Density Estimation Under Constraints

Peter HALL and Brett PRESNELL

We suggest a general method for tackling problems of density estimation under constraints. It is, in effect, a particular form of the weighted bootstrap, in which resampling weights are chosen so as to minimize distance from the empirical or uniform bootstrap distribution subject to the constraints being satisfied. A number of constraints are treated as examples. They include conditions on moments, quantiles, and entropy, the latter as a device for imposing qualitative conditions such as those of unimodality or "interestingness." For example, without altering the data or the amount of smoothing, we may construct a density estimator that enjoys the same mean, median, and quartiles as the data. Different measures of distance give rise to slightly different results.

Key Words: Biased bootstrap; Cressie–Read distance; Curve estimation; Empirical likelihood; Entropy; Kernel methods; Mode; Smoothing; Weighted bootstrap.

1. INTRODUCTION

In some descriptive applications of density estimation it is necessary to construct estimators that enjoy the same basic properties—for example, location, scale, or distance from Normality—as the original, unsmoothed dataset. A case in point is estimation of location in financial data, such as household income or house prices, where the median is the most often-used measure of location and can be substantially less than the mean. Standard kernel methods produce a density estimator with the same mean as the data, but having different median. The techniques suggested in this article allow the kernel estimator to be modified so that it has the same mean and median as the unsmoothed data, or the same standard deviation and interquartile range, or indeed having all four quantities the same if desired.

In some other applications of density estimation the constraints are more qualitative in nature. An example is the condition that the density estimator be unimodal. The methods introduced in this article allow unimodality to be imposed through a quantitative condition on entropy. More generally, entropy is one of several types of constraint that may be imposed on the "interestingness" of a density estimator, so as to produce estimates that might have arisen had the sample been a little different. For a discussion of interestingness constraints in connection with exploratory projection pursuit see, for

Peter Hall is Professor, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia (Email: halpstat@fac.anu.edu.au). Brett Presnell is Associate Professor, Department of Statistics, University of Florida, Gainesville, FL 32611-8545 (Email: presnell@stat.ufl.edu).

Journal of Computational and Graphical Statistics, Volume 8, Number 2, Pages 259–277

example, Friedman, Stuetzle, and Schroeder (1984) and Friedman (1987). Importantly, our methods allow such exploratory analyses to be conducted without adjusting the level of smoothing.

Our approach to constrained density estimation also solves a problem involving the smoothed bootstrap, where it is sometimes desired to resample from a continuous distribution rather than the relatively rough empirical distribution. This problem is related to the wild bootstrap approach to resampling; see, for example, Härdle (1990, p. 247) and Mammen (1992, pp. 16–17). In such cases one would usually not wish the smoothing step to alter the main properties of the bootstrap; for example, properties of coverage or level accuracy of confidence or testing procedures. These properties are directly related to third and fourth cumulants of the sample, and so one would wish to construct an estimator of the sampling density that gave rise to exactly the same (standardized) skewness and kurtosis as the original data. Our methods permit this to be done relatively easily. Indeed, if a kernel density estimator is employed, then resampling from our constrained density estimate is equivalent to an explicitly defined weighted-bootstrap algorithm with a little noise added for smoothing.

The type of constraint that may be treated using our method is limited only by the computational ease with which it can be imposed. In this regard, kernel-type density estimators are particularly easy to work with, and constraints on linear combinations of moments or quantiles are relatively simple to impose, since then both the estimator and the constraints are linear in functions of individual data values. Indeed, when imposing r linear constraints on a kernel density estimator for a sample of size n, an optimization problem in $n - 1$ dimensions may be reduced to one in only $r + 1$ dimensions, and the weights may be computed by application of the Newton–Raphson algorithm in this lower-dimensional space. Nonlinear constraints, such as those involving entropy, are perhaps most easily dealt with in the context of the original $(n - 1)$-dimensional problem. (We refer here to dimension of the numerical optimization problem, not of the statistical estimation problem.)

Our approach is based on weighting the data in a nonuniform way, replacing the identical weights n^{-1} in a kernel estimator by weights $p_1, \ldots, p_n$ that are determined by a minimum-distance argument. This method has connections to works by Efron (1981) on tilting in nonparametric problems; Owen (1988, 1990) on empirical likelihood; Chen (1997) on empirical-likelihood density estimation; and Hall and Presnell (1999) on general biased-bootstrap methods. In essence, it finds that weighted empirical distribution, with atoms at data values, which is nearest to the empirical (or uniform bootstrap) distribution in the sense of a general distance measure (such as those considered by Cressie and Read (1984), Read (1984), and Read and Cressie (1988)), subject to the constraints being satisfied. Then it constructs the nonparametric density estimator in which data values are weighted according to this distribution.

The use of different distance measures can have an effect, although usually only slight, on results. For example, when experimenting with ρ in the interval $[-2, 2]$ we sometimes found that small, secondary modes of a density estimate became less pronounced if we used larger values of ρ. Also, when $\rho = 2$ there was a tendency for negative weights to be assigned to data in the tails of the distribution, and this could be discerned in a tendency for the estimate to make a slight dip below the axis in that tail.

However, the shapes and structures of estimates were remarkably similar for different values of ρ.

Previous work on moment constraints for density estimators has concentrated on correcting for variance inflation in kernel density estimation by rescaling the data, sometimes in conjunction with a change in bandwidth. See Jones (1991) for an insightful account of the effect on overall performance of the density estimator of adjustments to the scale of the data, as well as for discussion of the literature. In contradistinction to the methods that Jones described, our approach alters neither the data nor the amount of smoothing used to construct the estimator, only the relative weights ascribed to individual data values in the estimator. Neither does it destroy the feature that the estimator integrates to 1, which property is in effect imposed as one of the constraints.

Alternative methods for density estimation under unimodality constraints include the "smoothed likelihood" approach of Bickel and Fan (1996), the "primal-dual bases algorithm" method of Meyer (1997), and the "iterated transformation" technique of Cheng, Gasser, and Hall (1999). A range of issues related to multimodality and unimodal density estimation were discussed by Scott (1992, Sec. 9.2).

2. METHODOLOGY

2.1 WEIGHTED DENSITY ESTIMATORS

The method is based on replacing uniform empirical weights n^{-1} by more general weights p_i. It may be introduced in terms of the weighted bootstrap, as follows. Assume that the basic density estimator $\tilde{f}$, computed from a random sample $\mathcal{X} = \{X_1, \ldots, X_n\}$, is a bootstrap estimator in the sense that it may be written as as

$$\tilde{f}(x) = E\{\phi(X_1^*, \ldots, X_n^*; x) | \mathcal{X}\}\,, \tag{2.1}$$

where ϕ is a known function and $\{X_1^*, \ldots, X_n^*\}$ is a resample drawn independently and uniformly from the empirical distribution determined by $\mathcal{X}$. In this prescription we do not insist that the smoothing parameters used to construct $\tilde{f}$ be regarded as functions of data. In practice they will be, and that could be taken into account in the definition of $\tilde{f}$, but when applying the bootstrap to curve estimation one usually does not increase numerical labor by recomputing the bandwidth for each new resample.

The simplest example of an estimator of the form (2.1) is arguably a generalized kernel estimator, where $\phi(x_1, \ldots, x_n; x) = \sum_i L(x_i, x)$ and the function L depends on n and incorporates the smoothing parameter. Examples include regular kernel estimators, histogram estimators, and orthogonal series estimators (such as wavelet estimators that do not involve thresholding). In this case the prescription at (2.1) produces $\tilde{f}(x) = \sum_i L(X_i, x)$.

Given a vector $p = (p_1, \ldots, p_n)$ of probabilities for a multinomial distribution on $X_1, \ldots, X_n$, define $\{X_1^\dagger, \ldots, X_n^\dagger\}$ to be a resample drawn by sampling from $\mathcal{X}$ using the weighted bootstrap with these probabilities. That is, the n resampling operations are independent, and in each of them, X_i is drawn with probability p_i. Define

$$\tilde{f}(x|p) = E\{\phi(X_1^\dagger, \ldots, X_n^\dagger; x) | \mathcal{X}\}\,,$$

which we call the biased-bootstrap form of $\tilde{f}$. Taking $p = p_{\text{unif}} = (n^{-1}, \ldots, n^{-1})$ we recover the basic estimator $\tilde{f}$, defined at (2.1).

Suppose there are r constraints, of the form $T_j(f) = t_j$ for $1 \le j \le r$, where each T_j is a functional operating on the population density f, and $t_1, \ldots, t_r$ are constants. The biased-bootstrap form of the constraints is

$$T_j\{\tilde{f}(\cdot|p)\} = t_j \quad \text{for} \quad 1 \le j \le r\,. \tag{2.2}$$

Let $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$ denote the value of p that minimizes the distance from p to p_{unif} subject to $\sum_i p_i = 1$ and conditions (2.2). Then, our constrained estimator of $f(x)$ is $\hat{f}(x) = \tilde{f}(x|\hat{p})$.

In principle, our assumption that the basic estimator $\tilde{f}$ is expressible by (2.1) is stronger than necessary. Some highly nonlinear methods, such as soft-thresholded wavelet estimators, do not admit that representation but are candidates for our approach. However, computation can be a significant issue in highly nonlinear cases.

Appropriate definitions of distance include the power-divergence measures introduced by Cressie and Read (1984) in the context of goodness-of-fit tests for multinomial distributions. In the present setting the power-divergence distance between p_{unif} and p may be taken as

$$D_\rho(p) = \{\rho\,(1-\rho)\}^{-1}\left\{n - \sum_{i=1}^n (np_i)^\rho\right\},$$

where $-\infty < \rho < \infty$ and $\rho \ne 0, 1$. We define D_0 and D_1 by taking limits, getting

$$D_0(p) = -\sum_{i=1}^n \log(np_i)\,,$$

and

$$D_1(p) = n\sum_{i=1}^n p_i\,\log(np_i)\,,$$

which are, respectively, the Kullback–Leibler divergence between p_{unif} and p, and between p and p_{unif}. The case of D_0 is referred to in this article simply as Kullback–Leibler distance. Minimizing $D_0(p)$ subject to a set of constraints is equivalent to maximizing empirical likelihood (Owen 1988, 1990), $\prod_i p_i$, subject to the constraints. The case $\rho = \frac{1}{2}$ corresponds to Hellinger distance, and $D_2(p)$ is proportional to L_2 distance between p and p_{unif}.

The vector $\hat{p}$ always defines a proper probability distribution when $\rho < 2$, but not necessarily otherwise, owing to the potential for some of the $\hat{p}_i$'s to be negative. In a sense, $\rho > 0$ provides greatest robustness against outliers, as may be seen from the following argument. Suppose the estimator and constraint are both linear, from which it follows that the constraint may be expressed as $\sum_i p_i Y_i = t$ (see (2.4)), where Y_i denotes the effect of the constraint on the ith component of the estimator. Consider the result of allowing one or more of the values of Y_i to diverge to $\pm\infty$. Then, unless the associated p_i converges to 0, $\sum_i p_i Y_i$ will also diverge. But if $\rho \le 0$, then allowing

$p_i \to 0$ entails $D_\rho(p) \to \infty$. And the rate of divergence can be quite fast for strictly negative values of ρ.

In practice, however, ρ has to be large negative before significant effects of the choice of ρ are observed in the case of density estimation. The reason is that large values of Y_i correspond to an abnormally extreme value of the smoothing parameter (e.g., an unduly small value of the bandwidth), not to outliers in a more conventional statistical sense. Therefore, provided the smoothing parameter is chosen sensibly, the sorts of problems alluded to above seldom cause difficulty. We experimented with different ρ's, and found that the effects on density estimates of allowing ρ to vary in the interval $[-2, 2]$ were usually only minor. However, we did encounter numerical problems with values of ρ outside the range $-2 \le \rho \le 2$.

2.2 LINEAR ESTIMATORS AND LINEAR CONSTRAINTS

For estimators that are linear in the p_i's, such as those based on kernels or orthogonal series, we may write

$$\tilde{f}(x|p) = \sum_{i=1}^{n} p_i \, K_i(x) \, , \tag{2.3}$$

where K_i is a known function depending on X_i. In particular, for kernel estimators, $K_i(x) = h^{-1} \, K\{(x - X_i)/h\}$, where K is the kernel and h the bandwidth. We assume throughout that K is a continuous, compactly supported, symmetric probability density.

Often the functional T_j is also linear in the p_i's, for example having the form $T_j(f) = \int \tau_j \, f$, where τ_j is a known function. In the case of a linear estimator under linear constraints, conditions (2.2) simplify to

$$\sum_{i=1}^{n} p_i \, T_j(K_i) = t_j \, . \tag{2.4}$$

An elementary application of Lagrange multipliers shows that if distance is measured by D_ρ then, provided $\rho \ne 1$, the resulting multinomial probabilities are given by

$$p_i(c) = \left\{ \sum_{j=0}^{r} c_j \, T_j(K_i) \right\}^{-1/(1-\rho)} , \tag{2.5}$$

where $c = (c_0, \ldots, c_r)$, the constants $c_0, \ldots, c_r$ are determined by conditions (2.4) for $j = 0, \ldots, r$, and we define $T_0(K_i) = 1$ and $t_0 = 1$. When $\rho = 1$ we have instead $p_i(c) = \exp\{\sum_j \, c_j \, T_j(K_i)\}$. It may be shown that when $\rho = 0$, $c_0 = n - (c_1 t_1 + \cdots + c_r t_r)$, and so (2.5) may be reduced to r unknowns.

More generally, to find c satisfying the constraints we may apply a Newton–Raphson argument, computing successively the $(r+1)$-vectors $c^{(1)}, c^{(2)}, \ldots$ according to the algorithm $c^{(m+1)} = c^{(m)} + M(c^{(m)})^{-1} \, V(c^{(m)})$, where $V(c) = (V_j)$ and $M(c) = (M_{jk})$ are, respectively, the $(r+1)$-vector and $(r+1) \times (r+1)$ matrix of which the elements are $V_j = \sum_i \, p_i(c) \, T_j(K_i) - t_j$ and $M_{jk} = \sum_i \, p_i(c)^2 \, T_j(K_i) \, T_k(K_i)$.

Once the p_i's have been computed it is a simple matter to resample from the distribution with density $\hat{f}$, if desired. Indeed, at each resampling step one simply draws a data value from $\mathcal{X}$ according to the multinomial distribution with probabilities $p_1, \ldots, p_n$, and adds to it the value of hY, where the Y's are independently distributed (independent also of the variables drawn in the multinomial resampling step) and have the distribution with density K.

2.3 EXAMPLES OF LINEAR CONSTRAINTS

Constraints on the mean and variance, and/or on the mode and interquartile range, are motivated by obvious desires to ensure that the distribution corresponding to a density estimator enjoys the same empirical measures of location and scale as the sample from which it was computed. Not only is this important from the viewpoint of descriptive statistics, it can be critical to the analytic performance of methods based on density estimators. For example, the level accuracy of Silverman's (1981) bandwidth test for unimodality is significantly enhanced by ensuring that the distribution corresponding to the density estimator has the same variance as the sample. See also Jones (1991).

One reason for imposing constraints on higher order moments is to ensure that in applications of the smoothed bootstrap (see, e.g., Silverman and Young 1987; Young 1990), performance of the bootstrap is not impaired by smoothing the data. Bear in mind that, as revealed by Edgeworth expansions for bootstrap statistics (see, e.g., Hall 1992), one explanation for the good performance of bootstrap methods for confidence and testing procedures is that third and fourth empirical moments are root-n consistent for the respective population moments. This result fails if we resample from the distribution with density $\tilde{f}$, and choose the smoothing parameter so as to get good performance for estimating f. The reason is that the bias of $\tilde{f}$, which is of larger order than $n^{-1/2}$, contributes a term of the same size to each of the moments of the distribution with density $\tilde{f}$. We may alleviate this problem by substantially undersmoothing when computing $\tilde{f}$, but then $\tilde{f}$ does not perform well as an estimator of f, and a significant part of the attraction of the smoothed bootstrap is lost. However, using the methods suggested in the following, and without altering the bandwidth, we may constrain $\tilde{f}$ so that the first few moments are identical to their sample counterparts—allowing us to enjoy the best of both worlds.

Constraints involving single moments are linear. If in addition the estimator $\tilde{f}$ is linear, given by (2.3), then the constraint that the jth moment with respect to the density estimator $\tilde{f}(\cdot|p)$ equals the jth sample moment is expressible in the form (2.4) with

$$T_j(K_i) = \int_{\mathcal{S}} y^j \, K_i(y)\, dy = \sum_{k=0}^{\langle j/2 \rangle} \binom{j}{2k} X_i^{j-2k} \, h^{2k} \, \kappa_{2k} \,, \tag{2.6}$$

where $\mathcal{S}$ equals the support of the sampling distribution, $\langle j/2 \rangle$ denotes the integer part of $j/2$, and $\kappa_l = \int y^l \, K(y)\, dy$. (The first identity in (2.6) holds for any linear estimator, and the second for a kernel estimator with symmetric kernel K, provided $\mathcal{S}$ may be taken as the whole real line.) Furthermore, $t_j = n^{-1} \sum_i X_i^j$. The constraint that $\hat{f}$ integrate to 1 has the same form, with $j = 0$. When $j \geq 1$ and $\tilde{f}(\cdot|p)$ is a kernel estimator, constraining

the jth moment of the distribution with density $\tilde{f}(\cdot|p)$ to equal the jth sample moment is equivalent to asking that

$$\sum_{i=1}^{n} p_i \sum_{k=0}^{\langle j/2 \rangle} \binom{j}{2k} X_i^{j-2k} h^{2k} \kappa_{2k} = n^{-1} \sum_{i=1}^{n} X_i^j .$$

When imposing a constraint on values of cumulants with indexes $k_1 < \cdots < k_m$ we may not wish to restrict all moments of orders less than k_m. In this case, computations can conveniently be carried out by optimizing over the unconstrained parameters in the estimating-equations approach described by Qin and Lawless (1995).

Many constraints involving quantiles are also linear. If $\tilde{f}$ is linear too, then the constraint that the qth quantile of the distribution with density $\tilde{f}$ equals the qth sample quantile, $\hat{\xi}_q$, is given by (2.4). There, with the subscript j suppressed in t_j and T_j, we take $t = q$ and

$$T(K_i) = \int_{\mathcal{S} \cap \{y < \hat{\xi}_q\}} K_i(y)\, dy = L\{(\hat{\xi}_q - X_i)/h\} . \tag{2.7}$$

The second identity in (2.7) holds for kernel estimators, provided $\mathcal{S}$ is the whole real line and L denotes the distribution function corresponding to the density K. If we want the range between the q_1th and q_2th quantiles (for $q_1 < q_2$) to be preserved, we take instead $t = q_2 - q_1$ and, for kernel estimators, $T(K_i) = L\{(\hat{\xi}_{q_2} - X_i)/h\} - L\{(\hat{\xi}_{q_1} - X_i)/h\}$.

The main effect of imposing constraints such as (2.6) and (2.7) is to adjust the bias of $\tilde{f}$. The impact on stochastic error is negligible. Thus, for a kernel estimator, the principal effect of constraining is to add a deterministic function of which the size is that of the square of bandwidth. Moreover, the values of p_i chosen by our biased-bootstrap algorithm are uniformly close to those for the unbiased bootstrap—that is, to n^{-1}. Details of these results will be given in Section 4.

If the bandwidth is kept fixed then, as more and more constraints are imposed, there comes a point where no further constraints may be fitted. For example, if we constrain the jth moment of the distribution with density $\tilde{f}(\cdot|p)$ to equal the jth sample moment, then there is a finite value k such that we can fit the constraints for all $j \le k$ but not for $j = k + 1$. We may always increase the value of k by reducing the bandwidth—of course, k diverges to ∞ as $h \to 0$, and the corresponding constrained distribution estimator converges to the usual empirical distribution function. In practice, constraints are usually imposed for reasons other than enhancing the mean-square performance of $\tilde{f}(\cdot|p)$ as an estimator of f, but if this were the target then one could select both the bandwidth and the number of fitted moments using a simple modification of the usual least-squares cross-validation algorithm.

2.4 ENTROPY AS A CONSTRAINT

In the case of second-order methods—for example, those based on nonnegative kernels—the trade-off between bias and variance is felt most seriously at peaks and troughs, where the sign of the second derivative of the density forces the curve estimate to pull away from from the true curve, toward a measure of the "average" position of

the curve. This problem may be diminished by decreasing bandwidth, but that can have other, deleterious side effects, for example by enhancing spurious fluctuations in the density estimator in places where data are sparse. An alternative approach is to note that the oversmoothing of peaks and troughs, rendering the density estimator generally flatter, also increases entropy. We may constrain entropy and force it to decrease from the value it assumes for a conventional kernel density estimator.

From some viewpoints the problem of constraining entropy, equal to $-E\{\log f(X)\}$, is similar to that of constraining a more conventional moment, such as $E(X^2)$. Both are functionals of the density and both may be estimated root-n consistently. And in both cases, naively substituting a regular nonparametric density estimator for f, and evaluating the functional, will not achieve the optimal root-n convergence rate. In the case of entropy see, for example, Joe (1989) and Hall and Morton (1993). Therefore, we could compute a relatively accurate estimator, $\hat{\epsilon}$ say, of entropy, for example by using leave-one-out methods, and then constrain a weighted but otherwise conventional kernel density estimator so that its entropy, given by

$$T\{\tilde{f}(\cdot|p)\} = -\sum_{i=1}^{n} p_i \int K_i(x) \log\left\{\sum_{j=1}^{n} p_j K_j(x)\right\} dx\,,$$

was equal to $\hat{\epsilon}$.

However, the uses to which entropy is typically put in curve estimation—for example, the exploration of "interestingness"—are generally less explicitly quantitative than this. In contradistinction to the case of moment constraints, one would usually be interested in a range of values of entropy, not a specific value such as $\hat{\epsilon}$. In Section 3 we shall use the notion of entropy, and the fact that it reflects the "interestingness" of a density, as a device for constructing a univariate density estimator.

Other nonlinear constraints that are of practical interest include skewness and kurtosis; for example, in the context of test statistics for Normality. See, for example, Jarque and Bera (1987). When optimizing subject to nonlinear constraints, it is often simplest to optimize in the $(n-1)$-dimensional problem, using a protected Newton–Raphson algorithm (see Sec. 3.3). Alternatively, approximate methods such as those of Wood, Do, and Broom (1996) might be considered.

3. EXAMPLES AND SIMULATIONS

For the examples and simulations in this section, bandwidths were computed using the "level 2" plug-in rule of Sheather and Jones (1991), as this was found to yield generally good results. As usual, in practice it may be desirable to try a variety of bandwidths. Kullback–Leibler ($\rho = 0$) distance was used to compute p, and except in Section 3.3, the biweight kernel was employed throughout. As a rule of thumb applicable beyond the examples treated here we would suggest, in cases where heavy-tailedness is not a problem and sample size is moderately large, computing a density estimator that was constrained so that the corresponding mean, median, and variance were equal to their empirical values. (Alternatively, variance could be replaced by interquartile range.)

These constraints are the most common ones defining the location and scale associated with a dataset.

3.1 EXAMPLES

Figures 1 and 2 show density estimates based on the salaries of 353 players in the National Basketball Association in 1991, published in the newspaper, *USA Today*. With a sample size this large, the plug-in bandwidth (here equal to 3.36) is often small enough that sample moments and moderate quantiles are not greatly affected by smoothing, although shortly we shall query the effects of smoothing on small modes. The negligible impact of smoothing on quartiles is born out by the top pair of panels in Figure 1, where

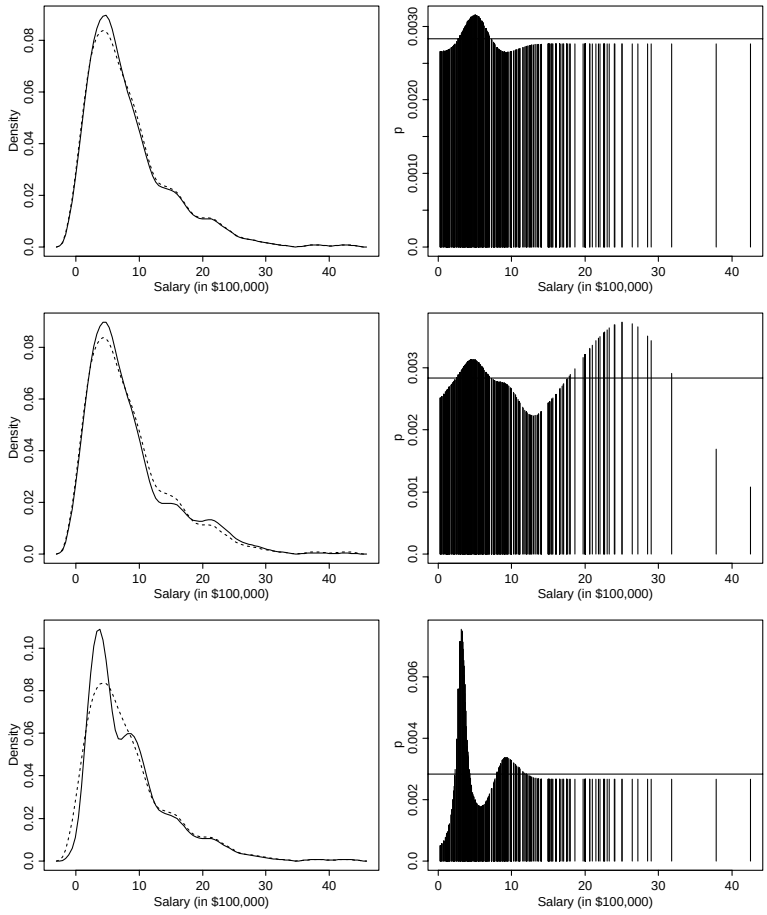

Figure 1. NBA Salary Data. Left side: unconstrained (dotted) and constrained (solid) density estimates. Right side: weights assigned to the observations (horizontal line at 1/n). Top: constraining median and quartiles to equal sample values. Middle: constraining median, quartiles, and first three moments to equal sample values. Bottom: constraining median and quartiles to equal sample values, and also constraining 12/353 = .011th quantile to equal 1.0.

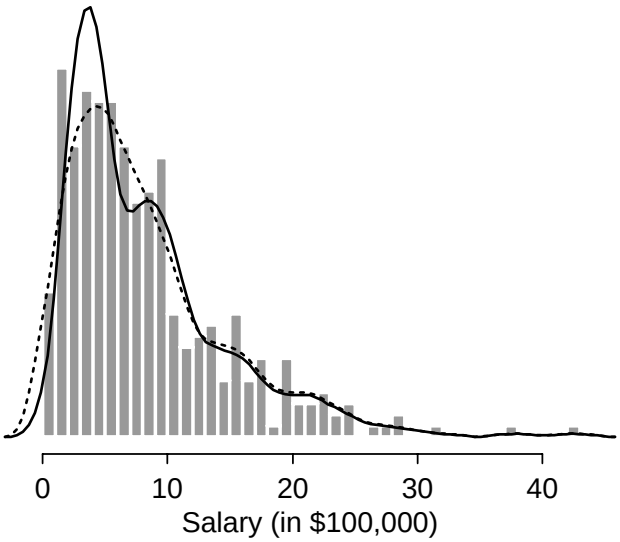

Figure 2. NBA Salary Data. Estimators from bottom left-hand panel of Figure 1 superimposed on a histogram of the data.

constraining the 25th, 50th, and 75th percentiles of the kernel density estimate to equal the corresponding sample quantiles is seen to change the estimate very little. In this respect, the fact that the quartiles are not extreme quantiles, and are not closely spaced, plays a role. As shown in the middle pair of panels, additionally constraining the first three moments to equal their sample values affects the p_i's substantially, but not the density estimate itself.

No matter how many constraints of this type one imposes, a solution of the optimization problem always exists provided the bandwidth is sufficiently small. In the present example, constraining the first four moments and the three quartiles is not possible unless one reduces the bandwidth by at least 6%, and then the resulting density estimate is quite different from the ordinary kernel estimate. Constraining the first four moments alone is feasible without any change to the bandwidth, and the main effect on the density estimate is to slightly increase its peakedness. A constraint on the first five moments requires a reduction in bandwidth of almost 70%, however, and this dramatically affects the "wiggliness" of the estimate. More details about constraints of this type will be given in Section 3.2.

It is tempting to ask whether \$1M—that is, \$1,000,000—could play the role of a ceiling in a significant number of salary negotiations. If it did, then it is conceivable that a second mode, or at least a bump, might exist in the neighborhood of \$1M. There is no evidence of this in the first two rows of Figure 1, but it is possible that evidence has been destroyed by biases introduced by smoothing. A potential major source of bias is the fact that smoothing with a fixed bandwidth and no boundary correction leads to a density estimate that places significant mass on negative salaries. That is, mass has been smeared too far to the left of \$1M. To remedy this problem we noted that only 11 salaries were under \$100,000, with another two exactly equal to \$100,000, so that

\$100,000 corresponds to approximately the $12/353 = .034$th quantile. Constraining the .034th quantile to equal \$100,000 is a very plausible way of reducing the leftwards bias; and it produces clear evidence of a second mode at approximately \$1M, well away from the location of the constraint. See the last row of panels in Figure 1.

Incidentally, constraining the .65, .70, and .75th quantiles to equal their sample values also brings out this feature. (Note that \$1,000,000 is roughly equal to the .70th sample percentile.) Constraining the qth quantile of the density estimate to equal the qth quantile of the data, for any single value of q in the range $1/353 \leq q \leq .05$, also produces a second mode. Its peakedness decreases as q increases. Beyond about $q = .05$ the new mode degenerates to a shoulder. Constraining the .50, .70, and .75th quantiles to equal their sample values also produces a shoulder. Therefore, we argue, there is evidence of a second mode in the vicinity of \$1M, but in the unconstrained density estimator it is destroyed by biases introduced through smoothing.

3.2 SIMULATION STUDY OF MOMENT CONSTRAINTS

To explore the effects of moment constraints for small samples we generated 1,000 samples of sizes $n = 50$ and $n = 100$ for the following four densities from Wand and Jones (1995, tab. 2.2), listed there in order of increasing difficulty for kernel density estimation: (a) Normal, (b) asymmetric bimodal Normal mixture $\frac{3}{4}N(0,1) + \frac{1}{4}N(\frac{3}{2}, \frac{1}{9})$, (c) symmetric bimodal Normal mixture $\frac{1}{2}N(-1, \frac{4}{9}) + \frac{1}{2}N(1, \frac{4}{9})$, and (d) chi-squared with six degrees of freedom.

The qualitative effect of constraining the first three moments to equal their sample values is generally small. Constraining four moments affects the estimates more seriously, however, particularly with samples of size $n = 50$. Constraining moments generally leads to slightly increased variability in the body of the distribution, and decreased variability in the tails. In the case of matching four moments with sample size $n = 50$, the increased variability in the middle of the distribution can be substantial.

The results for $n = 100$ are summarized in Figures 3 through 8. Figures 3–6 show pointwise medians and 5th and 95th percentiles of the density estimators. Figures 7 and 8, computed by averaging the results of the simulation study, compare the pointwise bias and standard deviation of the estimators for the four densities. Distribution-specific conclusions, based on many observations of density estimators for individual realizations as well as the summaries provided by the figures, are given in the following. There we discuss performance of constrained density estimators in terms of their ability to approximate the true density, although of course we recognize that moment matching would generally not be implemented for that explicit purpose.

1. For the Normal distribution, matching the first two sample moments improves the estimate mainly by reducing bias, although there is a slight overall reduction in variability. Matching the first three sample moments has only a small additional effect, producing a slight increase in bias relative to matching two moments, but still significantly less bias than the unconstrained estimate. In the center of the

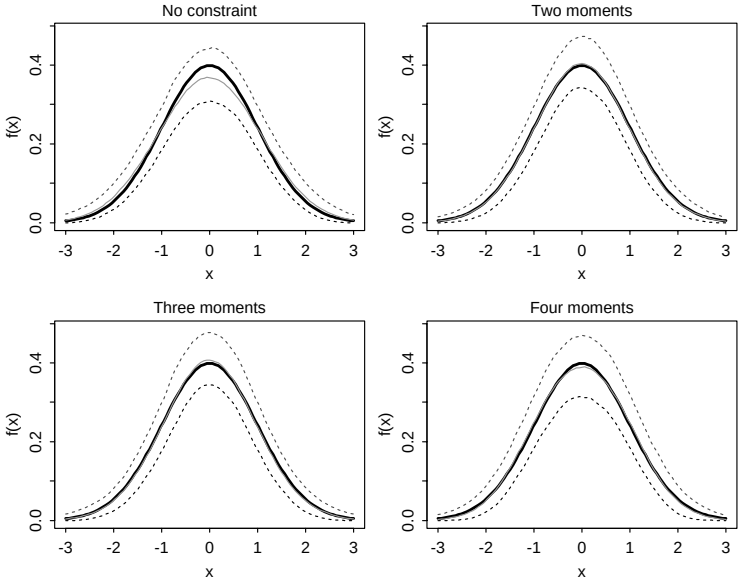

Figure 3. Pointwise median (solid line) and 5th and 95th percentiles (dotted lines) when estimating Normal density (heavy line) with sample size n = 100. Panels show the results of matching the first 0 (1), 2, 3, and 4 sample moments, respectively.

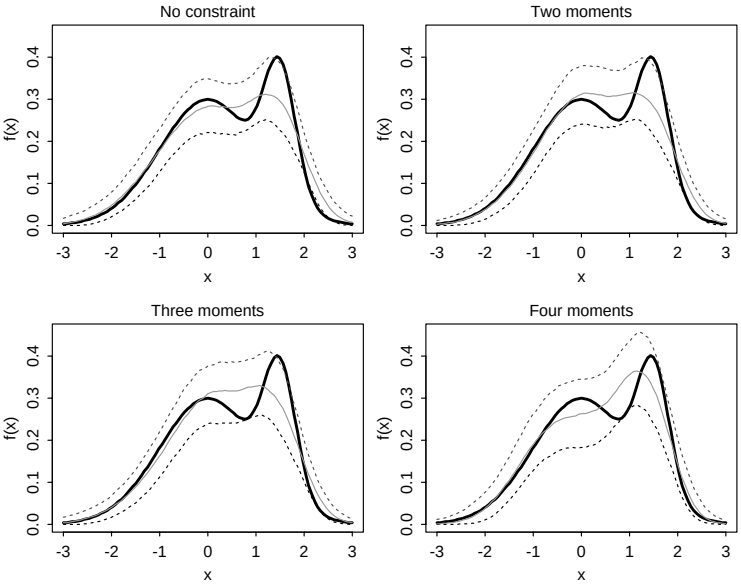

Figure 4. Pointwise median (solid line) and 5th and 95th percentiles (dotted lines) when estimating Normal mixture (b) (heavy line) with sample size n = 100. Panels show the results of matching the first 0 (1), 2, 3, and 4 sample moments, respectively.

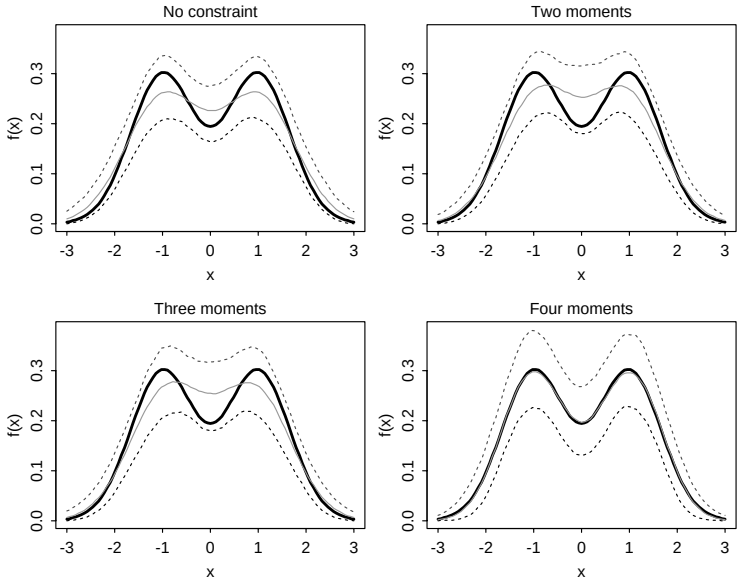

Figure 5. Pointwise median (solid line) and 5th and 95th percentiles (dotted lines) when estimating Normal mixture (c) (heavy line) with sample size $n = 100$. Panels show the results of matching the first 0 (1), 2, 3, and 4 sample moments, respectively.

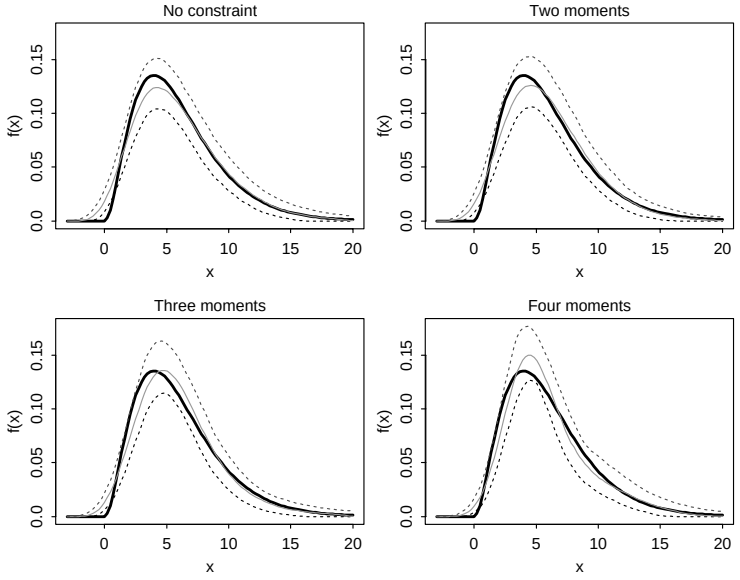

Figure 6. Pointwise median (solid line) and 5th and 95th percentiles (dotted lines) when estimating chi-squared (6) density (heavy line) with sample size $n = 100$. Panels show the results of matching the first 0 (1), 2, 3, and 4 sample moments, respectively.

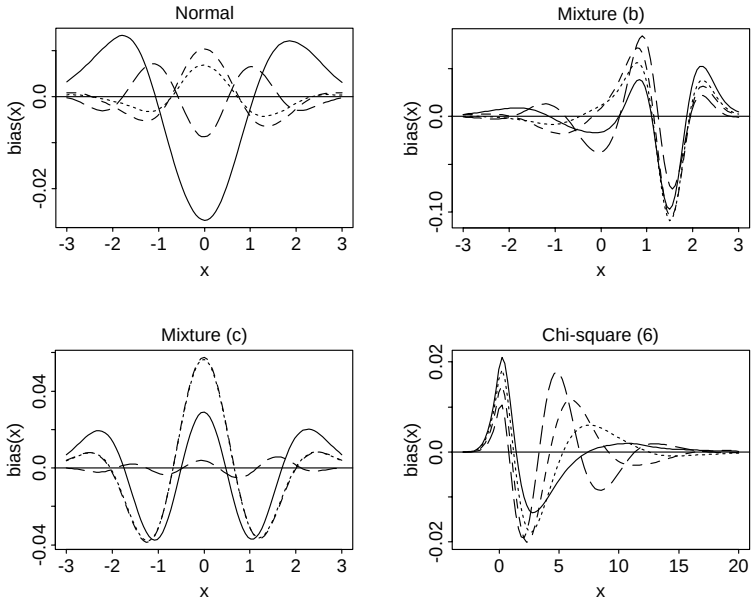

Figure 7. Bias of density estimates with $n = 100$. Unconstrained estimate (solid curve) and moment-con- strained estimates: first two moments (short dashes); three moments (medium dashes); four moments (long dashes).

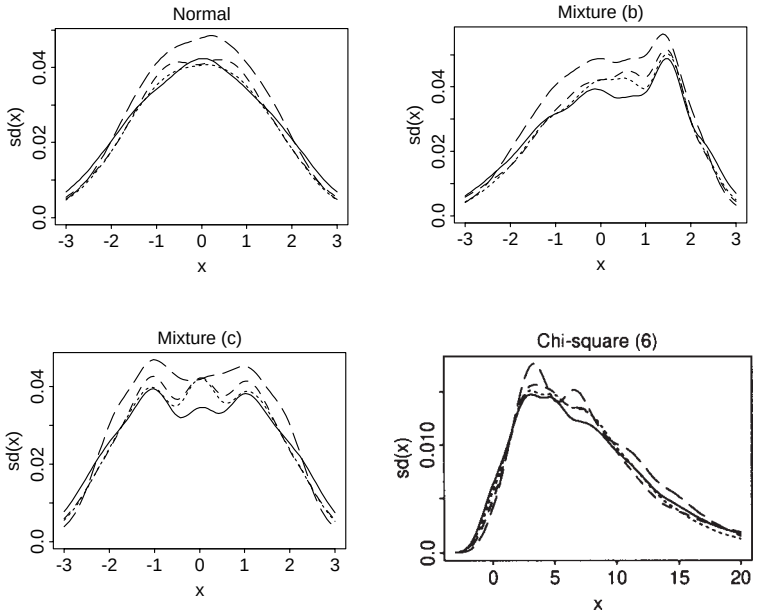

Figure 8. Standard deviations of density estimates with $n = 100$. Unconstrained estimate (solid curve) and moment-constrained estimates: first two moments (short dashes); three moments (medium dashes); four moments (long dashes).

distribution, variability increases slightly with the number of moment constraints. Qualitatively, the constrained estimates were more sharply peaked because of variance correction, but otherwise not much affected.

2. For Normal mixture (b), estimates constrained by fitting four moments did better at representing the primary mode and markedly less well at representing the trough, relative to unconstrained estimates. Qualitatively, estimates based on fitting all four moments performed best, usually getting the shape of the distribution right although sometimes mistaking the trough for a slope between the two modes.
3. Strikingly, for mixture (c) constraining the first four moments practically eliminated bias, more than offsetting the increase in variability and making the four-matched-moments estimate arguably the best of the four considered. Matching two or three moments produced less bias in the vicinity of the two modes, but more near the trough, relative to the unconstrained case.
4. For the chi-squared (6) distribution, constraining the first three moments improved estimation of the maximum height of the density and of values to the left of the mode, with some worsening to the right of the mode. Matching all four moments often resulted in slight over-estimation of the peakedness of the density, and lead to increased variability.

3.3 ENTROPY AND UNIMODALITY

Since the entropy constraint is highly nonlinear in the p_i's, constraining entropy is computationally more difficult than constraining moments and/or quantiles. For the work discussed in the following we used a protected Newton–Raphson algorithm.

Changing the entropy of a kernel density estimate by more than a small amount can drastically alter the estimated density. In particular, reducing the entropy increases peakedness and reduces bumps in the tails. Combining this observation with the fact that increasing bandwidth also tends to reduce the number of modes, while decreasing peakedness, suggests the following method for producing a unimodal density estimate. (1) Compute a kernel density estimate with a bandwidth designed for good overall performance, and compute the maximum height of the estimate. (2) Increase bandwidth by a small amount. Then, if necessary, decrease the entropy of the resulting estimate until its maximum height is at least as large as that of the original estimate. (Note that both these actions tend to produce a unimodal estimate. Using the height of the original estimate as a threshold serves to ensure that the new, constrained estimate is biased downwards to no greater an extent than was the original, unconstrained form.) If the resulting density estimate is unimodal, stop; otherwise, continue increasing bandwidth and decreasing entropy in small increments until a unimodal estimate is obtained. To remove shoulders in the latter estimate one may iterate a little further.

Using the Normal kernel we applied this algorithm to all 1,000 of the the $n = 50$ chi-squared (6) samples from the simulation study in the previous section, without experiencing any any difficulties. That is to say, on each occasion the algorithm converged to a unimodal density estimate. Results for the first four samples for which the standard kernel density estimate was multimodal, and the corresponding data weights, are shown

in Figure 9. To generate those curves we stopped as soon as a unimodal estimate was achieved, and so a shoulder is visible in each estimate. Fifty-four percent of the 1,000 density estimates were multimodal.

For the Normal kernel, increasing the bandwidth is guaranteed to monotonically reduce the number of modes (Silverman 1981). When the bandwidth is first increased by a small amount, it can happen that the maximum height of the density estimate

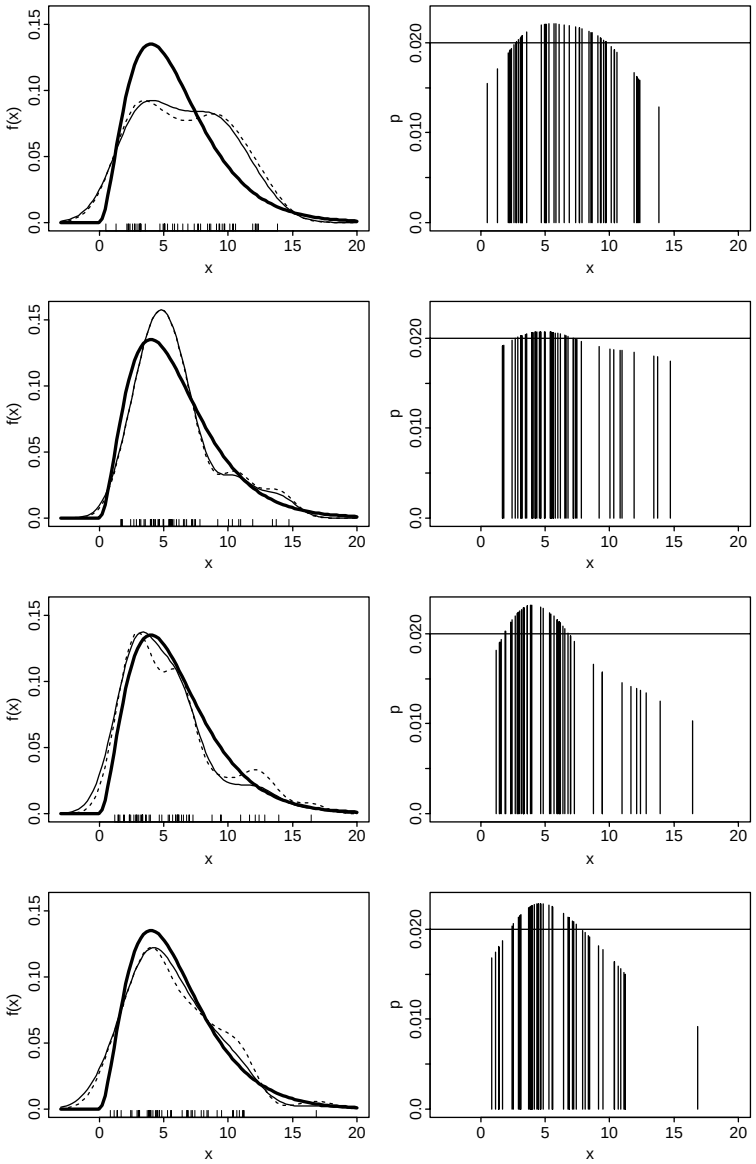

Figure 9. Constrained unimodal density estimates (solid curve) for first four samples of n = 50 chi-squared (6) observations for which the plug-in bandwidth yielded a multimodal kernel density estimate (short dashes).

actually increases, so that no entropy constraint is needed in the first few iterations of the algorithm. For data from very long-tailed distributions (but not the chi-squared (6) distribution, in our experience), it may not always be possible, after increasing bandwidth, to reduce the entropy of the density estimate enough to reach the original maximum height. This difficulty can be overcome by using a locally varying bandwidth, providing greater smoothing in the tail than near the peak. Since our algorithm is quite effective in producing unimodality, the locally varying bandwidth need not be chosen in a sophisticated way.

4. OUTLINE OF THEORY

We shall develop theoretical properties of the principal methods described in Sections 2.1–2.3. Our main results are that, after imposing the constraints, (1) the components of $\hat{p}$ are close to n^{-1}, in relative terms, in the sense that $\max_i n\hat{p}_i$ converges to one in probability as $n \to \infty$; (2) the difference between the basic estimator $\tilde{f}$ and its constrained form $\hat{f}$ is mostly in terms of an adjustment to bias, with negligible alterations to error-about-the-mean; and (3) the size of the bias-adjustment term is of the same order as the original bias of $\tilde{f}$. In deriving these results we consider only the case of kernel estimators and linear constraints, of the form (2.6) or (2.7). It is assumed that there are r constraints. For brevity we work only with Kullback–Leibler distance, although other cases are virtually identical. Regularity conditions will be stated during the proof.

Define the r-vectors $S = (S_1, \ldots, S_n)^T$ and $\Delta = (\Delta_1, \ldots, \Delta_r)^T$ by $S_i = (T_1(K_i), \ldots, T_r(K_i))^T$ and $\Delta_j = n^{-1} \sum_i T_j(K_i) - t_j$, and let $A = (A_{jk})$ and $\hat{A} = (\hat{A}_{jk})$ be the $r \times r$ matrices given by $A_{jk} = E\{T_j(K_1)\, T_k(K_1)\}$ and $\hat{A}_{jk} = n^{-1} \sum_i T_j(K_i)\, T_k(K_i)$. We assume that each $E\{T_j(K_1)^2\} < \infty$ and A is nonsingular. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_r)^T$ be another r-vector, and observe from (2.5) (and discussion of the case $\rho = 0$ just below that formula) that in the case of Kullback–Leibler distance we may write $p_i = n^{-1}(1 + \epsilon^T S_i)^{-1}$, where ϵ is determined by the constraints. The latter have the form

$$t_k = \sum_{i=1}^n p_i\, T_k(K_i) = n^{-1} \sum_{i=1}^n \left\{1 - \epsilon^T S_i + (\epsilon^T S_i)^2 - \ldots\right\} T_k(K_i)\,,$$

for $1 \le k \le r$, whence it follows that $\epsilon = \hat{A}^{-1}\Delta + O_p(\|\Delta\|^2)$. Because $\hat{A} = A + O_p(n^{-1/2})$, then

$$p_i = n^{-1}\left\{1 - \Delta^T A^{-1} S_i + O_p\left(n^{-1/2}\|\Delta\|\,\|S_i\| + \|\Delta\|^2\,\|S_i\|^2\right)\right\},$$

where the remainder is of the stated order uniformly in i. This establishes claim (1) in the first paragraph of this section. Moreover,

$$\sum_{i=1}^n p_i\, K_i = n^{-1} \sum_{i=1}^n K_i - \Delta^T\, A^{-1}\, v + R\,, \tag{4.1}$$

where

$$R = O_p\left\{n^{-1} \sum_{i=1}^n \left(n^{-1/2}\|\Delta\|\,\|S_i\| + \|\Delta\|^2\,\|S_i\|^2\right) K_i\right\}$$

and $v = (v_1, \ldots, v_r)^T$ is the vector defined by $v_j = E\{T_j(K_1)\, K_1\}$.

Equivalently to (4.1), $\hat{f} = \tilde{f} - \Delta^T\, A^{-1}\, v + R$. Assume that f has two bounded, continuous derivatives on its support, which equals an interval. Then, v generally converges to a fixed, nonzero vector v^0, say, as n tends to infinity (or equivalently, as the amount of smoothing used to define each K_i converges to 0). For example, if the estimator is evaluated at x, and a given component of v corresponds to the case where the functional T_j is defined by (2.6) or (2.7), then the corresponding component of v^0 equals $x^j\, f(x)$ or $I(x < \xi_q)\, f(x)$, respectively, where ξ_q denotes the qth quantile of the distribution with density f. (In the case of (2.6) we assume that $E(X^{2j}) < \infty$, and for (2.7), that f is bounded above zero in the neighborhood of ξ_q.) The presence of the indicator function in the case of (2.7) leads to an asymptotic discontinuity in the second derivative of $\hat{f}$, but not in the estimator itself or in its gradient.

The stochastic component of Δ is generally of smaller order than the expected value of Δ, which in turn is of the same size as the bias of $\tilde{f}$. To appreciate why, we again go back to the examples at (2.6) and (2.7), and note that if the estimator is evaluated at x and if a given component of Δ corresponds to the case where T_j is defined by (2.6), then that component equals $\frac{1}{2} h^2 j(j-1) \kappa_2\, E(X_1^{j-2}) + o_p(h^2) + O_p(n^{-1/2})$; while if $T = T_j$ is defined by (2.7), then the corresponding component of Δ equals $\frac{1}{2} h^2 \kappa_2 f'(\xi_q) + o_p(h^2) + O_p(n^{-1/2})$. Therefore, in each case, the jth component of Δ equals $h^2 g_j + o_p(h^2) + O_p(n^{-1/2})$, where g_j and g will denote functions not depending on h or n. Hence, $\hat{f} = \tilde{f} + h^2 g + o_p\{(nh)^{-1/2} + h^2\}$. Since the bias of $\tilde{f}$ is of size h^2, and the variance is of size $(nh)^{-1}$, then we have established claims (2) and (3) in the first paragraph.

ACKNOWLEDGMENTS

We are grateful to three referees and an associate editor for their helpful comments. This research was carried out during the second author's tenure as Research Associate in the Centre for Mathematics and its Applications, Australian National University.

[Received January 1998. Revised July 1998.]

REFERENCES

Bickel, P. J., and Fan, J. (1996), "Some Problems on the Estimation of Unimodal Densities," *Statistica Sinica*, 6, 23–45.

Chen, S. X. (1997), "Empirical Likelihood-Based Kernel Density Estimation," *Australian Journal of Statistics*, 39, 47–56.

Cheng, M.-Y., Gasser, T., and Hall, P. (1999), "Nonparametric Density Estimation Under Unimodality and Monotonicity Constraints," *Journal of Computational and Graphical Statistics*, 8, 1–21.

Cressie, N. A. C., and Read, T. R. C. (1984), "Multinomial Goodness-of-Fit Tests," *Journal of the Royal Statistical Society*, Ser. B, 46, 440–464.

Efron, B. (1981), "Nonparametric Standard Errors and Confidence Intervals" (with discussion), *Canadian Journal of Statistics*, 36, 369–401.

Friedman, J. H. (1987), "Exploratory Projection Pursuit," *Journal of the American Statistical Association*, 82, 249–266.

Friedman, J. H., Stuetzle, W., and Schroeder, A. (1984), "Projection Pursuit Density Estimation," *Journal of the American Statistical Association*, 79, 599–608.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge: Cambridge University Press.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, New York: Springer.

Hall, P., and Morton, S. (1993), "On the Estimation of Entropy," *The Annals of the Institute of Statistical Mathematics*, 45, 69–88.

Hall, P., and Presnell, B. (1999), "Intentionally-Biased Bootstrap Methods," *Journal of the Royal Statistical Society*, Ser. B, 61, 143–158.

Jarque, C. M., and Bera, A. K. (1987), "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, 55, 163–172.

Joe, H. (1989), "Estimation of Entropy and Other Functionals of a Multivariate Density," *The Annals of the Institute of Statistical Mathematics*, 41, 683–687.

Jones, M. C. (1991), "On Correcting for Variance Inflation in Kernel Density Estimation," *Computational Statistics and Data Analysis*, 11, 3–15.

Mammen, E. (1992), *When Does Bootstrap Work?* New York: Springer.

Meyer, M. C. (1997), "An Extension of the Mixed Primal-Dual Bases Algorithm to the Case of More Constraints Than Dimensions," unpublished manuscript.

Owen, A. B. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249.

——— (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120.

Qin, J., and Lawless, J. (1995), "Estimating Equations, Empirical Likelihood and Constraints on Parameters," *Canadian Journal of Statistics*, 23, 145–159.

Read, T. R. C. (1984), "Small-Sample Comparisons of the Power Divergence Goodness-of-Fit Statistics," *Journal of the American Statistical Association*, 79, 929–935.

Read, T. R. C., and Cressie, N. A. C. (1988), *Goodness-of-Fit Statistics for Discrete Multivariate Data*, New York: Springer.

Scott, D. W. (1992), *Multivariate Density Estimation—Theory, Practice and Visualization*, New York: Wiley.

Sheather, S. J., and Jones, M. C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society*, Ser. B, 53, 683–690.

Silverman, B. W. (1981), "Using Kernel Density Estimates to Investigate Multimodality," *Journal of the Royal Statistical Society*, Ser. B, 43, 97–99.

Silverman, B. W., and Young, G. A. (1987), "The Bootstrap: To Smooth or not to Smooth," *Biometrika*, 74, 469–479.

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

Wood, A. T. A., Do, K.-A., and Broom, B. M. (1996), "Sequential Linearization of Empirical Likelihood Constraints With Application to U-Statistics," *Journal of Computational and Graphical Statistics*, 5, 365–385.

Young, G. A. (1990), "Alternative Smoothed Bootstraps," *Journal of the Royal Statistical Society*, Ser. B, 52, 477–484.