distribution, f^Y. The index of the highest resolution space was chosen to be $J = 7$ so that a total of 128 (2^7) scaling coefficients are generated by the Debauchies-Lagarias algorithm. Coiflet basis with six-tap filter was used. At the top of Figure 1 the projection estimate of true underlying density based on the observed sample, is shown. At the bottom panel the thresholded wavelet estimator, which is almost undistinguishable from the theoretical density, is depicted.

FIGURE 1 ABOUT HERE

Example 2. In this example we illustrate how the estimate (5) performs in the case of a stratified sample. We assume an underlying $X \sim Beta(2,2)$ density. We consider three strata and observe a sample $y_{1,1}, \ldots, y_{1,500} \sim f_1^Y$, $y_{2,1}, \ldots, y_{2,500} \sim f_2^Y$, and $y_{3,1}, \ldots, y_{3,500} \sim f_3^Y$, where

$$\begin{aligned} f_1^Y(y) &= y^2(1-y), \\ f_2^Y(y) &= y^{5/2}(1-y), \\ f_3^Y(y) &= y^3(1-y), \end{aligned}$$

for $y \in [0,1]$, that is, $Y_1 \sim Beta(3,2)$, $Y_2 \sim Beta(7/2,2)$, and $Y_3 \sim Beta(4,2)$. The biasing functions are given by $g_1(y) = y$, $g_2(y) = y^{3/2}$, and $g_3(y) = y^2$. At the top of Figure 2, the projection estimate (5) is depicted. The picture on the bottom left shows the thresholded estimate and that on the bottom right shows the local linear smoothed estimate. Notice that in this case the local linear smoother performs better than the wavelet shrinkage, due to smooth nature of the underlying density.

FIGURE 2 ABOUT HERE

Example 3. Here we emphasize the difference in behavior between the thresholded and local linear smoothed estimates, when the underlying density is not smooth. Let us consider a piece-wise linear density function $f^X(y)$:

$$f^X(y) = \begin{cases} 8y, & y \in [0, \frac{1}{4}) \\ 4(1-2y), & y \in [\frac{1}{4}, \frac{1}{2}) \\ 8y-4, & y \in [\frac{1}{2}, \frac{3}{4}) \\ 8-8y, & y \in [\frac{3}{4}, 1] \end{cases}$$

We assume a size biasing function $g(y) = y$, and consider a biased sample of size 1000. In the top panel of Figure 3, both the underlying (solid line) and observed (dashdotted line) densities are shown. In addition, the projection estimate is depicted in dotted line. Because of the shape of size biasing function, more samples are observed in the right part of the density, and thus this part is better estimated. In the bottom left panel the thresholded estimate is shown (Daubechies' filter with parameter 6 was applied), while the panel at the bottom right depicts the local linear smoothed version. With the exception of the estimated density for values of y less than $1/4$, the thresholded estimate performs better in this case, in the sense the peaks of the density are better captured.

FIGURE 3 ABOUT HERE

Example 4. In this example we show how the naive estimator mentioned in Section 1 compares to the proposed wavelet-based estimator. Given a biased sample $y_1, \ldots, y_n$ the naive estimator is computed by first estimating the observed density $f^Y(y)$, and then dividing it by the biasing function $g(y)$, that is

$$\hat{f}^X_{naive}(y) = \frac{\mu \hat{f}^Y(y)}{g(y)}.$$

According to Efromovich (2004b) this naive estimator is rate inadmissible whenever the biasing function is not as smooth as the underlying density f^X. Specifically, we consider a smooth f^X ($Beta(2,2)$), and a discontinuous biasing function,

$$g(y) = \begin{cases} 1, & y \in [0, 0.4) \\ y, & y \in [0.4, 1] \end{cases}$$

A total of 520 biased observations were generated, and the local linear smoothed projection estimate was computed. This is shown at the top of Figure 4. The naive estimator was computed as well; it is shown at the bottom of Figure 4. The unsatisfactory performance of the naive estimator around $x = 0.4$ is a consequence of the discontinuity of $g(y)$, which however, does not affect the wavelet-based estimate.

In the simulations with both the density and biasing functions infinitely differentiable, similar performance results were obtained by naive and smoothed estimates. Since the smoothness of f^X is typically unknown, it is advisable to avoid the naive estimator.

FIGURE 4 ABOUT HERE

Example 5. In our last simulational example, we consider the problem of estimating strata dependent biasing functions if the biasing for just one stratum is known. Let us pretend that in Example 2, only the first biasing function, $g_1(y) = y$, was known. In order to estimate $g_2(y)$ and $g_3(y)$ up to a constant, we first estimate $f^X(y)$, given $g_1(y)$, and $f_1^Y(y)$ by (4). Then, the estimates of $g_m(y)$, for $m = 2, 3$ are proportional to

$$\frac{f_m^Y(y)}{\widehat{f^X}(y)}.$$

Figure 5 depicts the approximated and true functions $g_2(y)$, and $g_3(y)$, up to a constant, in the log-scale. With usual wavelet transforms, some numerical problems were encountered resulting in peaks. In order to avoid them, we recommend boundary-corrected forward (and inverse) wavelet transform.

FIGURE 5 ABOUT HERE

5 An application to automobile accident data

In this section we apply our methodology to a real-life problem. We consider the alcohol levels of male drivers in fatal driving accidents during 1975. This constitutes a typical example of size-biased data because drunk drivers in fatal accidents (resulting in death of a person) have a larger chance of being "sampled", that is, getting their blood alcohol level measured. The sample is divided in two strata: drivers under thirty and over thirty years old. The motivation is that the same alcohol content may have different influence on young and older drivers.

The data set was obtained from the National Highway Traffic Safety Administration Department of Transportation, in the United States, and available in `www.nhtsa.dot.gov`. The data set was part of the The Fatality Analysis Reporting System (FARS) formally referred to as the Fatal Accident Reporting System. It is a collection of files documenting all qualifying fatal crashes since 1975 that occurred within the 50 states, the District of Columbia, and Puerto Rico. To be included in the census of crashes, a crash had to involve a motor vehicle traveling on a trafficway customarily open to the public, and must result in the death of a person (occupant of a vehicle or a nonmotorist) within 30 days of the crash. The FARS analytic reference

guide has several sections including the accident, vehicle and person files. Our variables of interest were, within the Person File, the age and the result of alcohol test (named BAC, from Blood Alcohol Concentration). The BAC is a continuos variable expressed in grams/100ml.

We considered two stratified samples of total size 379, belonging to BACs of male drivers less than or equal to 30 years old and more than 30 years old. Upper panel in Figure 5 shows the histogram of BACs for the younger drivers. Lower values of BACs are observed in comparison to the the bottom panel, showing the BACs for older drivers. The reason for this could be that the "young" sample contains significant proportion of inexperienced drivers, so that experience plays the role in a crash and typical sample shows smaller BACs values. However, in the case of older drivers, the accident can be mainly attributed to alcohol intoxication.

FIGURE 5 ABOUT HERE

The difference observed in the strata histograms, concerning to the experience factor, leads us to choose the following biasing functions:

$$\begin{aligned} g_y(t) &:= \sqrt{t} \\ g_o(t) &:= t, \end{aligned}$$

where "y" and "o" stand for young and old strata, respectively. In this way, smaller observed BAC values have associated similar biasing functions for both strata, but larger BAC values will be characterized by a larger biasing function, for the older drivers case. This elicitation of biasing functions is similar in spirit to Bayesian prior elicitation: the choice is subjective and reflects influence of historic data, nature of the phenomenon, and common sense.

Given the stratified sample, we applied formula (7) to obtain an estimate of the underlying density of the BAC values. Figure 6 depicts the estimate of the underlying density, under the previously defined biasing functions. A bimodal density estimate is obtained. The smaller mode could be interpreted as representing accidents in which the alcohol may not be the cause; however, the larger mode would be related to accidents caused by alcohol intoxication.

FIGURE 6 ABOUT HERE

6 Concluding remarks

In this paper we have developed a wavelet density estimator, in the context of stratified sized-biased data. Under mild conditions on the underlying and biasing functions, the defined estimator is consistent in the mean integrated squared error sense. We have shown how the Daubechies-Lagarias algorithm can be used to evaluate the scaling function at any point; this makes our estimator easily computable for numerical applications. We have illustrated the performance of the estimator with several simulated-data examples and eventually applied the developed methodology to the estimation of a real underlying density function, representing the alcohol concentration in male drivers involved in fatal accident in the USA.

In our future work we plan to implement other shrinkage and smoothing policies, and compare the performance of the estimators under the proposed shrinkage rules. The rate-optimality of the proposed wavelet estimator is under current investigation. In the spirit of a reproducible research all codes written in Matlab and used in this work are available at

`http://www2.isye.gatech.edu/~brani/wavelet.html`.

7 Appendix

A. Proofs of Consistency

Proof of Theorem 3.1.
First, let us point out that consistency of the estimator $\widehat{f}^X_{m,J}(y)$, for all strata $m = 1, \ldots, M$, implies that of the overall estimator $\widehat{f}^X_J(y)$. Because of additivity, the properties of $\widehat{f}^X_J = \sum_{m=1}^M \alpha_m \widehat{f}^X_{m,J}(y)$ follow immediately from those of $\widehat{f}^X_{m,J}$, for $m = 1, 2, \ldots, M$. Thus,

$$MISE\left(\widehat{f}^X_J(y)\right) = \sum_{m=1}^{M} \alpha_m^2 MISE\left(\widehat{f}^X_{m,J}(y)\right).$$

Since $f(y) \in \mathcal{L}_2$ and *Parseval's identity* is applicable, the coefficients $c_{J,k}$ and $d_{j,k}$ belong to ℓ_2,

$$\sum_{k\in\mathbb{Z}} c^2_{J,k} + \sum_{j=J}^{\infty}\sum_{k\in\mathbb{Z}} d^2_{j,k} = \int f(y)^2 dy = \|f\|^2 < \infty. \tag{8}$$

To simplify notation we will consider the proof of consistency for stratum s such that $J(s) = J$, which was defined in (6), and thus we drop index of

the stratum s. The proof for the rest of strata is analogous, since if $J \to \infty$, $J(m) \to \infty$, for all m. We first show that the $\hat{c}^s_{J,k}$ is an unbiased estimate of $c_{J,k}$.

$$\begin{aligned}
\mathbb{E}(\hat{c}^s_{J,k}) &= \mu_s \mathbb{E}\left[\frac{\phi_{J,k}(Y)}{g_s(Y)}\right] \\
&= \mu_s \int_0^1 \frac{f^Y_s(y)\phi_{J,k}}{g_s(y)} dy \\
&= \int_0^1 f^X(y)\phi_{J,k}(y)dy = c_{J,k}.
\end{aligned}$$

Next we find an upper bound on the variance of $\hat{c}^s_{J,k}$.

$$\begin{aligned}
\mathbf{Var}(\hat{c}^s_{J,k}) &= \frac{1}{n_s}\mathbf{Var}\left(\frac{\mu_s\phi_{J,k}(Y)}{g_s(Y)}\right) \\
&= \frac{1}{n_s}\left(\int_0^1 \frac{\mu_s^2\phi^2_{J,k}(y)}{g_s^2(y)} f^Y(y)dy - c^2_{J,k}\right) \\
&= \frac{1}{n_s}\left(\int_0^1 \frac{\mu_s\phi^2_{J,k}(y)}{g_s(y)} f^X(y)dy - c^2_{J,k}\right) \\
&\leq \frac{1}{n_s}\left(B\int_0^1 \phi^2_{J,k}(y)dy - c^2_{J,k}\right) \leq \frac{1}{n_s}(B - c^2_{J,k}) \\
&\leq \frac{B}{n_s},
\end{aligned}$$

where constant B is as in formulation of Theorem 3.1.

From the orthogonality of $\phi_{J,k}$ and $\psi_{j,k}$,

$$\int \phi_{J,k}(y)\psi_{j,k}(y)dy = 0, \quad \text{and} \quad \int \psi_{j,k}(y)\psi_{j',k}(y) = 0, \quad \text{for } j' \neq j,$$

and the Parseval's identity (8), it follows that

$$\begin{aligned}
&\int_0^1 \left(f^X(y) - \widehat{f}^X_{s,J}(y)\right)^2 dy \\
= &\int_0^1 \left(f^X(y)\right)^2 dy + \int_0^1 \left(\sum_{k\in\mathbb{Z}} \hat{c}^s_{J,k}\phi_{J,k}(y)\right)^2 dy - 2\int_0^1 f^X(y)\sum_{k\in\mathbb{Z}} \hat{c}^s_{J,k}\phi_{J,k}(y)dy \\
= &\sum_{k\in\mathbb{Z}} c^2_{J,k} + \sum_{j\geq J}\sum_{k\in\mathbb{Z}} d^2_{j,k} + \sum_{k\in\mathbb{Z}}\left(\hat{c}^s_{J,k}\right)^2 - 2\sum_{k\in\mathbb{Z}} c_{J,k}\hat{c}^s_{J,k} \\
= &\sum_{k\in\mathbb{Z}}(\hat{c}^s_{J,k} - c_{J,k})^2 + \sum_{j\geq J}\sum_{k\in\mathbb{Z}} d^2_{j,k}.
\end{aligned}$$

It may be concluded that the expected $\mathcal{L}_2$-error is bounded since

$$\begin{aligned}
&\mathbb{E}\left\{\int_0^1 \left(f^X(y) - \widehat{f}^X_{s,J}(y)\right)^2 dy\right\} \\
&= \sum_k \mathbb{E}\left\{\left(c_{J,k} - \hat{c}^s_{J,k}\right)^2\right\} + \sum_{j\geq J}\sum_{k\in\mathbb{Z}} d^2_{j,k} \\
&= \sum_{k=1}^{k(n_s)} \mathbf{Var}(\hat{c}^s_{J,k}) + \sum_{j=J}^{\infty}\sum_{k\in\mathbb{Z}} d^2_{j,k} \\
&\leq \frac{k(n_s)B}{n_s} + \sum_{j=J}^{\infty}\sum_{k\in\mathbb{Z}} d^2_{j,k}.
\end{aligned}$$

Because $f^X(y) \in \mathcal{L}_2([0,1])$, (8) implies

$$\lim_{J\to\infty} \sum_{j\geq J}\sum_{k\in\mathbb{Z}} d^2_{j,k} = 0.$$

By assumption $k(n_s)/n_s \to 0$, thus the expected $\mathcal{L}_2$-error converges to zero, or equivalently, the estimate $\hat{f}^X_{s,J}$ is $\mathcal{L}_2$-consistent, and so the overall estimator (5) is $\mathcal{L}_2$-consistent as well. □

Proof of Theorem 3.2.
As in the proof of Theorem 3.1, it is sufficient to show consistency for a stratum s, such that $J(s) = J$, as defined in (6). The values $J_0(1), \ldots, J_0(M)$ can be chosen so that

$$J_0(s) = J_0 = min\{J_0(1), \ldots, J_0(M)\}.$$

Notice that by definition, $J_0(m) < J(m)$, for all m, so that $J_0 < J$. The estimator corresponding to stratum s is thus given by

$$\widetilde{f}^X_{s,J}(y) = \sum_{k\in\mathbb{Z}} \hat{c}^s_{J_0,k}\phi_{J_0,k}(y) + \sum_{J_0\leq j<J}\sum_{k\in\mathbb{Z}} d^{s^\star}_{j,k}\psi_{j,k}(y).$$

By mimicking the proof of Theorem 3.1, and considering the wavelet expansion of $f^X(y)$,

$$f^X(y) = \sum_{k\in\mathbb{Z}} c_{J_0,k}\phi_{J_0,k}(y) + \sum_{j=J_0}^{\infty}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(y),$$

we obtain

$$\begin{aligned}
&\int_0^1 \left(f^X(y) - \widetilde{f}^X_{s,J}(y)\right)^2 dy \\
\leq\ & \frac{k(n_s)B}{n_s} + \sum_{j\geq J_0,k} d^2_{j,k} + \sum_{J_0\leq j<J}\sum_{k\in\mathbb{Z}} (d^{s^\star}_{j,k})^2 - 2\sum_{J_0\leq j<J}\sum_{k\in\mathbb{Z}} d_{j,k}d^{s^\star}_{j,k} \\
\leq\ & \frac{k(n_s)B}{n_s} + 2\sum_{J_0\leq j}\sum_{k\in\mathbb{Z}} d^2_{j,k},
\end{aligned}$$

which goes to 0 as $J_0 \to \infty$. □

B. Daubechies-Lagarias algorithm

The Daubechies-Lagarias algorithm provides a fast and economic calculation of the value of scaling and wavelets functions at some point. Except for the Haar wavelet, all compactly supported orthonormal families of wavelets, such us, the Daubechies, Symmlet, Coiflet, and others, have no a closed form. Based on Daubechies and Lagarias (1991), local pyramidal algorithm it is described as follows. Let ϕ be the scaling function of a compactly supported wavelet generating an orthonormal multiresolution analysis (Mallat, 1999). Suppose the support of ϕ is $[0, l]$. Let $x \in (0, 1)$ and define $dyad(x) = \{d_1, d_2, \ldots, d_n, \ldots\}$ as the set of 0-1 digits in dyadic representation of x, that is, $x = \sum_{j=1}^{\infty} d_j 2^{-j}$. Define also $dyad(x, n) = \{d_1, \ldots, d_n\}$.

Let $h = (h_0, \ldots, h_l)$ be the vector of wavelet filter coefficientes. Define the following matrices of size $l \times l$,

$$T_0 = \sqrt{2}(h_{2i-j-1})_{1\leq i,j\leq l}, \qquad T_1 = \sqrt{2}(h_{2i-j})_{1\leq i,j\leq l}.$$

Then, it can be shown that the element a_{11} of matrix A defined as ,

$$A = \lim_{n\to\infty} T_{d_1} \cdot T_{d_2} \cdot \ldots \cdot T_{d_n}$$

is equal to $a_{11} = \phi(x)$.

To compute the value of the mother wavelet ψ at some point x, needed in calculations of $\hat{d}_{j,k}$ in (3), the result in Pinheiro and Vidakovic (1997) can be used,

Theorem 7.1. *Let x be an arbitrary real number. Let the orthonormal MRA be given by wavelet filter coefficients $\{h_0, h_1, \ldots, h_N\}$ for $N = 2n - 1$. Let be the vector u of size N defined as*

$$u(x) = \{(-1)^{1-[2x]} h_{i+1-[2x]}, i = 0, \ldots, N-1\}.$$

If for some i the index $i + 1 - [2x]$ is negative or larger than N then the corresponding component of u is zero. Let the vector v be defined as

$$v(x,n) = \frac{1}{N}\mathbf{1}' \prod_{i \in dyad([2x],n)} T_i.$$

Then,

$$\psi(x) = \lim_{n\to\infty} u(x)'v(x,n).$$

Alternatively, the values of detail coefficients in the decomposition (3) can be obtain by Mallat's algorithm applied on $\hat{c}_{J,k}$ in (2).

Acknowledgements

Pepa Ramírez thanks to the support by the Comunidad de Madrid-UC3M (Project No. 2006/03565/001), the Ministerio de Educación y Ciencia (Project No. 2007/04438/001), and the Fulbright Commission for the grant to visit the Georgia Insitute of Technology. Brani Vidakovic's work was supported in part by grant NFS-ATM 0724524 at Georgia Insitute of Technology. The authors thank Editor, Associate Editor and an anonymous referee for their comments that substantially improved and clarified the manuscript.

References

[1] Antoniadis, A., Carmona, R., 1991. Multiresolution analysis and wavelets for density estimation. Technical report, University of California, Irvine.

[2] Chacón, J.E., Rodríguez-Casal, A., 2005. On the L_1-consistency of wavelet density estimates. The Canadian Journal of Statistics 33(4), 489–496.

[3] Cox, D. R., 1969. Some Sampling Problems in Technology. In New Developments in Survey Sampling, Johnson N. L. and Smith Jr. H. Editors, Wiley Interscience, New York, 506–527.

[4] Daubechies, I., Lagarias, J., 1991. Two-scale difference equations II. Local regularity, infinite products of matrices and fractals, SIAM Journal of Mathematical Analysis 23(4), 1031–1079.

[5] De la Uña-Álvarez, J., Rodríguez-Casal, A., 2007. Nonparametric estimation from length-biased data under competing risks. Computational Statistics & Data Analysis 51, 2653-2669.

[6] Delyon, B., Juditsky, A., 1993. Wavelet estimators, global error measures: Revisited. Technical Report 782, IRISA-INRIA, France.

[7] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D., 1995. Wavelet shrinkage: Asymptopia? (with discussion). Journal of the Royal Statistical Society, Series B 57(2), 301–369.

[8] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D., 1996. Density estimation by wavelet thresholding. The Annals of Statistics 23, 508–539.

[9] Doukhan, P., 1988. Formes de Toeplitz associées à une analyse multiéchelle. C. R. Acad. Sci. Paris. Sér. A. 306, 663–666.

[10] Doukhan, P., and Léon, J.R., 1990. Déviations quadratique d'estimateurs de densité par projections orthogonales. C. R. Acad. Sci. Paris Sér. I Math. 310, 425–430.

[11] Efromovich, S., 1999. Nonparametric Curve Estimation. Methods, Theory, and Applications. Springer Series in Statistics.

[12] Efromovich, S., 2004a. Distribution estimation for biased data. Journal of statistical planning and inference 124, 1–43.

[13] Efromovich, S., 2004b. Density estimation for biased data. The Annals of Statistics 32(3), 1137–1161.

[14] El Barmi, H., Simonoff, J.S., 2000. Transformation-based estimation for weighted distributions. Journal of Nonparametric Statistics 12(6), 861–878.

[15] Fan, J., 1992. Design-adaptative nonparametric regression. Journal of the American Statistical Association (87), 998–1004.

[16] Fan, J., 1993. Local linear regression smoothers and their minimax efficiencies. The Annals of Statistics 21(1), 196–216.

[17] Feller W., 1971. Introduction to probability theory and its applications, 2nd ed. John Wiley and Sons, New York NY.

[18] Jones, M.C., 1991. Kernel density estimation for length-biased data. Biometrika 78(3), 511–519.

[19] Kerkyacharian, G., Picard, D., 1992. Density estimation in Besov spaces. Statistics & Probability Letters 13, 15–24.

[20] Kerkyacharian, G., Picard, D., 1993. Density estimation by kernel and wavelets method: Optimality of Besov spaces. Statistics & Probability Letters 18, 327–336.

[21] Mallat, S.G., 1999. A Wavelet Tour of Signal Processing. 2nd Edition, San Diego: Academic Press.

[22] Nikolaidou, C., Sapatinas, T., 2006. Wavelet density estimation for weighted data. Technical report TR-01-2006.

[23] Pinheiro, A., Vidakovic, B., 1997. Estimating the square root of a density via compactly supported wavelets. Computational Statistics & Data Analysis 25, 399–415.

[24] Ross, S.M., 1983. Stochastic Processes. John Wiley and Sons, New York, NY.

[25] Vardi, Y., 1982. Nonparametric estimation in the presence of length bias. The Annals of Statistics 10(2), 616–620.

[26] Vidakovic, B., 1999. Statistical Modeling by Wavelets. Wiley, NY.

[27] Walter, G.G., 1992. Approximating of the delta function by wavelets. Journal of Approximation Theory 71, 329–343.

[28] Walter, G.G., 1994. Wavelets and other orthogonal systems with applications, CRC Press Inc,. Boca Raton, FL.

[29] Walter, G.G., Ghorai, J., 1992. Advantages and disadvantages of density estimation with wavelets. Computing Science and Statistics. Proceedings of the 24rd Symposium on the Interface, 234–243.

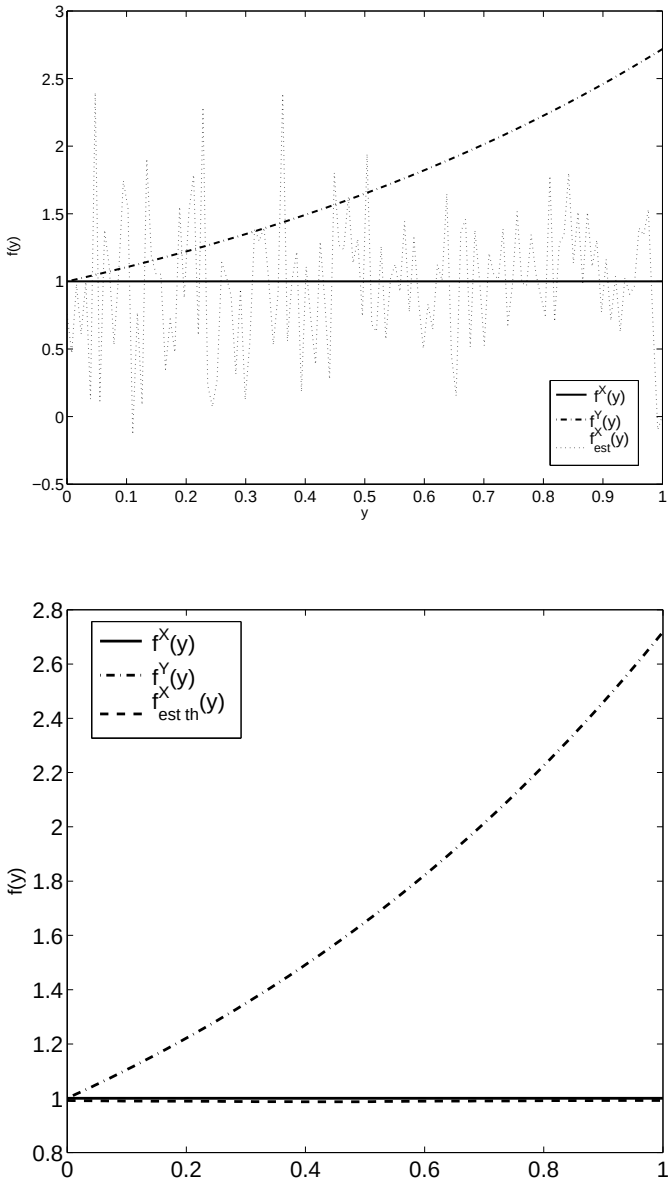

Figure 1: Underlying true (*solid line*), biased (*dashdotted line*), and estimated (*dotted lines*) densities. Top: Before thresholding. Bottom: After universal thresholding.

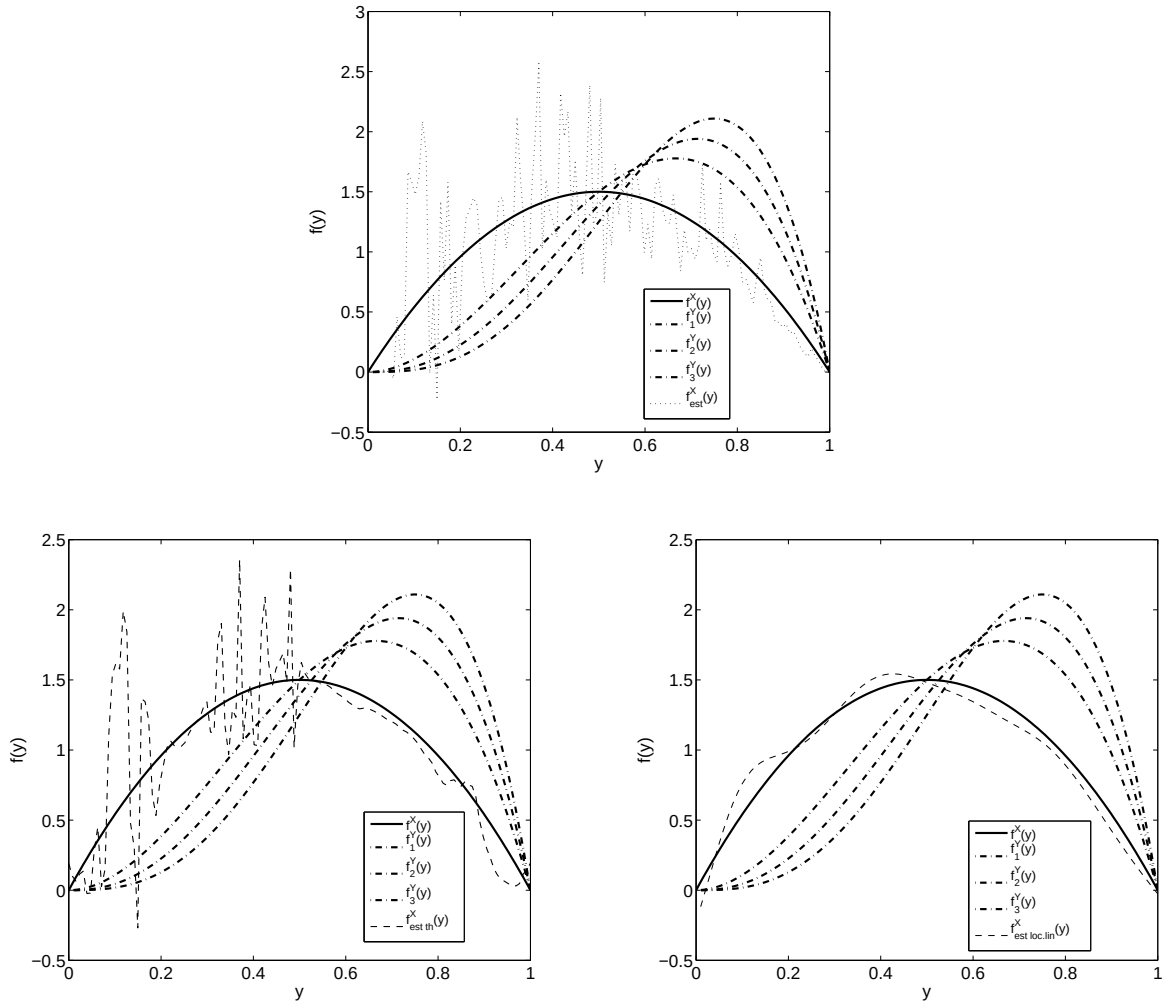

Figure 2: Underlying true (*solid line*) and biased (*dashdotted lines*) densities. Top: projection estimate (*dotted line*). Bottom left: estimate after universal thresholding (*dashed line*). Bottom right: estimate after local linear smoothing (*dashed line*).

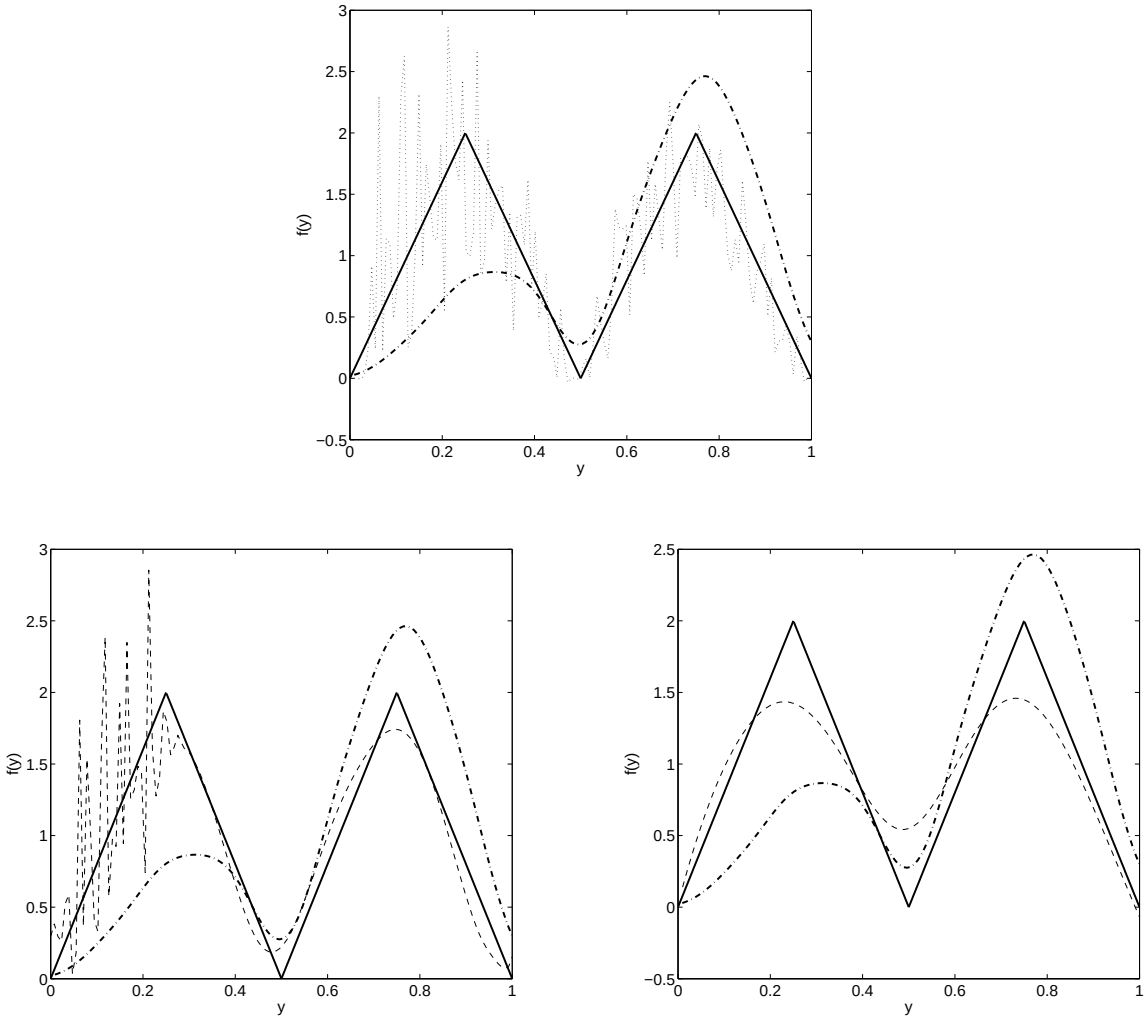

Figure 3: Underlying true (*solid line*) and observed biased (*dashdotted lines*) densities. Top: projection estimate (*dotted line*). Bottom left: estimate after universal thresholding (*dotted line*). Bottom right: estimate after local linear smoothing (*dotted line*).

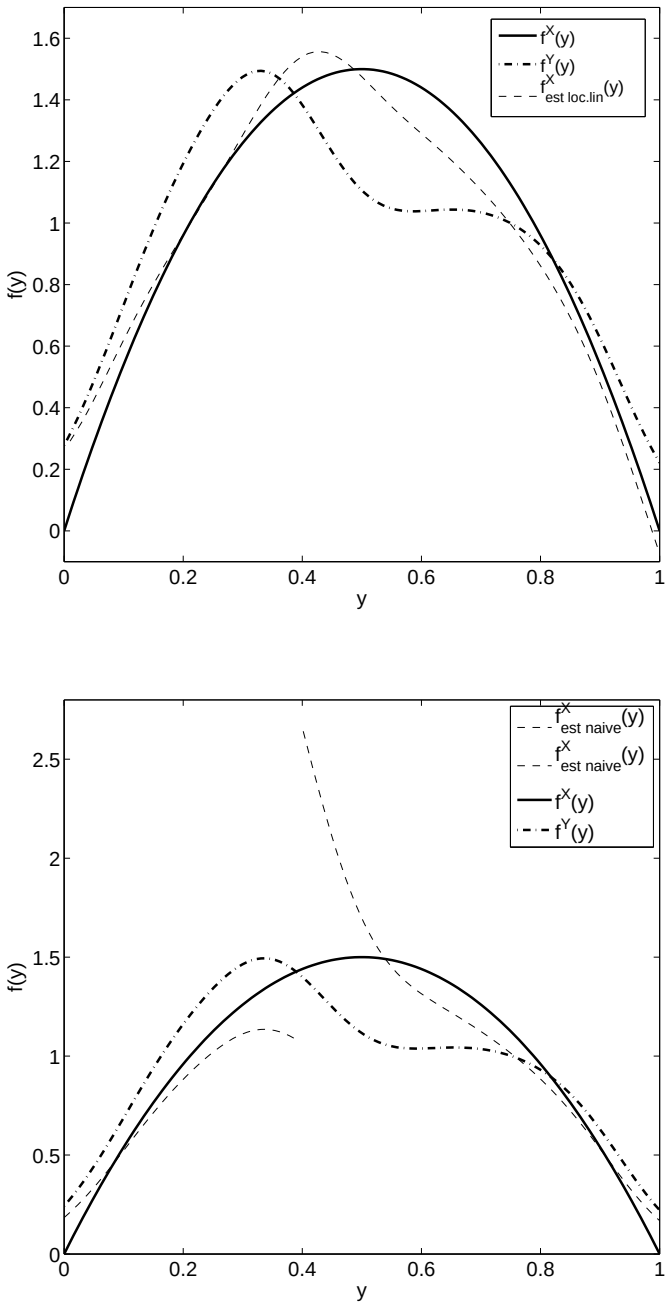

Figure 4: Underlying (*solid line*) and observed biased (*dashdotted lines*) densities. Top: projection estimate (*dashed line*) after local linear smoothing (*dotted line*). Bottom: naive estimate (*dashed line*)

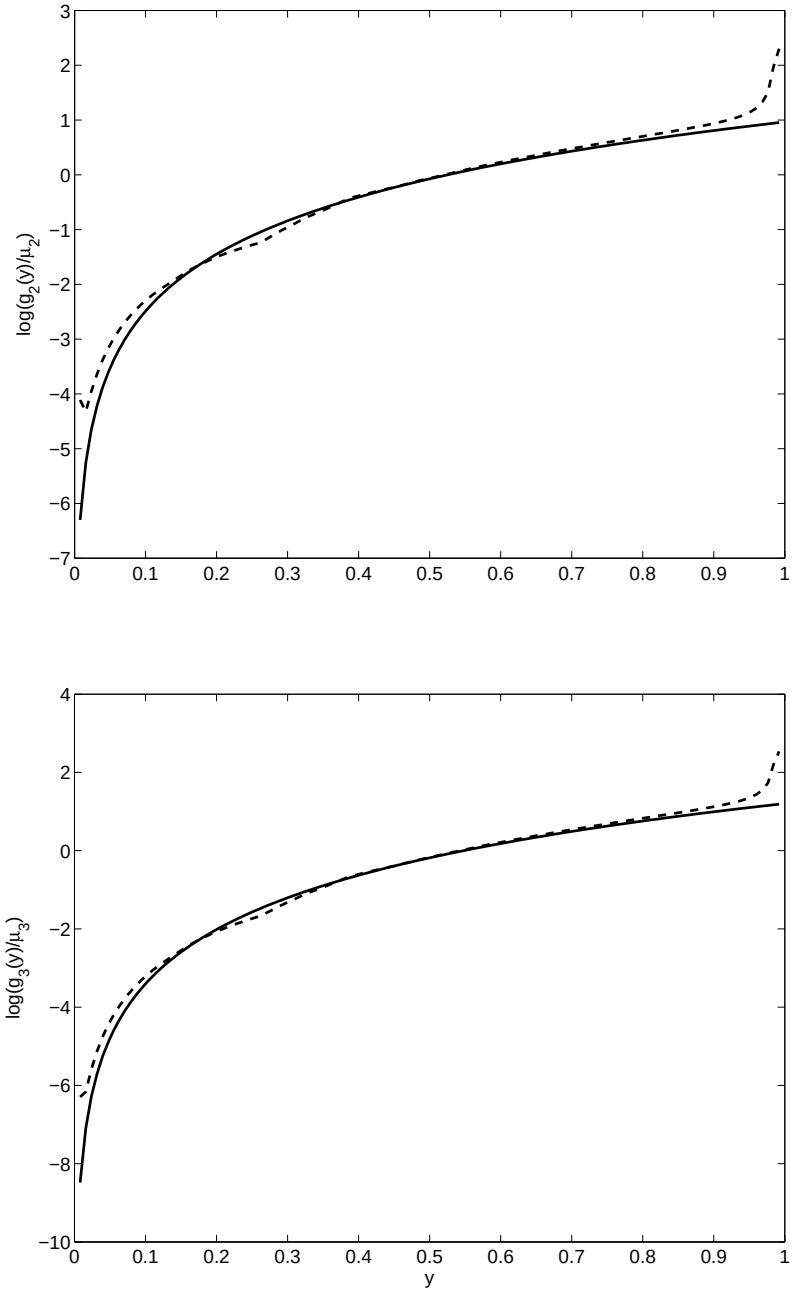

Figure 5: True (*solid line*) and estimated (*dotted line*) biasing functions $g_2(y)$ and $g_3(y)$ in log-scale.

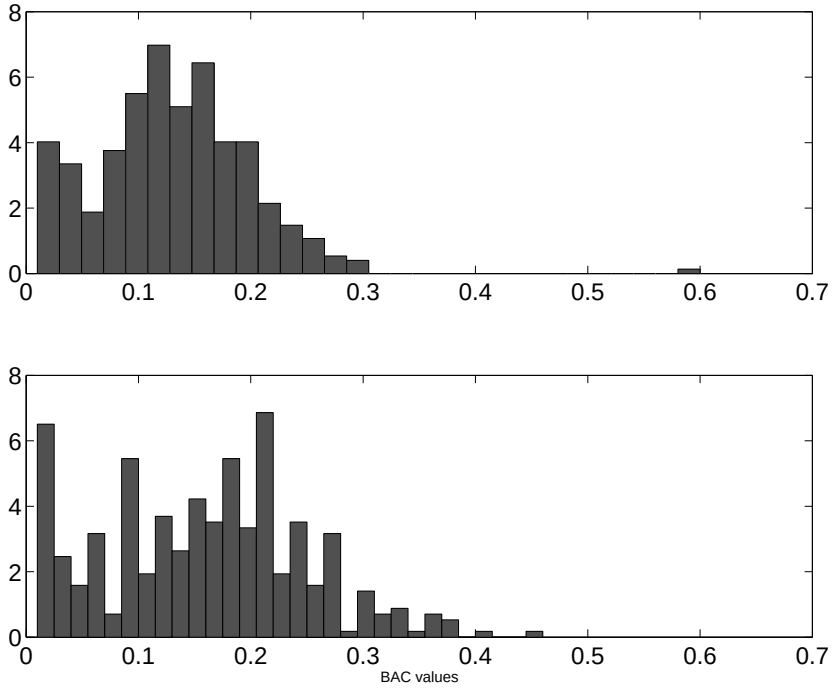

Figure 6: Histograms of BAC values for drivers of age less than or equal to 30 (top) and more than 30 (bottom).

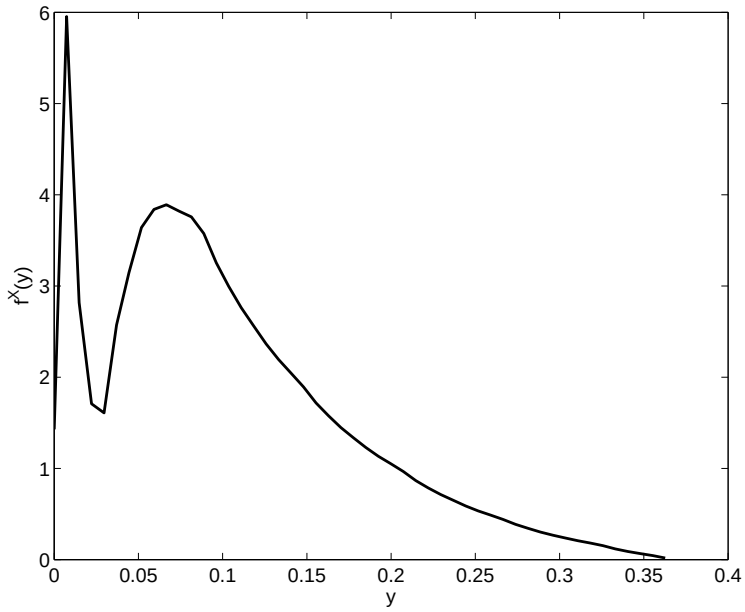

Figure 7: Estimated underlying density for blood alcohol concentration values in men in fatal driving accidents.