# Nonparametric Density Estimation Under Unimodality and Monotonicity Constraints

Ming-Yen Cheng, Theo Gasser, and Peter Hall

We introduce a recursive method for estimating a probability density subject to constraints of unimodality or monotonicity. It uses an empirical estimate of the probability transform to construct a sequence of maps of a known template, which satisfies the constraints. The algorithm may be employed without a smoothing step, in which case it produces step-function approximations to the sampling density. More satisfactorily, a certain amount of smoothing may be interleaved between each recursion, in which case the estimate is smooth. The amount of smoothing may be chosen using a standard cross-validation algorithm. Unlike other methods for density estimation, however, the recursive approach is robust against variation of the amount of smoothing, and so choice of bandwidth is not critical.

**Key Words:** Curve estimation; Isotonic regression; Iteration; Kernel methods; Mode; Mode testing; Probability transform; Recursion; Smoothing; Turning point.

## 1. INTRODUCTION

The number of modes in a distribution provides an indication of the number of components in the population which it represents. Therefore, testing for the number of modes is a way of assessing the number of subpopulations, and often the null hypothesis is that the population is homogeneous—that is, the sampling distribution has a single mode. Early approaches to mode testing were based on fitting parametric mixture models, usually Normal mixtures (e.g., Cox 1966), but more recently a variety of nonparametric tests has become available (e.g., Silverman 1981; Hartigan and Hartigan 1985; Müller and Sawitzki 1991). Assuming that a nonparametric test does not reject the null hypothesis of unimodality, there is interest in estimating the unimodal density nonparametrically. In this article we suggest a new approach to this problem, based on treating a general unimodal density as a transformation of a known, unimodal template. We introduce a recursive method for estimating the transformation, and show how to adapt this technique to the problem of density estimation under the constraint of monotonicity.

Ming-Yen Cheng is Associate Professor, Department of Mathematics, National Taiwan University, Taipei 106, Taiwan (E-mail: cheng@math.ntu.edu.tw). Theo Gasser is Professor, Department of Biostatistics, University of Zürich, Sumastrasse 30, 8006 Zürich, Switzerland. Peter Hall is Professor, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

©1999 *American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*
*Journal of Computational and Graphical Statistics, Volume 8, Number 1, Pages 1–21*

One popular test for unimodality, the bootstrap or bandwidth test of Silverman (1981), is based directly on nonparametric estimation of the sampling density. It might be thought that the test could form the basis for unimodal estimation of the density, assuming that the null hypothesis was not rejected. Unfortunately this is not the case, for at least two reasons. First, the unimodal density estimate produced by the test is smoothed in a global way that is determined almost entirely by local features of the density at the mode. This is clear from asymptotic properties of the test, for example those derived by Mammen, Marron, and Fisher (1992). It may be shown numerically that this can result in substantial oversmoothing in some places, and undersmoothing in others, relative to what is desirable to produce an estimator that is at least visually accurate. Smoothing less than the amount determined by the bandwidth test will destroy the unimodality property.

Second, the bandwidth test is particularly susceptible to data clusters away from the central part of the distribution, and often requires a very large bandwidth if it is to produce a density estimator with a single mode. That bandwidth may even diverge to infinity with increasing sample size; this would happen if the sample were from a Student's $t$ distribution, for example. In practice, this difficulty is usually overcome by confining attention to data from the center of the distribution, but this means that the density estimator is only available in the central region, not for the full support of the density.

Wang (1995) addressed $L^1$ theory of density estimation under qualitative constraints. Bickel and Fan (1996) suggested several methods based on "smoothed likelihood" for estimating a density subject to the constraint of unimodality. Their approach is relatively amenable to asymptotic analysis, but requires explicit estimation of the mode. Meyer (1997) proposed methods for estimating nonparametric regression means under qualitative constraints, starting with the "primal–dual bases algorithm" of Fraser and Massam (1989). Her approach involves a degree of subjective choice, or at least choice from outside the algorithm, for example about where points of inflection lie. Concave and convex parts of the function are fitted separately. There could be an arbitrarily large number of these, even under the constraint of unimodality. The literature on nonparametric density or regression estimation under monotonicity, or isotonic regression, is extensive. It includes, for example, Grenander's (1956) "nonparametric maximum likelihood approach," which produces a step function estimator of a monotone density; and recent work of Woodroofe and Sun (1993) and Sun and Woodroofe (1996), which develops penalized versions of Grenander's method. In this context we should note the monograph of Robertson, Wright, and Dykstra (1988) on inference under "monotonicity constraints."

We should also mention the recent contributions of Qian (1994) on least-squares methods, Tantiyaswasdikul and Woodroofe (1994) on isotonic smoothing splines, and Shi (1995) on isotonic regression in $L_1$. In principle, any regression method can be converted into a density estimation method by binning the data and regressing on the bin counts. This approach has a number of disadvantages, however, including diminished control over the smoothing step and the potential for negative estimators, particularly in regions where the density is low. Our recursive approach to monotone density estimation works particularly well when the density vanishes at its lower end.

We shall introduce our method in Section 2, and describe its numerical performance in Section 3. Section 4 will outline its theoretical properties.

## 2. METHODOLOGY

Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote a sample from a distribution with density $f$, and let $g$ be a known probability density function of the same type as $f$. That is, $g$ is unimodal if we wish our estimator $\hat{f}$ to have that property, and monotone increasing (or decreasing) if we want that for $\hat{f}$. Let $F$ (an unknown function) and $G$ (known) be the respective distribution functions corresponding to $f$ and $g$, and let $\widehat{F}$ be the empirical distribution function of the sample.

To simplify our exposition we shall suppose that $f$ and $g$ are both supported on the interval $\mathcal{I} = [0, 1]$. Our methods apply without difficulty to heavy-tailed densities with infinite support, where we would usually proceed as though the common support of $f$ and its template were equal to (or a little greater than) the interval between the largest and smallest data values. Numerical examples illustrating this point will be given in Section 3.

Leaving aside for a moment the stochastic aspects of the problem, suppose we have constructed a sequence of densities $f_1 \equiv g, f_2, \ldots, f_j$, each successively closer to $f$. Let $F_j$ denote the distribution function corresponding to $f_j$. Then, $f_j = f$ is a fixed point of the transformation that takes $f_j$ to the density $f_{j+1}$ that equals a constant multiple of $f_j\{F_j^{-1}(F)\}$, the constant being uniquely defined by the requirement that $f_{j+1}$ is a density. Therefore, if $f_j$ was close to $f$ then we expect $f_{j+1}$ to be even closer. Section 4 will explore these theoretical issues in more detail, but in the meantime we base our estimator $\hat{f}$ on an empirical version of the heuristic argument just above, producing a sequence of density estimators $\hat{f}_1 \equiv g, \hat{f}_2, \hat{f}_3, \ldots$ converging to $\hat{f}$, as follows.

Suppose we have constructed the sequence as far as $\hat{f}_j$. Define $\widehat{F}_j$ to be the distribution function corresponding to $\hat{f}_j$, and put $\tilde{f}_{j+1} = \hat{f}_j\{\widehat{F}_j^{-1}(\widehat{F})\}$ and $\hat{f}_{j+1} = \tilde{f}_{j+1} / \int_{\mathcal{I}} \tilde{f}_{j+1}$. Thus, if we define $T_j$ to be the mapping from $\mathcal{I}$ to $\mathcal{I}$ that may be represented by $\widehat{F}_j^{-1}(\widehat{F})$, then $\hat{f}_{j+1}$ is proportional to $g_j = g(T_1 \ldots T_j)$ for $j \geq 0$, where $T_1 \ldots T_j$ denotes the compound transformation $T_1(T_2(\ldots(T_j)\ldots))$. Assuming that the sequence $\{\hat{f}_j\}$ converges, we let $\hat{f}$ denote its limit.

In the absence of smoothing, this algorithm produces a sequence of functions that resemble successive step-function approximations to $f$; see panel (b) of Figure 1. The sequence converges to a limit, which in $L^p$ terms is a reasonably accurate estimate of $f$. However, its discontinuities (arising from the jump discontinuities of $\widehat{F}$) make it unattractive as an estimate of a smooth density.

We experimented with smoothing before, during, and after the algorithm, and found the second approach to produce the best results. It amounts to interleaving a smoothing step between each component $T_i$ in the compound transformation $U_j = T_1 \ldots T_j$, as shown in the next paragraph. Kernel smoothing before the iteration part of the algorithm produces an estimate which, like a kernel density estimate, is very sensitive to choice of the smoothing parameter. By way of contrast, the estimate derived by smoothing during the algorithm is only mildly susceptible to variation of the bandwidth, as we shall show in Section 3. Kernel smoothing after iteration was not really effective unless the bandwidth was chosen carefully to vary with location, since the jumps that need to be smoothed out were of different sizes. Additionally, smoothing before or after iteration produced edge-effect problems, particularly in the case of monotone functions. While these could be removed after a little manipulation, it was better not to have to deal with them at all.

Let $K$ be a continuous, compactly supported density, such as the Epanechnikov kernel (see, e.g., Wand and Jones 1995, p. 30), let $L$ denote the corresponding distribution function, and let $h$ be a bandwidth. Suppose we have already computed the smoothed version, $S_{j-1}$, of $U_{j-1}$. Simply adjoining the step-function transformation $T_j$, by using

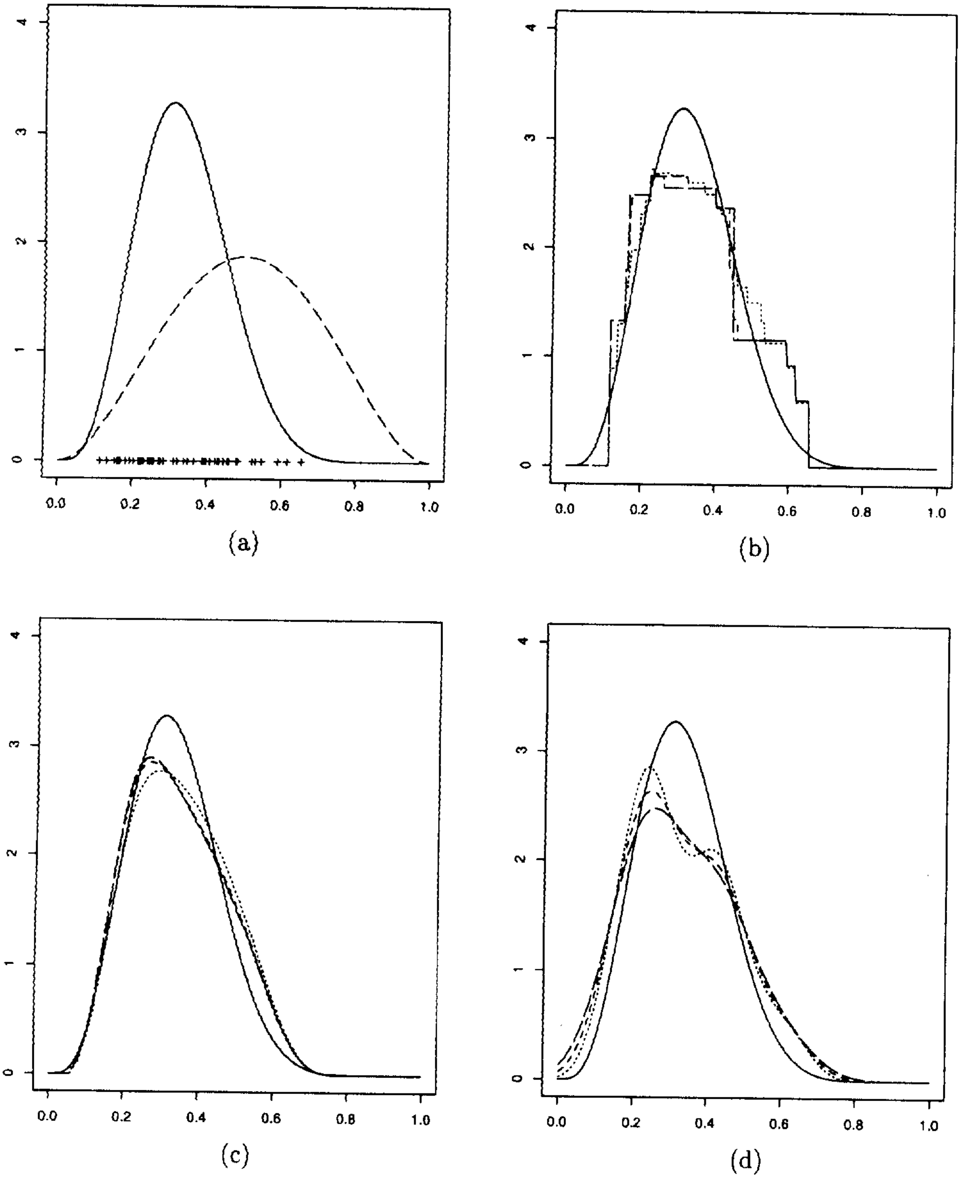


*Figure 1. Comparison of unimodal density estimates using recursion with no smoothing (panel (b)), recursion with smoothing (panel (c)), and kernel methods (panel (d)). The recursion estimates $\hat{f}_3, \hat{f}_6, \hat{f}_9$ are represented by the dotted, dashed, and long-dashed lines, respectively. Panel (a) depicts the true $f$ (unbroken line) and template $g$ (broken line).*

$S_{j-1}T_j = S_{j-1}(T_j)$ as an approximation to $U_j = U_{j-1}T_j$, produces a discontinuous transformation once more. Therefore, we construct instead the following smoothed version of $S_{j-1}T_j$,

$$s_j(x) = \int L\{(x-t)/h\}\, d(S_{j-1}T_j)\,,$$

which we compute in the form

$$\begin{aligned} s_j(x) &= \sum_i \big[L\{(x-t_i)/h\}I(|x-t_i|<\epsilon) + I(x-t_i \geq \epsilon)\big] \\ &\qquad \times \big\{(S_{j-1}T_j)(t_{i+1}) - (S_{j-1}T_j)(t_i)\big\}\,, \qquad (2.1) \end{aligned}$$

where $\ldots < t_i < t_{i+1} < \ldots$ are points on a fine grid on the interval $\mathcal{I}$, and $\epsilon = \epsilon(x) = \min\{h, x, 1-x\}$. Owing to small discrepancies at the boundaries, produced by smoothing, this transformation may not take $\mathcal{I}$ exactly to $\mathcal{I}$. Hence, we adjust by making a linear transformation, obtaining finally:

$$S_j(x) = \frac{s_j(x) - s_j(0)}{s_j(1) - s_j(0)}\,, \quad \text{for} \quad x \in \mathcal{I}\,.$$

In place of the unsmoothed functions $g_j = g(T_1 \ldots T_j) = g(U_j)$ and $\hat{f}_{j+1} = g_j / \int_{\mathcal{I}} g_j$ introduced three paragraphs earlier, we redefine $g_j = g(S_j)$ and $\hat{f}_j$ as the same functional of $g_j$. The sequence $\hat{f}_1, \hat{f}_2 \ldots$ converges to our smoothed estimator, $\hat{f}$. See panel (c) of Figure 1.

It is worth stressing the obvious point that, by their definitions, the transformations $S_j$, $T_j$ and $U_j$ are all monotone increasing. This property guarantees that, if the template $g$ is unimodal or monotone, the same is true of each iterate $g_j$ and hence of all members of the sequence of iterates $\hat{f}_1, \hat{f}_2, \ldots$.

Standard cross-validation arguments may be employed to select the bandwidth $h$ and any parameters on which the template, $g$, might depend. (Parameter-dependent templates can be useful for estimation subject to more general constraints, as we shall note in Section 3.) For example, suppose $g = g(\cdot|\theta)$ depends on a parameter vector $\theta$, and write $\hat{f}$ as $\hat{f}(\cdot|\theta, h)$ to express dependence of the limit of the sequence $\hat{f}_1, \hat{f}_2 \ldots$ on $h$ and $\theta$. Let $\hat{f}_{(i)}(\cdot|\theta, h)$ denote the version of $\hat{f}(\cdot|\theta, h)$ that we compute if we omit $X_i$ from the sample $\mathcal{X}$, and define

$$\mathrm{CV}(\theta, h) = \int \hat{f}(x|\theta, h)^2\, dx - 2\, n^{-1} \sum_{i=1}^{n} \hat{f}_{(i)}(X_i|\theta, h)\,. \qquad (2.2)$$

Choose $(\hat{\theta}, \hat{h})$ to minimize $\mathrm{CV}(\theta, h)$, and take $\hat{f}(\cdot|\hat{\theta}, \hat{h})$ as our final estimator of $f$. In theory a different bandwidth could be used for each iteration. However, the exceptional insensitivity of the estimator to choice of bandwidth (see Sec. 3) means that this is unnecessary, and so in the numerical results reported here we employed the same bandwidth at each step. This greatly simplified the procedure and reduced computing time to manageable proportions. Without this simplification, cross-validation would not really be practicable in the present context. The reader is referred to Scott (1992, sec. 6.5) and Wand and Jones (1995, sec. 3.3) for accounts of cross-validation in density estimation.

Panel (a) of Figure 1 depicts the true unimodal density $f$ (represented by the unbroken line, and repeated in panels (b)–(d)) and the template $g$ (broken line). The functions $f$ and $g$ are Beta densities with parameters $(5, 10)$ and $(3, 3)$, respectively. The sample is illustrated as a "rug" on the horizontal axis of panel (a). Sample size is $n = 50$. Panel (b) shows three members $\hat{f}_3, \hat{f}_6, \hat{f}_9$ of the sequence $\{\hat{f}_j\}$ when the smoothing step is omitted; and panel (c) depicts the same functions when the smoothing step is included.

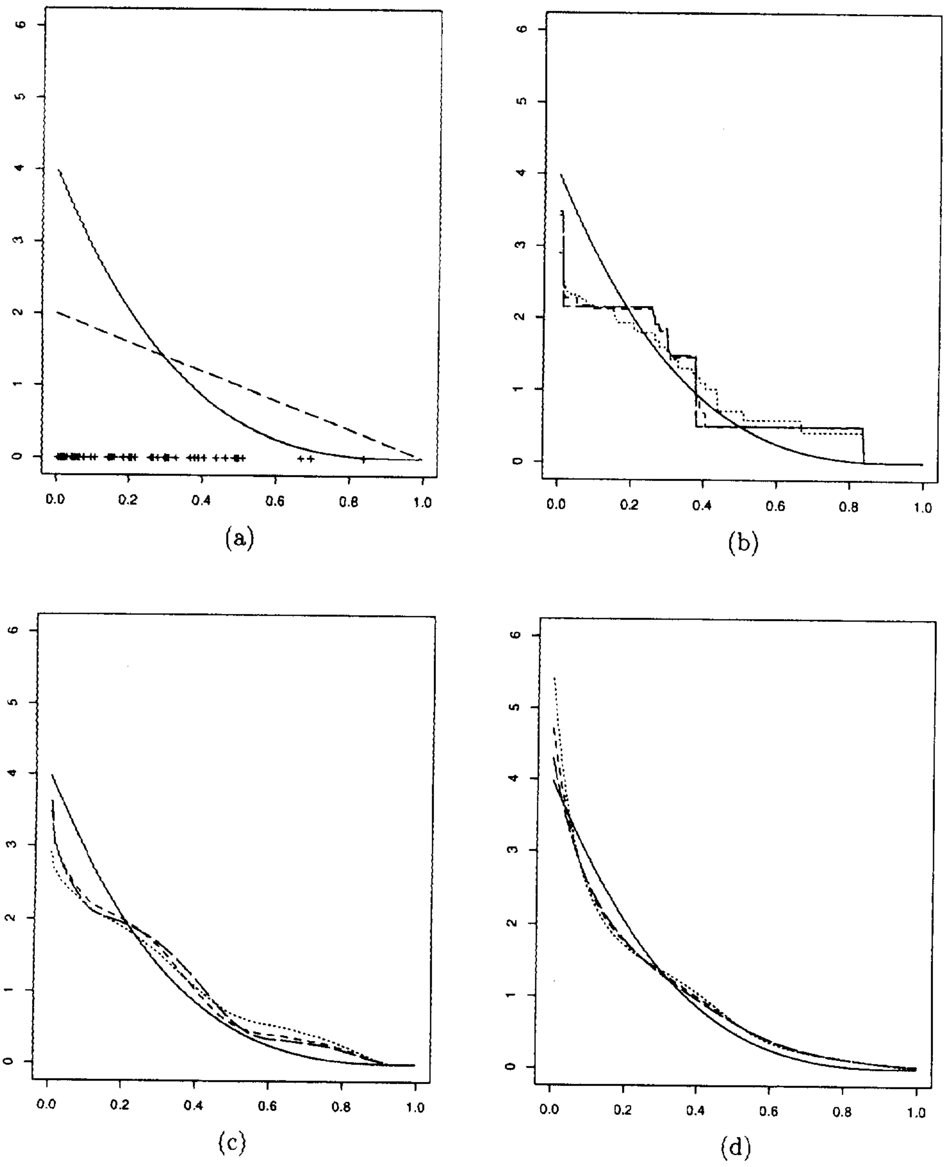


*Figure 2. Comparison of monotone decreasing density estimates using recursion with no smoothing (panel (b)), recursion with smoothing (panel (c)), and local log-linear methods (panel (d)). Panel (a) depicts the true f (unbroken line) and template g (broken line).*

(On the scale to which panels (b) and (c) are drawn, the functions $\hat{f}_j$ for $j \geq 10$ are indistinguishable from that with $j = 9$.) The bandwidth, $\hat{h}$, was chosen by cross-validation, and the Epanechnikov kernel was used for the computations described at (2.1).

Panel (d) of Figure 1 shows three kernel estimates, using the bandwidths $\hat{h}_{\text{crit}}$, .8 $\hat{h}_{\text{crit}}$, and 1.2 $\hat{h}_{\text{crit}}$. Here, following the methodology of Silverman's (1981) test for unimodality, $\hat{h}_{\text{crit}}$ is the smallest bandwidth that gives a unimodal density estimate. (In the simulation studies in this article, $\hat{h}_{\text{crit}}$ was almost always larger than the bandwidth that minimized mean squared error, which is why we use it throughout. When the true density was compactly supported, as in the case of Figure 1, we defined $\hat{h}_{\text{crit}}$ by counting the number of modes on the support.) The Standard Normal kernel was used, again as suggested by Silverman (1981). The kernel estimates in panel (d) have difficulty producing a unimodal approximation to $f$ without degrading its peak.

Figure 2 illustrates the same method in the case of a monotonicity constraint. There, panel (a) depicts the true density $f$ and the template $g$, being respectively the Beta (1, 4) density and $2(1-x), 0 < x < 1$. Both functions are monotone decreasing on $\mathcal{I}$. Panel (a) also shows the dataset, again of size $n = 50$. Panels (b)–(d) illustrate respectively the unsmoothed recursive estimator, the smoothed recursive estimator, and local log-linear methods (see, e.g., Simonoff 1996, secs. 3.3.1 and 3.4).

# 3. NUMERICAL PROPERTIES

## 3.1 SUMMARY OF PROPERTIES

Extensive simulation studies, which we summarize in Sections 3.2 and 3.3, show that the recursive method performs particularly well, without any vices, in cases where either (a) the density is unimodal and vanishes at the ends of its support (which may be finite or infinite); or (b) the density is monotone and vanishes at one end of its support (finite or infinite). In these cases, choice of template has little effect on the final estimate, provided the template is unimodal and compactly supported in case (a), and monotone, compactly supported and vanishing at its lower end in case (b). The main influence of the template is on the manner in which $\hat{f}$ decreases to zero at the ends of its support, in either of the two cases. If we are in doubt about the manner in which $f$ decreases, then it is better to use a template which decreases sharply, rather than smoothly, to zero.

To appreciate this point, note that our recursive method produces an estimate of the transformation $T$ such that $f = C\, g(T)$, for a positive constant $C$. (If the true density $f$ is unimodal or monotone, and if $g$ has the same respective property, then $T$ will be uniquely defined by this identity.) Therefore, $f' = C\, g'(T)\, T'$, where the prime denotes differentiation. Suppose $f$ decreases to zero steeply, in the sense that $f'(x)$ converges to a nonzero number as $x$ converges to a point $x_f$ (one of the ends of the support of $f$) where $f$ vanishes; but that the template, $g$, decreases smoothly at the corresponding endpoint, $x_g$, of its support, in that $g'(x) \to 0$ there. Since $0 \neq f'(x_f) = C\, g'\{T(x_f)\}\, T'(x_f)$, and $g'\{T(x_f)\} = g'(x_g) = 0$, then $T'(x_f)$ must be infinite. This will be evidenced by relatively poor performance of the estimate, as it

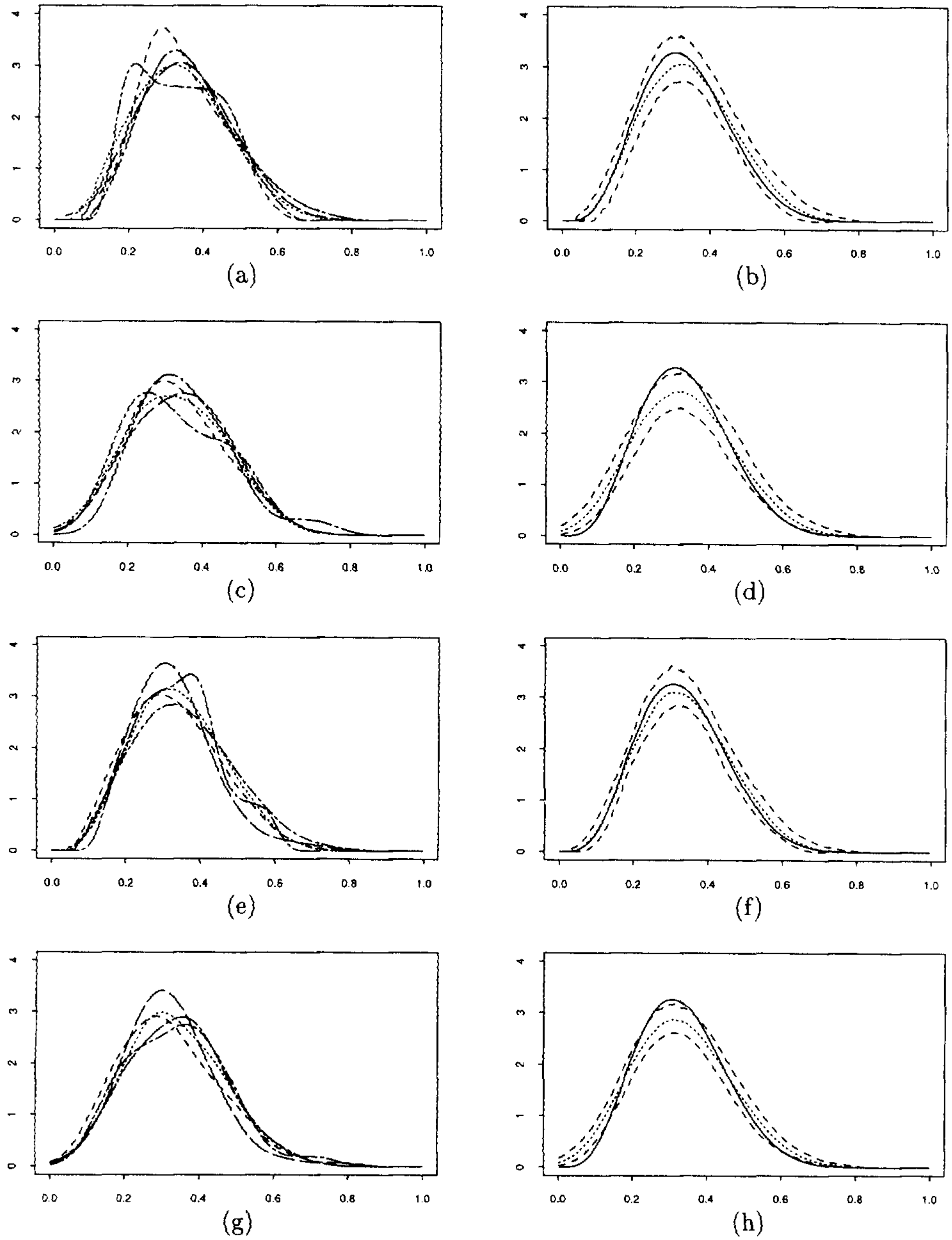


*Figure 3. Effect of sampling variation in smoothed recursive estimates, and comparison with kernel estimates, under the constraint of unimodality. Panels in the first two rows represent sample sizes $n = 50$, while those in the last two rows address $n = 100$. Panels in the first and third rows illustrate estimates obtained by recursion, while those in the second and fourth rows were obtained using kernel methods. Panels (a), (c), (e), and (g) depict estimates computed from five typical samples. Each of the other panels shows values of upper 10% points, lower 10% points and median (all in a pointwise sense) of 199 independent estimates. The true density $f$ is also depicted in these panels; it is represented by the unbroken line. Both $f$ and the template $g$ are as in Figure 1.*

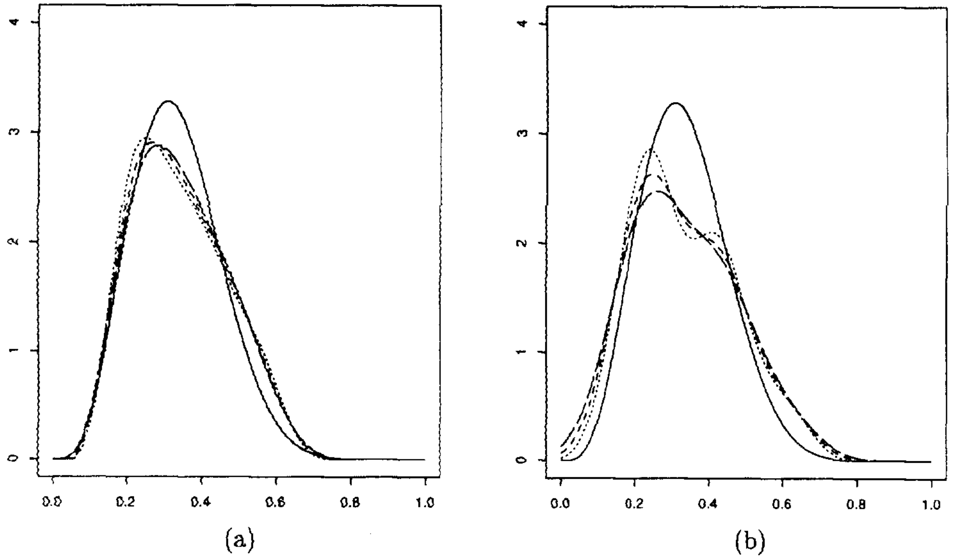


*Figure 4. Effect of bandwidth variation on density estimates constructed using smoothed recursion (panel (a)) and kernel methods (panel (b)). Solid lines represent $f$ and broken lines show the estimates. Sample size is $n = 50$. Both $f$ and $g$ are as in Figure 1.*

struggles to approximate a transformation which has unbounded slope. See the second row of three panels in Figure 8. This problem vanishes if the template decreases steeply; see Figure 7.

In cases which do not come under (a) and (b) above, the sequence $\hat{f}_1, \hat{f}_2, \ldots$ will converge to an accurate estimate of $f$ provided the values of $f$ and $g$ are in the same proportion at their *key points*. We define these points to be the ends of the support, and the mode (when the constraint is of unimodality). For example, if $f$ is unimodal and supported on the compact interval $[x_1, x_2]$, with its mode at $x_3$; and if the same is true of $g$, this time with $x_i$ replaced by $y_i$; then the key points of $f$ and $g$ are $x_1, x_2, x_3$ and $y_1, y_2, y_3$, respectively. We require the existence of a constant $C > 0$ such that $f(x_i) = C\,g(y_i)$ for $i = 1, 2, 3$. If $f$ and $g$ are monotone decreasing or increasing on $[x_1, x_2]$ and $[y_1, y_2]$, respectively, then we need $f(x_i) = C\,g(y_i)$ for $i = 1, 2$.

This constraint, which we call the *alignment condition*, is trivially satisfied in cases (a) and (b). That is why our method performs so well, without any bad habits, in those circumstances. When the alignment condition fails, the sequence $\hat{f}_1, \hat{f}_2, \ldots$ will still converge, with or without applying the smoothing step, but in large samples the limit will not be close to $f$. The problem may be overcome in at least two ways: by estimating the value of $f$ at its key points (e.g., using boundary kernels or locally parametric methods; see Simonoff [1996, secs. 3.3.1 and 3.4]), and choosing $g$ accordingly; or by incorporating the alignment condition through the vector $\theta$ of parameters included in the estimator $\hat{f}(x|\theta, h)$ introduced in Section 2. We have found that using a "naive" template $g$ is usually not adequate, since it tends to be too far distant from the true density $f$, and results in convergence to a density estimate that does not resemble $f$.

The subsequent sections illustrate these points, showing that the recursive method

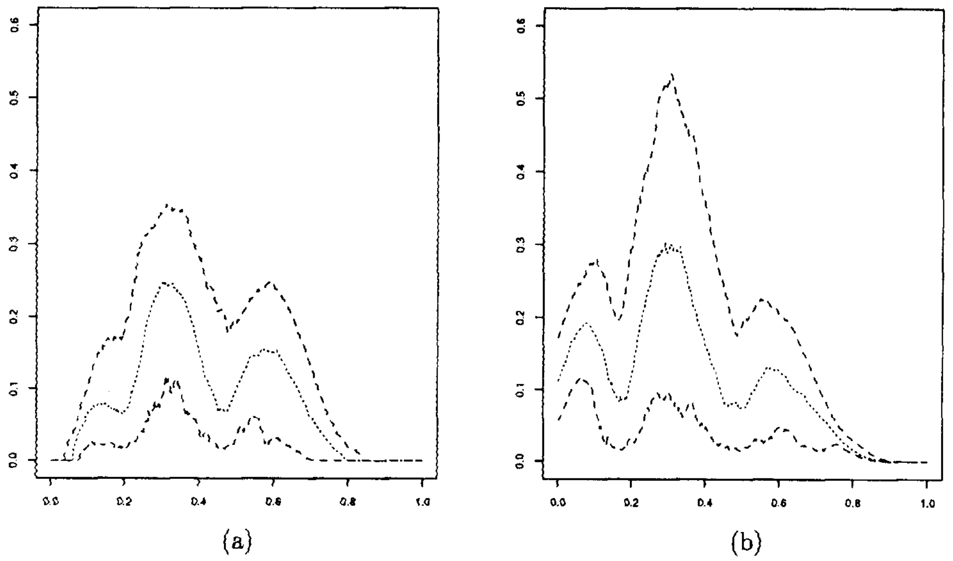


*Figure 5. Graph of 10% and 90% points, and median, of distance between suprema and infima of density estimates for bandwidths in the range* $.8\hat{h} \le h \le 1.2\hat{h}$*, as a function of* $x$*. Panel (a) addresses the case of smoothed recursion, and panel (b), the case of standard kernel estimation. Sample size is* $n = 50$*. Both* $f$ *and* $g$ *are as in Figure 1.*

provides a flexible tool for solving problems in constrained density estimation. It may be extended to a wider range of contexts, for example to estimation under the constraint that $f$ has $m$ modes, provided $m$ is known. Here the alignment condition must hold at $2m + 1$ key points (the $m$ modes, the positions of the $m - 1$ local minima between the modes, and the two extremities of the support of $f$). The recursive method may be implemented in the same way in all cases, but it requires more care when $m \ge 2$.

### 3.2 CONSTRAINT OF UNIMODALITY

Figure 1 treated the case of compactly supported $f$ and $g$ when both densities decrease smoothly to zero at the ends of their support. Figures 3–6 address the variability of these estimates. In the first four panels of Figure 3 we show, for samples of size $n = 50$, (a) smoothed recursive density estimates for five typical samples; (b) upper 10% points, lower 10% points, and the median (all in a pointwise sense) for 199 simulated smoothed recursive estimates; (c) standard kernel density estimates, using the same samples as (a); and (d) upper 10% points, lower 10% points and the median for 199 standard kernel estimates, for the same samples as (b). The next four panels, (e)–(h), of Figure 3 illustrate the same quantities for estimates computed from samples of size $n = 100$. (Throughout this article, five "typical" samples were chosen to be those that produced 20%, 35%, 50%, 65%, or 80% quantile of integrated squared error, approximated by summation, of the smoothed recursive estimator, among 199 independent samples.)

The estimates on which the figure is based represent $\hat{f}_{15}$ in the sequence of smoothed estimates $\{\hat{f}_j\}$, but are virtually identical for larger numbers of recursions. Section 3.4

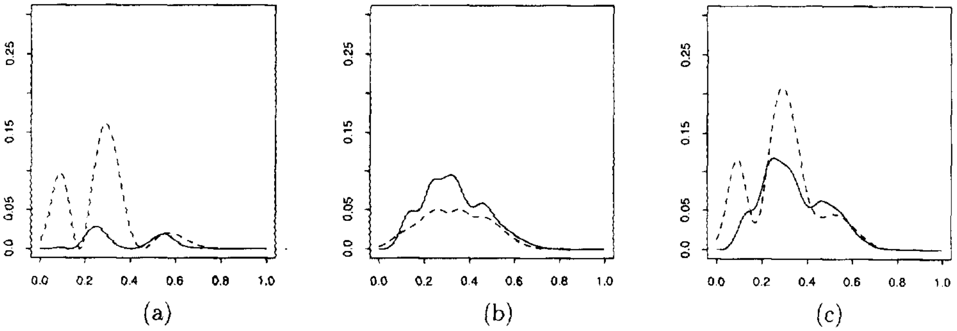


*Figure 6. Graphs of squared bias (panel (a)), variance (panel (b)), and mean squared error (panel (c)), each as a function of $x$ (on the horizontal axis). The solid line represents the case of smoothed recursive estimation, and the broken line, standard kernel estimation. Sample size is $n = 100$. Both $f$ and $g$ are as in Figure 1.*

will provide more detail about the convergence properties of $\{\hat{f}_j\}$. In panels (a), (b), (e), and (f) we chose the bandwidth by cross-validation for each data set. For panels (c), (d), (g), and (h) we used the smallest bandwidth, $\hat{h}_{\text{crit}}$, such that the density estimate was unimodal, and employed the Standard Normal kernel. In panels (a), (c), (e), and (g), each line type represents the same sample.

In some cases one might not be overly concerned by standard kernel density estimates that had small additional modes. Then one might be content to select the bandwidth by cross-validation, for example, rather than use $\hat{h}_{\text{crit}}$. However, this approach penalizes standard kernel methods in other ways, since they are particularly susceptible to variation in the smoothing parameter. By way of comparison, the unimodality constraint imposed by the smooth template has the effect of providing substantial resistance to variation in the smoothing parameter. The constraint confers a degree of smoothness that has to be supplied in another way in more general problems of density estimation. In particular, the number of "degrees of freedom" of movement available to a unimodal density estimator is greatly restricted, relative to a standard kernel estimator.

To illustrate this point, Figure 4 depicts different curve estimates computed from the same data set used to produce panels (b)–(d) of Figure 1. Both panels show estimates computed for the bandwidths $h = .8\hat{h}, 1.0\hat{h}, 1.2\hat{h}$, where the curves in panel (a) are derived by smoothed recursion, and have $\hat{h}$ equal to the cross-validation bandwidth; and the curves in panel (b) are derived using the standard kernel method, and have $\hat{h}$ equal to the critical bandwidth $\hat{h}_{\text{crit}}$. The estimates based on smoothed recursion are seen to be substantially less variable, as functions of bandwidth, than those founded on standard kernel methods.

Corroborating this evidence, Figure 5 graphs the upper 10% points, lower 10% points, and median, as functions of $x$, of the values of

$$D(x) \equiv \sup_{.8\hat{h} \le h \le 1.2\hat{h}} \hat{f}(x|h) - \inf_{.8\hat{h} \le h \le 1.2\hat{h}} \hat{f}(x|h)$$

computed from the 199 independent samples used to produce panels (b) and (d) of Figure 3. Panel (a) of Figure 5 is for the case of smoothed recursion, and panel (b) for the case of standard kernel estimation. Interpretation of $\hat{h}$ is the same as in Figure 4. As

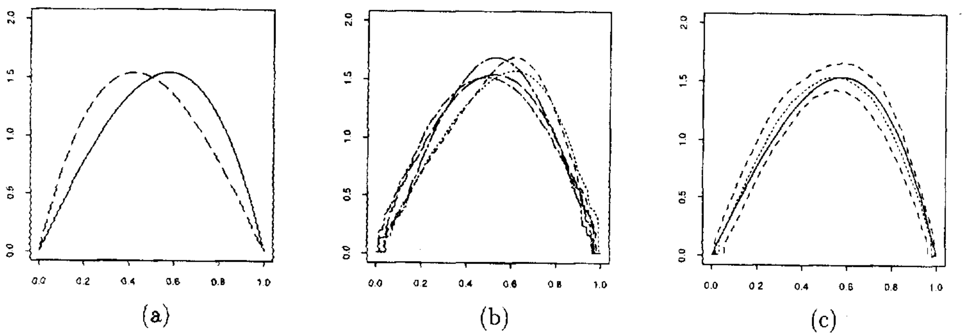


*Figure 7. Illustration of smoothed recursion when $f$ and $g$ both decrease steeply at the ends of their support. The functions $f$ and $g$ are depicted in panel (a). Panel (b) shows five typical density estimates for sample size $n=100$, with bandwidths chosen by cross-validation. Panel (c) shows 10% and 90% points, and the median (all in a pointwise sense), of 199 independently sampled, smoothed recursive estimates. The true density $f$ is also depicted in panel (c).*

in the case of Figure 4 it is seen that smoothed recursion estimators are substantially less sensitive to variation in the bandwidth than are standard kernel estimators.

Figure 6 presents plots of squared bias (panel (a)), variance (panel (b)), and mean squared error (panel (c)), as a function of position, $x$, for estimators computed using smoothed recursion or standard kernel methods. The density $f$ and template $g$ were as in Figure 3, as too was the method of bandwidth choice. Each expected value was calculated as the average of 199 simulated values of its argument. It is clear from the figure that, for most values of $x$, the estimator computed using smoothed recursion enjoys lesser bias and mean squared error than the standard kernel estimator, constrained to be unimodal. The kernel estimator has lesser variance, but this is because an overly large bandwidth has had to be chosen to enforce unimodality, and as a result the kernel estimator suffers more serious bias problems.

Figures 3–6 have addressed the density illustrated in Figure 1, using the template illustrated there. They show that our smoothed recursion estimators improve on kernel estimators in terms of sensitivity to bandwidth choice as well as more conventional measures of statistical performance. A substantially more complex kernel method, based on local bandwidth choice, might be competitive with smoothed recursion. It would be difficult to implement automatically in a simulation study, however.

Next we examine three other types of $f$ and $g$: compactly supported densities, when both functions decrease steeply to zero (Fig. 7); compactly supported $f$ and $g$ when one but not the other decreases steeply to zero (Figure 8); and infinitely supported $f$ and compactly supported $g$ decreasing smoothly to zero (Figures 9 and 10). Throughout Figures 7–10 the bandwidth was chosen by cross-validation.

Figures 3 and 7 show that, provided the template is chosen to match the way in which $f$ decreases at the ends of its support, the smoothed recursive density estimate captures the properties of $f$ very well. Even in cases where the template decreases steeply to zero, but $f$ decreases smoothly, smoothed recursion accurately approximates $f$; see panels (a)–(c) of Figure 8. However, the estimate can behave poorly if the ends of the template decrease to zero substantially more smoothly than those of $f$; see panels (d)–(f)

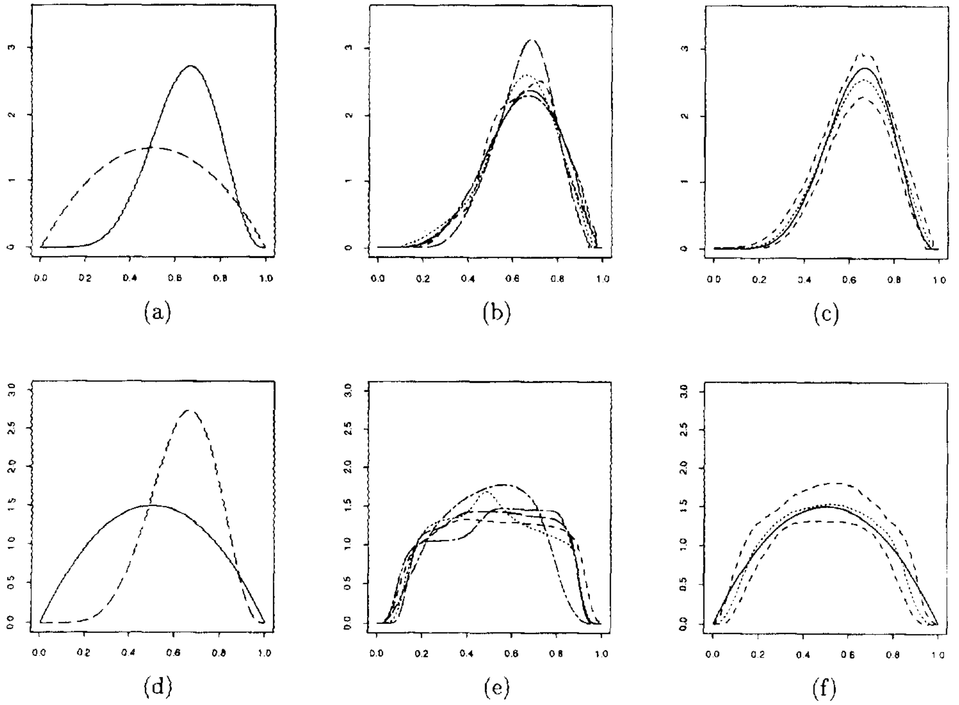


*Figure 8. Illustration of smoothed recursion when one of $f$ and $g$, but not the other, decreases steeply at the ends of its support. Panel (b) depicts five typical density estimates for the functions $f$ and $g$ illustrated in panel (a), and panel (c) graphs the 10% and 90% points, and the median (all in a pointwise sense), of 199 independently sampled estimates. Panels (e) and (f) depict the same quantities as panels (b) and (c), respectively, except that now $f$ and $g$ are as in panel (d). Sample size is $n$=100.*

of Figure 8. "When in doubt, use a template that decreases sharply at the ends of its support."

Figures 9 and 10 address the performance of smoothed recursive estimates, and kernel estimates, computed from samples drawn from Student's $t$ distribution with 5 degrees of freedom. The template in the smoothed recursive case was the Beta (3,4) density, although similar results are obtained with a template that decreased steeply to zero at the ends of its support. The two rows in Figure 9 illustrate the case of smoothed recursive and kernel estimators, respectively. Panels (a) and (c) in that figure depict five typical estimates, while panels (b) and (d) show 10% and 90% points, as well as the median, of 199 estimates. To compute the results illustrated in panels (a) and (b) the density was assumed to have support a little larger than the range of the data. In effect, we transformed this interval to $\mathcal{I} = [0, 1]$ and then applied the method used to produce earlier figures for data from compactly supported $f$'s. It is not necessary to do this explicitly, however, since it is automatically accomplished by the algorithm described in Section 2.

Panels (c) and (d) of Figure 9 were computed using kernel density estimates with bandwidth $\hat{h}_{\text{crit}}$. (In the case of the $t_5$ density we defined $\hat{h}_{\text{crit}}$ by counting the number

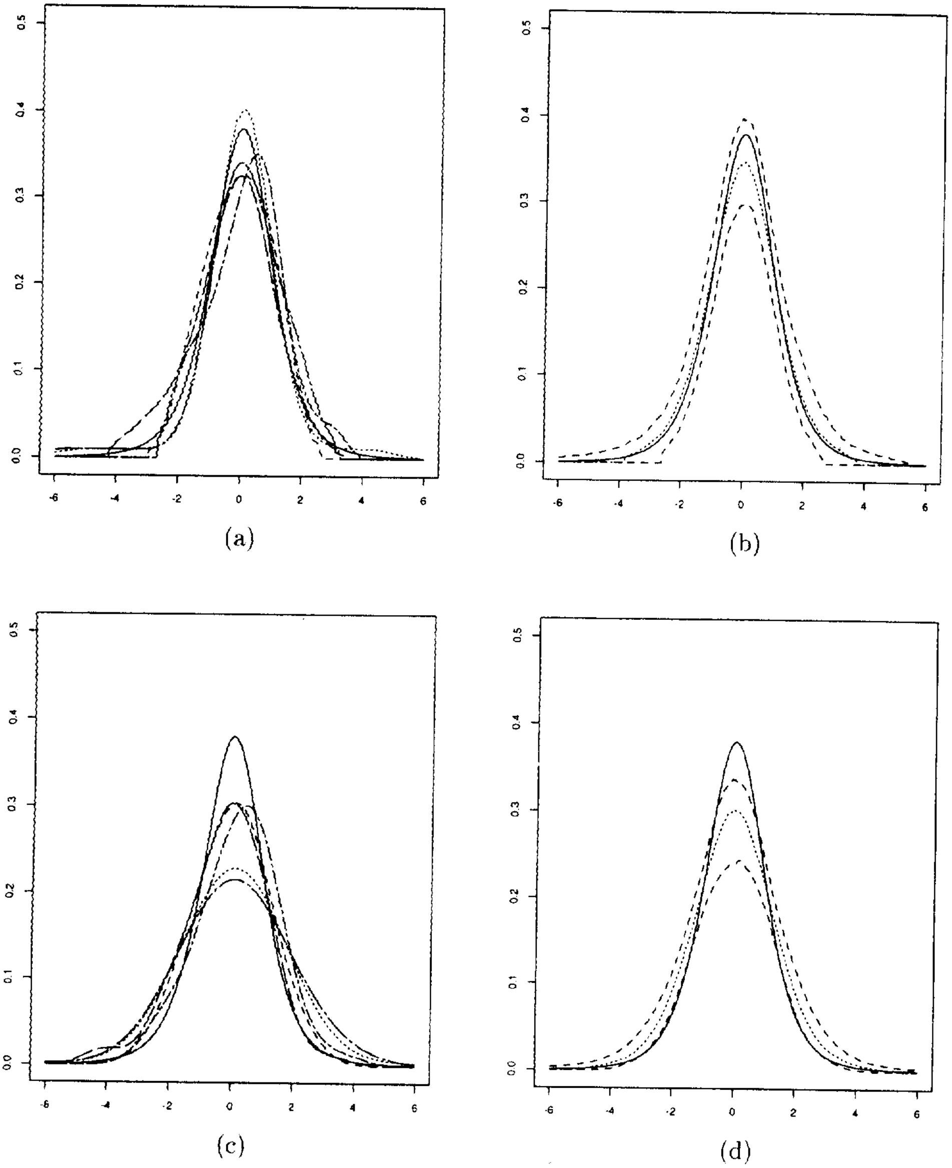


*Figure 9. Illustration of smoothed recursion (panels (a) and (b)) when f (depicted by the solid line in panel (a)) is infinitely supported, in the case of a unimodality constraint. For computational purposes the support of f was taken to be slightly larger than that of the data. Panel (a) depicts five typical estimates, and panel (b) graphs 10% and 90% points, and the median, of 199 estimates. Panels (c) and (d) depict the analogous results for kernel estimates (shown as broken lines in panel (c)), using the bandwidth $\hat{h}_{\text{crit}}$ and based on the same samples. Sample size is n=100.*

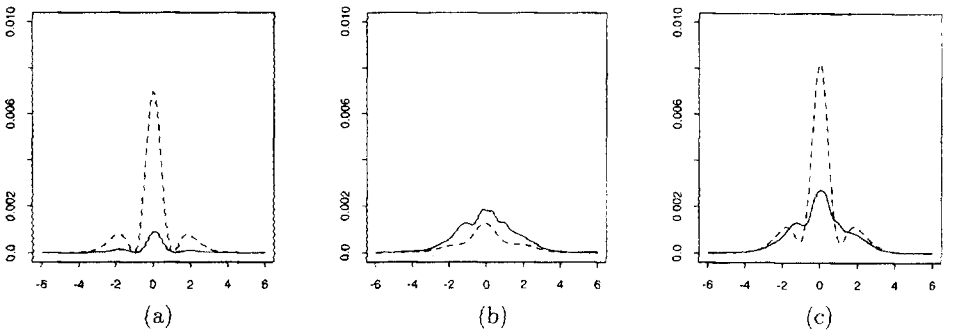


*Figure 10. Graphs of squared bias (panel (a)), variance (panel (b)), and mean squared error (panel (c)), each as a function of $x$ (on the horizontal axis). Both $f$ and $g$ are as in Figure 9, and sample size is $n$=100. The solid line represents the case of smoothed recursive estimation, and the dashed line, standard kernel estimation.*

of modes on the interval $[-6, 6]$. If we were to use the whole real line, then the density estimate would be much flatter in shape.) Here, outlying data values have forced the use of a very large bandwidth in order to ensure unimodality, with the result that the shape of the estimate at the mode is seriously in error. The inaccuracy of kernel estimators is also demonstrated by Figure 10, which plots squared bias (panel (a)), variance (panel (b)), and mean squared error (panel (c)), as a function of position, $x$, for estimators computed using smoothed recursion or standard kernel methods. All parameter settings were the same as in Figure 9. Panel (c) of Figure 10 starkly illustrates the accuracy problems experienced by kernel methods when the bandwidth is chosen so as to enforce unimodality. As in the case of Figure 6, such a bandwidth produces very low variance, but at the expense of particularly high bias.

### 3.3 CONSTRAINT OF MONOTONICITY

Figure 11 illustrates the variability of our recursive estimates under the constraint of monotonicity. The density $f$ and template $g$ are as in Figure 2. The curves in panels (a) and (e) represent the smoothed version of $\hat{f}_{15}$ for five typical samples of sizes $n = 50$ and 100, respectively. In each case, bandwidths were chosen by cross-validation. Panels (c) and (g) depict local log-linear, kernel-type locally parametric estimates (e.g., Hjort and Jones 1996; Loader 1996; Simonoff 1996, p. 64ff) for the samples used in panels (a) and (e), respectively. This choice was made because it gives better performance at the ends of the interval than standard kernel methods. Panels (b) and (f) (in the case of smoothed recursion) and (d) and (h) (for local log-linear estimation) show values of upper 10% points, lower 10% points, and the median, all in a pointwise sense, for 199 independent estimates. To obtain the local-linear results (i.e., panels in the second and third rows) we used the Standard Normal kernel and bandwidths that are optimal in a mean integrated squared error sense. (Therefore, the bandwidth did not vary from one sample to another.)

The variability of smoothed recursive estimators with changing bandwidth, relative to kernel-type estimators, is visually even less in the case of monotonicity than it was

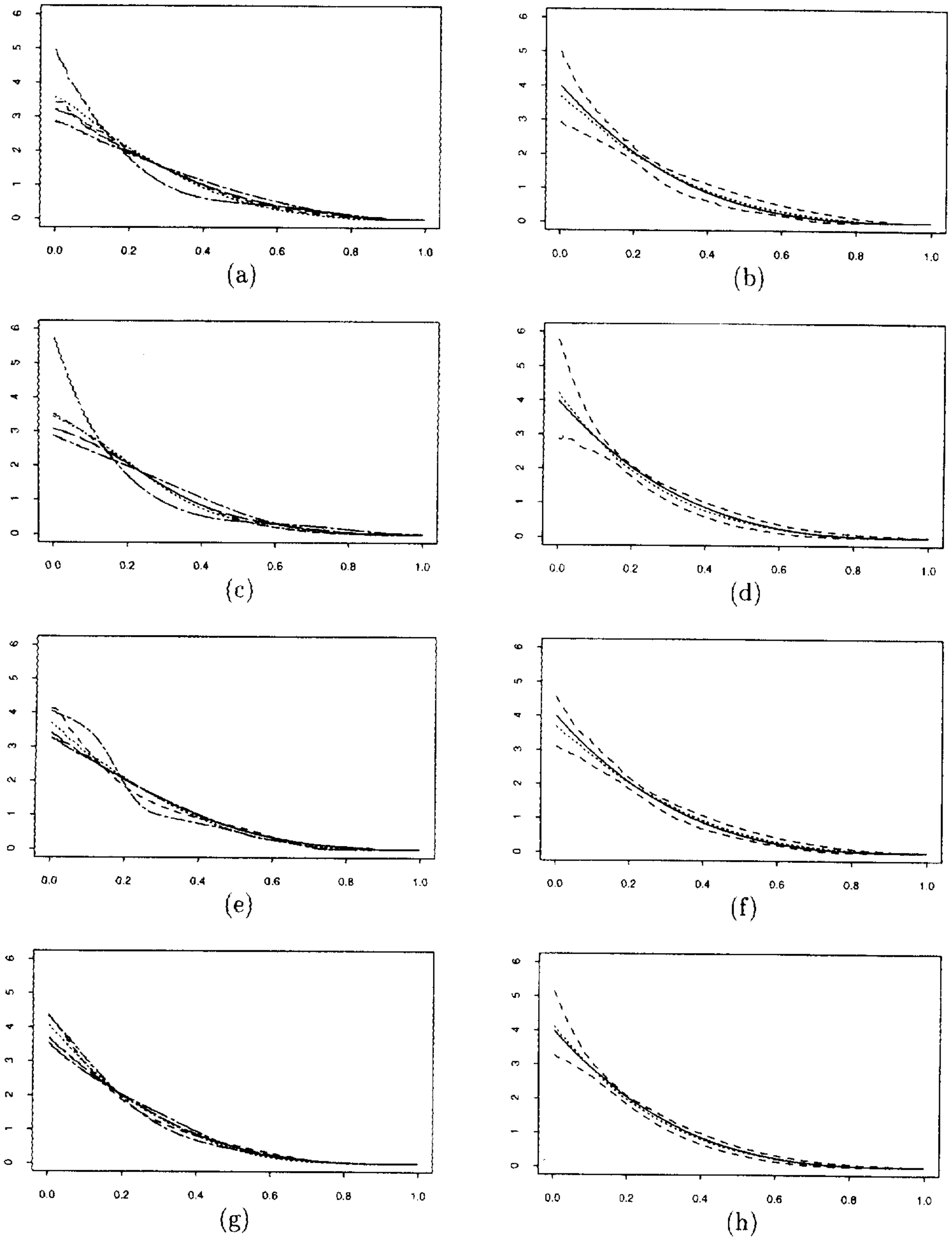


*Figure 11. Effect of sampling variation in smoothed recursive estimates, and comparison with kernel estimates, under the constraint of unimodality. Panels are as for Figure 3, except that now the densities $f$ and $g$ are as in Figure 2, and in place of standard kernel estimation we used local log-linear (locally parametric) methods.*

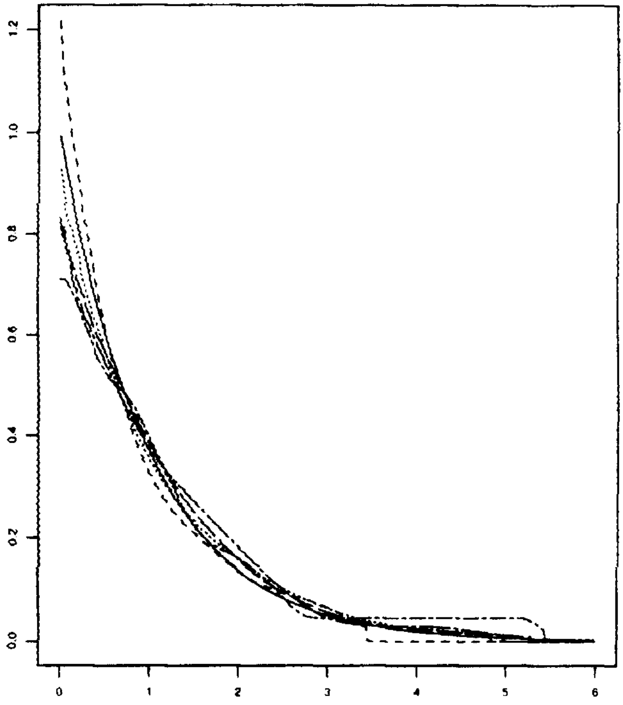


*Figure 12. Illustration of smoothed recursion when $f$ (solid line) is infinitely supported, in the case of a monotonicity constraint. The upper right-hand end of support of $f$ was taken to be a little larger than the largest order statistic.*

under the constraint of unimodality. This is apparently due to the fact that the second derivative of the true density vanishes, and so an unconstrained, kernel-type estimator is less resistant to visually obvious perturbations in the zeroth and first derivatives. Moreover, the influence of the manner of decrease of the template on that of the density estimator $\hat{f}$ is the same under monotonicity as it was in the case of unimodality. That is to say, if the template decreases gently to zero but the true density decreases steeply, then our smoothed recursive method performs poorly. The reason is the same as in the case of unimodality—for a slowly decreasing $g$ and a steeply decreasing $f$, the transformation $T$ that we are trying to estimate has infinite derivative, and so cannot be approximated accurately. For the sake of brevity we do not illustrate these properties here. Note, however, that they suggest the use of a linear template.

Figure 12 illustrates the case of estimating a monotone density with infinite support, using a template with finite support. The true density $f$ was exponential with unit mean, and the template, $g$, was identical to that in Figure 2. The method of computation was the monotone-template version of that employed for Figure 9, in that we took the support of $f$ to be the interval between the left-hand end of the true support and a point slightly larger than the largest data value. This approach assumes that the left-hand end of the support is known, which would usually be the case in practice. Alternatively, one may replace the left-hand endpoint by that data value which is furthest to the left. This approach works well in practice. In each case, the smoothed version of $\hat{f}_{15}$ is shown, with the bandwidth

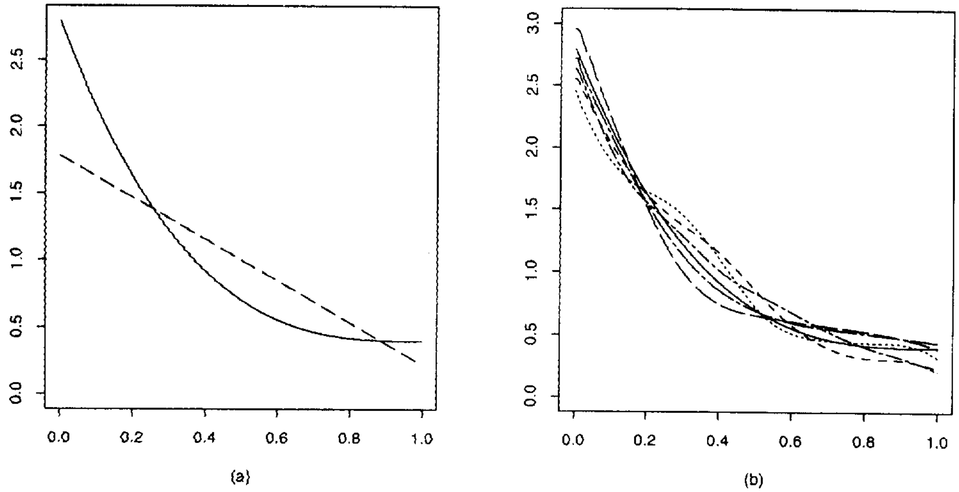


*Figure 13. Illustration of methods for achieving the alignment condition when it is not trivially satisfied. Panel (a) depicts the true $f$ (unbroken line) and template $g$ (broken line, for a particular value of its parameter $\theta$). Panel (b) shows smoothed recursive density estimates computed using the cross-validation method at (2.2).*

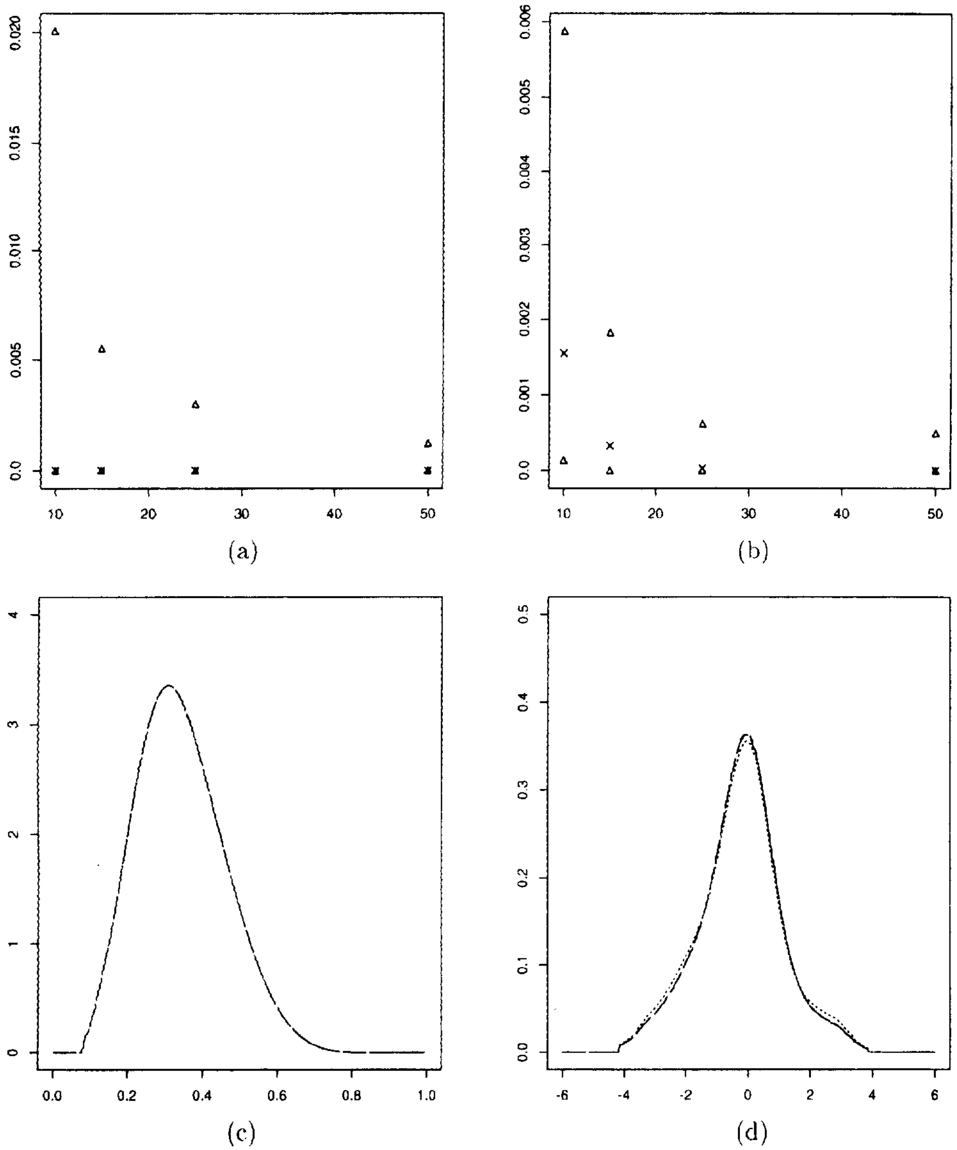
chosen by cross-validation and with sample size $n = 100$.

When the true, monotone density $f$ does not take the value 0 at one of the ends of its support, care has to be taken to ensure that the alignment condition holds. Figure 13 illustrates one of the ways suggested in Section 3.1 for overcoming this problem. Throughout, we took the true density to be $f(x) = 2.4\ (1-x)^3 + .4$ for $0 < x < 1$, and the template $g$ to be linear, defined by $g(0) = a_1$, $g(1) = a_2$ and $\int_{\mathcal{I}} g = 1$. These curves are shown in panel (a). Panel (b) depicts smoothed recursive density estimates for which the ratio of the left- to the right-hand ends of the template was determined by cross-validation, using the criterion at (2.2). In each case the parameter $\theta$ was taken to be $a_1/a_2$, the smoothed version of $\hat{f}_{15}$ is shown, and sample size is $n = 500$. The first method suggested in Section 3.1, based on plugging in estimates of $f$ at the endpoints 0 and 1, usually performs less well than that employed for panel (b). This is apparently because the bias of plug-in estimators has quite a different structure from the bias of smoothed recursive estimators, and so estimates at the ends of $[0, 1]$ and estimates over the rest of the interval are, in effect, misaligned in a systematic way.

### 3.4 Convergence of Iterates

We report here the results of a simulation study exploring the rate of convergence of the iterates $\hat{f}_1, \hat{f}_2, \ldots$. For the sake of brevity we confine attention to only two cases: the compactly supported, unimodal density treated in Figures 1 and 3–6; and the density with unbounded support, featured in Figures 9 and 10. The template $g$ used in those examples was also employed here. In each case, and for each of 199 samples of size $n = 100$, we computed the supremum of the distance apart of successive density iterates $\hat{f}_j$: $D_j = \sup|\hat{f}_{j+1} - \hat{f}_j|$. Panels (a) and (b) of Figure 14 graph the 10% point, 90% point, and median of the 199 values of $D_j$, for $j = 10, 15, 25$, and 50. Panels (c) and

(d) depict the estimates $\hat{f}_5$, $\hat{f}_{10}$ and $\hat{f}_{15}$ in the case of the sample which produced the 99th largest (out of 199) value of $D_{10}$. (We do not plot $\hat{f}_j$ for larger values of $j$ since, on the scale of panels (c) and (d), they are not distinguishable from one another.) In each case the bandwidth was chosen by cross-validation. These results make it clear that convergence of $\hat{f}_j$ occurs rapidly.



(a)

(b)

(c)

(d)

*Figure 14. Rate of Convergence of Iterates $\hat{f}_j$. In the respective cases of Figures 1 and 9, panels (a) and (b) illustrate the 10% point, 90% point, and median of 199 values of the supremum of the absolute difference between $f_j$ and $f_{j+1}$, for $j = 10, 15, 25$ and 50. In the same respective contexts, panels (c) and (d) depict the estimates $\hat{f}_5$, $\hat{f}_{10}$, and $\hat{f}_{15}$ for a "typical" sample, chosen as stated in the text.*

## 4. THEORETICAL RESULTS

Here we outline theoretical evidence for the numerical properties noted in Sections 2 and 3. We simplify the problem by treating the case of no noise. Thus, we suppose we are given a "true" density $f$, and a template density $g$, both of them supported on $\mathcal{I}$. Compute sequences $f_1, f_2, \ldots$ and $g_1, g_2, \ldots$ recursively, as follows. Suppose we have constructed the density $f_j$. (For the first step, we take $f_1 = g$.) Let $F$ and $F_j$ be the distribution functions corresponding to $f$ and $f_j$, respectively; define $T_j$ to be the mapping from $\mathcal{I}$ to $\mathcal{I}$ represented by $F_j^{-1}(F)$; and define $f_{j+1}$ to be the density that is proportional to $f_j(T_j)$. That is, $f_{j+1}$ is proportional to $g_j = g(U_j)$, where $U_j = T_1 \ldots T_j$. We wish to show that, as $j \to \infty$, $U_j \to T$, where $T$ is the unique monotone transformation such that $f$ is proportional to $g(T)$. (There will exist a unique $T$ provided $f$ and $g$ satisfy the alignment condition introduced in Section 3.1.)

We are aware of only one case where $f_j$ and $g_j$ may be written down explicitly; this is when $f$ and $g$ are both Beta $(\gamma, 1)$ or both Beta $(1, \gamma)$ densities. Define $h_\gamma(x) = \gamma x^{\gamma-1}$ for $x \in \mathcal{I}$, and let $f = h_\alpha$ and $g(x) = h_\beta$, where $\alpha, \beta > 0$. Then, $g\{F_1^{-1}(F)\}$ equals a constant multiple of $h_\gamma$, where $\gamma = \alpha + 1 - (\alpha/\beta)$. It follows that successive applications of the method suggested in Section 2 produce probability densities $f_1 = h_{\beta_1}, f_2 = h_{\beta_2}, \ldots$, where $\beta_j - \alpha = (\beta - \alpha)/(\beta\beta_1 \ldots \beta_{j-1})$ and $\beta_0 = \beta$. Therefore, successive iterates $f_j = h_{\beta_j}$ converge to $f$ at an exponentially fast rate. (The numerical work reported in Section 3.4 suggests that more generally, the rate of convergence may be only polynomially fast.)

Without the assumption that $f$ and $g$ are gamma densities, some progress may be made in the case where the densities $f$ and $g$ are uniformly close and are supported on $\mathcal{I}$. To this end, allow either or both of $f$ and $g$ to depend on the small positive constant $\epsilon$, and be such that

$$g = f + \epsilon\,\psi + o(\epsilon) \tag{4.1}$$

uniformly on $\mathcal{I}$, as $\epsilon \to 0$, where $\psi$ is a bounded function not depending on $\epsilon$. Then, the $j$-fold recursion of our horizontal alignment method takes $g$ to

$$g_j = g(U_j) = f + \epsilon\, L^j(\psi) + o(\epsilon)\,,$$

uniformly on $\mathcal{I}$ as $\epsilon > 0$, where $L$ is the linear operator defined by $L(\psi) = \psi - (\Psi - F\,\Psi_1)\,(f'/f)$, $\Psi(x) = \int_{[0,x]} \psi(y)\,dy$ and $\Psi_1 = \Psi(1)$.

We would like to be able to prove that under general conditions on $f$ and $\psi$, $L^j(\psi) \to 0$ as $j \to \infty$, for all $\psi$ except constant multiples of $f$. This result, which we shall call (R), would go some way towards establishing convergence more generally, although it is not within reach at present. We are able to prove the following, however. First, if $f'$ vanishes at no more than a finite number of isolated points then the only eigenvector of $L$ corresponding to eigenvalue $\lambda = 1$ is $f$ itself (up to constant multiples). If in addition $f$ is bounded away from 0 on $\mathcal{I}$ then there exist no eigenvalues other than $\lambda = 1$. For a finite-dimensional space this would be more than enough to establish result (R). If $f$ is linear then $L$ is a strict contraction on the set $\mathcal{B}$ of all functions $\psi$ that have a bounded derivative and satisfy $\int \psi = 0$ and $\psi(0)/f(0) = \psi(1)/f(1)$. (The latter condition is simply the assumption that $f$ and $g$ are in the same proportion at endpoints, which was assumed in our simulation study. The condition $\int \psi = 0$ follows from (4.1) if $f$ and $g$ are both densities.) A strict contraction means that there exists a norm $\|\cdot\|$ on $\mathcal{B}$ such that $\|L(\psi)\| < \|\psi\|$ for all $\psi \in \mathcal{B}$ except $\psi = 0$. It implies that $L^j(\psi) \to 0$ for all $\psi \in \mathcal{B}$, and so gives result (R).

## ACKNOWLEDGMENTS

We are grateful for the helpful, constructive comments of three reviewers and an Associate Editor.

*[Received May 1997. Revised August 1998.]*

## REFERENCES

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

Cox, D.R. (1966), "Notes on the Analysis of Mixed Frequency Distributions," *British Journal of Mathematical and Statistical Psychology*, 19, 39–47.

Fraser, D.A.S., and Massam, H. (1989), "A Mixed Primal-Dual Bases Algorithm for Regression Under Inequality Constraints. Application to Convex Regression," *Scandinavian Journal of Statistics*, 16, 65–74.

Grenander, U. (1956), "On the Theory of Mortality Measurement, II," *Skandinavisk Aktuarietidskrift*, 39, 125–153.

Hartigan, J.A., and Hartigan, P.M. (1985), "The DIP Test of Unimodality," *The Annals of Statistics*, 13, 70–84.

Loader, C.R. (1996), "Local Likelihood Density Estimation," *The Annals of Statistics*, 24, 1602–1618.

Hjort, N.L., and Jones, M.C. (1996), "Locally Parametric Nonparametric Density Estimation," *The Annals of Statistics*, 24, 1619–1647.

Mammen, E., Marron, J.S., and Fisher, N.I. (1992), "Some Asymptotics for Multimodality Tests Based on Kernel Density Estimates," *Probability Theory and Related Fields*, 91, 115–132.

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

Müller, D.W., and Sawitzki, G. (1991), "Excess Mass estimates and Tests for Multimodality," *Journal of the American Statistical Association*, 86, 738–746.

Qian, S. (1994), "Generalization of Least-Square Isotonic Regression," *Journal of Statistical Planning and Inference*, 38, 389–397.

Robertson, T., Wright, F., and Dykstra, R. (1988), *Order Restricted Inference*, New York: Wiley.

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

Shi, N.-Z. (1995), "The Minimal $L_1$ Isotonic Regression," *Communications in Statistics* Series A, 24, 175–189.

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

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

Sun, J., and Woodroofe, M. (1996), "Adaptive Smoothing for a Penalized NPMLE of a Non-increasing Density," *Journal of Statistical Planning and Inference*, 52, 143–159.

Tantiyaswasdikul, C., and Woodroofe, M. (1994), "Isotonic Smoothing Splines Under Sequential Designs," *Journal of Statistical Planning and Inference*, 38, 75–87.

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

Wang, Y. (1995), "The L1 Theory of Estimation of Monotone and Unimodal Densities," *Journal of Nonparametric Statistics*, 4, 249–261.

Woodroofe, M., and Sun, J. (1993), "A Penalized Maximum Likelihood Estimator of $f(0+)$ When $f$ is Non-increasing," *Statistica Sinica*, 3, 501–515.