Statistica Sinica **22** (2012), 681-701
doi:http://dx.doi.org/10.5705/ss.2010.309

NONPARAMETRIC ESTIMATION OF A SMOOTH DENSITY WITH SHAPE RESTRICTIONS

Mary C. Meyer

Colorado State University

Abstract: Given assumptions about shape and smoothness, a density is estimated non-parametrically using regression splines. Examples of shapes include decreasing, decreasing and convex, and unimodal with known mode. A least-squares criterion is used, so that the estimate is obtained with a single projection onto a convex cone. The convergence rates for the estimators are derived. For the case of unknown mode, a plug-in estimator may be used. If the mode estimator converges fast enough, the rate of the plug-in estimator is the same as for the known-mode estimator. Simulations show that, for small samples, the proposed estimators compare well with competing estimators.

Key words and phrases: Decreasing density, unimodal density, shape restrictions, cone projection, weighted least squares.

1. Introduction and Background

Density estimation using shape assumptions has long been of interest. The maximum likelihood estimator (MLE) for a non-increasing density was proposed by Grenander (1956): it is a step function with jumps allowed only at the observations; computation is straight-forward using the pooled adjacent violators algorithm (PAVA). It is known that the MLE is consistent on intervals not containing the mode (Prakasa Rao (1969)), but there is a spiking problem at the mode, where the estimator is too large. A maximum likelihood estimate using a penalty term for the value at the mode was proposed by Woodroofe and Sun (1993), who showed that the penalized MLE is consistent everywhere. The penalized MLE is also a step function, and can be computed using PAVA after the observations are slightly shifted.

The discontinuities of these estimators may be thought unsatisfactory if a smoothness assumption is warranted. Bickel and Fan (1996) considered the linear spline decreasing density estimator, with knots at the observations. The PAVA algorithm is again used, with observations shifted to be at the midpoints of the original observations. The estimator is continuous but there are still flat spots, and this version is again inconsistent at the mode. A consistent linear spline

decreasing density estimator was proposed by Meyer and Woodroofe (2004); the estimator is forced to be concave on an interval containing the mode. If the concave assumption is known to be valid over a given interval, then this estimator requires no user-defined penalty or smoothing parameter. Otherwise, the interval of concavity can be used as a penalty device and is allowed to go to zero as n increases.

The decreasing and convex density estimator was considered by Groeneboom, Jongbloed, and Wellner (2001) and Balabdaoui and Wellner (2010). This piecewise linear estimator has a tendency to spike at the mode. Balabdaoui (2007) gave a consistent estimate of the value of the density and its derivative at the mode, based on the values near the mode.

Maximum likelihood estimation of a unimodal density with known mode can be accomplished using two decreasing estimators on either side of the mode. For the unknown mode case, the maximum likelihood estimator does not exist because the likelihood is unbounded if the mode is allowed to vary. A plug-in estimator may be used (Bickel and Fan (1996)), or the alternative unimodal estimator of Meyer (2001) in which both sides are estimated at once using a partial ordering. If the density is log-concave then it is unimodal, so log-concavity might be a useful surrogate for unimodality. In the case of unknown mode, the log-concave estimator has the advantage of not having to specify the mode, but if the mode is known to be zero, for example, this cannot be specified. Recent work on log-concave density estimation was done by Pal, Woodroofe and Meyer (2007), Rufibach (2007), and Dümbgen and Rufibach (2009, 2011). Such standard unimodal densities as normal, Beta, and Gamma densities are log-concave; however, a mixture of normal random variables is not in general log-concave, nor are heavy-tailed densities such as the Cauchy or $t(2)$.

Several smoothed unimodal estimators have been proposed using kernel ideas. Fougères (1997) proposed an estimator that attains unimodality by a "rearrangement" of the kernel estimator. Eggermont and LaRiccia (2000) considered the derivative of the least concave majorant of the distribution function for a kernel estimator. Mammen et al. (2001) proposed a two-step estimator that projects the unconstrained estimator onto the constraint set. Hall and Huang (2002) adjusted the kernel weights to obtain a unimodal version. Hall and Kang (2005) proposed an estimator using "data sharpening," in which the observations are moved the "least amount" necessary so that the kernel estimator is unimodal. Dümbgen and Rufibach (2009) specify a smoothed version of their log-concave density estimator.

In this paper, smoothed shape-restricted density estimators are proposed for decreasing, unimodal, and convex densities. They are constructed using regression splines, so that the fit is a linear combination of smooth basis functions.

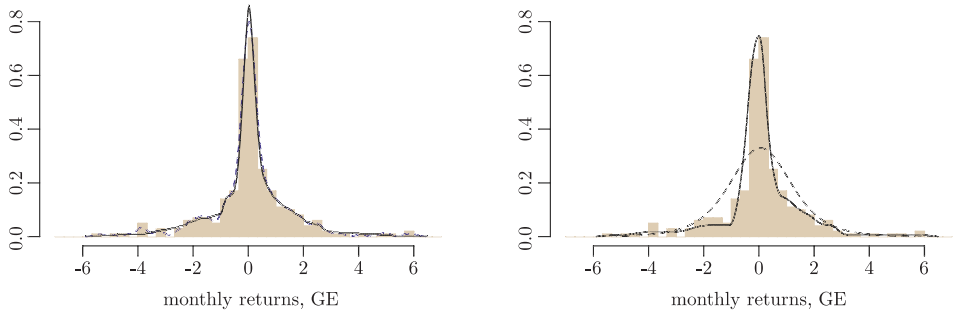

Figure 1. Monthly stock returns data from General Electric. Left: the standard kernel (dashed) and the Fougères rearrangement (solid). Right: the smoothed log-concave estimate (dashed) and the proposed unimodal spline estimate (solid).

They have the advantage of being parsimonious in the sense that the number of basis functions is small compared with the sample size n, but they are flexible enough to capture a wide variety of density forms. The least-squares criterion proposed by Groeneboom, Jongbloed, and Wellner (2001) is minimized over the set of linear combinations of basis functions, with the coefficients constrained to capture the shape assumptions. The estimator is obtained by a weighted projection onto a convex cone, and hence is computationally straightforward.

Examples of estimates of a unimodal density with mode at zero are shown in Figure 1, where a histogram of $n = 300$ months of stock returns from General Electric is shown. The data were downloaded from `http://finance.yahoo.com`. On the left is the standard kernel (dashed) estimate, with default bandwidth as chosen by the `R` function `density`. The Fougères rearrangement with mode zero is the solid curve, with bandwidth 80% of the default for the the kernel. The spurious bumps in the kernel estimate (typical for "heavy tailed" data) are smoothed out by the rearrangement, but the estimate is discontinuous at the mode, and its derivative is discontinuous at points where the rearrangement diverges from the original kernel estimate. On the right the proposed estimator (solid) is compared with the log-concave estimator provided by the `R` function `logConDens` (using the default parameters). Log-concavity is never a good assumption for "heavy-tailed" data.

In the next section, the sets of spline basis functions for the three shape assumptions are specified. The algorithm for the density estimator is presented in Section 3. In Section 4, results for convergence rates of the estimators are presented. Simulations results presented in Section 5 show that the estimators compare favorably to some established estimators for small to moderate-sized samples.

2. Regression Spline Basis Functions

Ramsay (1988) introduced monotone regression splines for least-squares regression and extensions. The spline basis functions are smooth monotone piecewise polynomials so that linear combinations of these basis functions, constrained to have non-negative coefficients, are again monotone. Meyer (2008) extended the method to other shape restrictions and showed that the estimator can be formulated as a projection onto a polyhedral convex cone, where the edges of the cone are the basis vectors. Here we provide basis functions for the decreasing, decreasing and convex, and unimodal density estimation. For the decreasing cases, the shape restrictions hold if and only if the linear combination of basis functions has non-negative coefficients; for the unimodal case an equality constraint is added to ensure continuity at the mode.

2.1. Decreasing case

Consider the estimation of a decreasing function on $[0, M_0]$, and define knots $0 = t_0 < t_1 < \cdots < t_k < t_{k+1} = M_0$. For quadratic splines, the $m = k + 3$ basis functions are

$$\delta_j(x) = \begin{cases} 1 & \text{for } 0 \le x < t_{j-1}, \\ 1 - \frac{(x-t_{j-1})^2}{(t_{j+1}-t_{j-1})(t_j-t_{j-1})} & \text{for } t_{j-1} \le x < t_j, \\ \frac{(t_{j+1}-x)^2}{(t_{j+1}-t_j)(t_{j+1}-t_{j-1})} & \text{for } t_j \le x < t_{j+1}, \\ 0 & \text{for } x \ge t_{j+1}, \end{cases}$$

$j = 1, \ldots, k$, plus

$$\delta_{k+1}(x) = \begin{cases} \frac{(t_1-x)^2}{(t_1-t_0)^2} & \text{for } t_0 \le x < t_1, \\ 0 & \text{for } x \ge t_1, \end{cases}$$

$$\delta_{k+2}(x) = \begin{cases} 1 & \text{for } 0 \le x < t_k, \\ 1 - \frac{(x-t_k)^2}{(t_{k+1}-t_k)^2} & \text{for } x_{j-1} \le x < x_j, \\ 0 & \text{for } x \ge t_{k+1}, \end{cases}$$

and $\delta_{k+3}(x) \equiv 1$. These span the space of piecewise quadratic functions on $[0, t_{k+1}]$. Because at each knot, there is only one basis function with non-zero derivative, and there is only one basis function that is non-zero at t_{k+1}, a linear combination is non-negative and non-increasing if and only if the coefficients are non-negative. Note that cubic basis functions are not suitable for monotone constraints, because there does not exist a set of linear constraints on the coefficients that is necessary and sufficient for monotonicity. The piece-wise quadratic non-increasing basis functions are shown in Figure 2(a) for $k = 6$ knots, marked

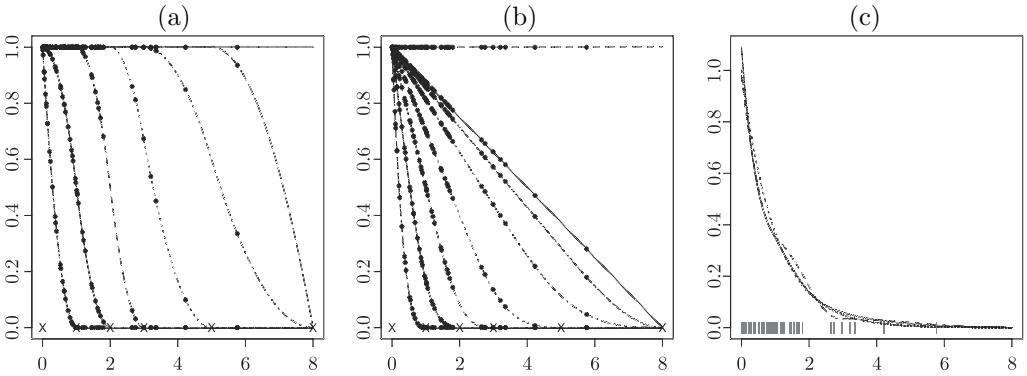

Figure 2. (a) The basis functions for the decreasing density estimator, with the values for the basis vectors indicated by the dots; the knots are marked by "X." (b) Decreasing and convex basis. (c) The least-squares decreasing (dashed) and decreasing convex (dark) density estimates for a sample of size $n = 50$ from the density shown as the dotted curve, with the data marked as vertical ticks.

with the symbol "X" on the plot. Given a random sample of points in $[0, M_0]$, we define basis vectors in which the elements of a basis vector consist of the basis functions evaluated at the values of a random sample. For a sample of size $n = 50$ from an exponential density, the basis vector values are shown in Figure 2(a) as the dots on the curves.

2.2. Decreasing and convex case

Cubic splines may be used for the estimation of a decreasing and convex function. Let $c_j = (t_{j-1} - t_{j+1})/2$ and $d_j = (t_{j+1}^2 - t_{j-1}^2 + t_j(t_{j+1} - t_{j-1}))/6$ for $j = 1, \ldots, k$. Then the $m = k + 4$ basis functions are

$$\delta_j(x) = \begin{cases} c_j x + d_j & \text{for } 0 \le x < t_{j-1}, \\ \frac{(x-t_{j-1})^3}{6(t_j - t_{j-1})} + c_j x + d_j & \text{for } t_{j-1} \le x < t_j, \\ \frac{(t_{j+1}-x)^3}{6(t_{j+1}-t_j)} & \text{for } t_j \le x < t_{j+1}, \\ 0 & \text{for } x \ge t_{j+1}, \end{cases}$$

for $j = 1, \ldots, k$, and

$$\delta_{k+1}(x) = \begin{cases} (t_1 - x)^3 & \text{for } 0 \le x < t_1, \\ 0 & \text{for } x \ge t_1, \end{cases}$$

$$\delta_{k+2}(x) = \begin{cases} c_{k+2} x + d_{k+2} & \text{for } 0 \le x < t_k, \\ \frac{(x-t_k)^3}{6(t_{k+1}-t_k)} + c_{k+2} x + d_{k+2} & \text{for } t_k \le x < t_{k+1}, \\ 0 & \text{for } x \ge t_{k+2}, \end{cases}$$

where $c_{k+2} = (t_k - t_{k+1})/2$ and $d_{k+2} = (2t_{k+1}^2 - t_k^2 - t_k t_{k+1})/6$. Finally, $\delta_{k+3}(x) = t_{k+1} - x$ and $\delta_{k+4}(x) = 1$. The functions δ_j, $j = 1, \ldots, m$, span the space of piecewise cubic spline functions. At each knot, there is only one basis function with positive second derivative and, further, there is only one basis function that is nonzero at t_{k+2}, and one basis function with nonzero slope at t_{k+2}. Hence, a linear combination is positive, decreasing, and convex if and only if the coefficients are non-negative. The piece-wise cubic decreasing and convex basis functions are shown in Figure 2(b), for the knots marked with "X." The values of the basis vectors (using the data generated for plot (c)) are marked with small circles.

2.3. Unimodal case

For estimation of a unimodal function with known mode, we place k_1 interior knots to the left of the mode, and k_2 interior knots to the right of the mode. The mode itself is a knot, and the exterior knots encompass the domain of the function, so there are $m = k_1 + k_2 + 3$ knots in all. Without loss of generality, we assume the mode is at the origin. We start with the basis functions for the monotone case, excluding the basis function that has negative slope at the origin. To the right of the mode, take $\delta_1, \ldots, \delta_{k_2+2}$ to be the monotone basis functions on those k_2 interior knots. Let $\delta_{k_2+3}, \ldots, \delta_{k_1+k_2+4}$ be the decreasing monotone basis functions defined on the k_2 interior knots to the left of the mode, where the formulas for the functions are modified to be increasing. The basis functions are shown in Figure 3 for $k_1 = k_2 = 3$, where the right-side basis functions are shown in plot (a) and the left side basis functions in plot (b). The knots are marked along the bottom with "$\wedge$" characters.

Because the maximum of each basis function is one and each has slope zero at the mode, a single inequality constraint ensures continuity of the linear combination and its derivative at the mode. Letting $m = k_1 + k_2 + 4$, a function $f(x) = \sum_{i=1}^m b_j \delta_j(x)$ is unimodal and continuous with continuous first derivative over the range of knots if and only if $b_j \geq 0$ for $j = 1, \ldots, m$ and the equality constraint $\sum_{j=1}^{k_2+2} b_j = \sum_{j=k_2+3}^{k_1+k_2+4} b_j$ holds. Let $\boldsymbol{v} = (-1, \ldots, -1, +1, \ldots, +1)'$, where the number of negative elements of $\boldsymbol{v}$ is $k_2 + 2$ and the number of positive elements is $k_1 + 2$. The set of the constraints on the vector $\boldsymbol{b}$ can be written as $\boldsymbol{b} \geq \boldsymbol{0}$ and $\boldsymbol{v}'\boldsymbol{b} = 0$. This forms a convex polyhedral cone contained in an $m - 1$ dimensional subspace of $\mathbb{R}^m$.

The coefficient vector $\boldsymbol{b}$ for the unimodal case can be written as $\boldsymbol{W}\boldsymbol{\phi}$, where the columns of the $m \times (m-1)$ matrix $\boldsymbol{W}$ span the linear subspace of $\mathbb{R}^m$ that is orthogonal to $\boldsymbol{v}$. The set

$$\left\{\boldsymbol{\phi} \in \mathbb{R}^{m-1} : \boldsymbol{W}\boldsymbol{\phi} \geq \boldsymbol{0}\right\}$$

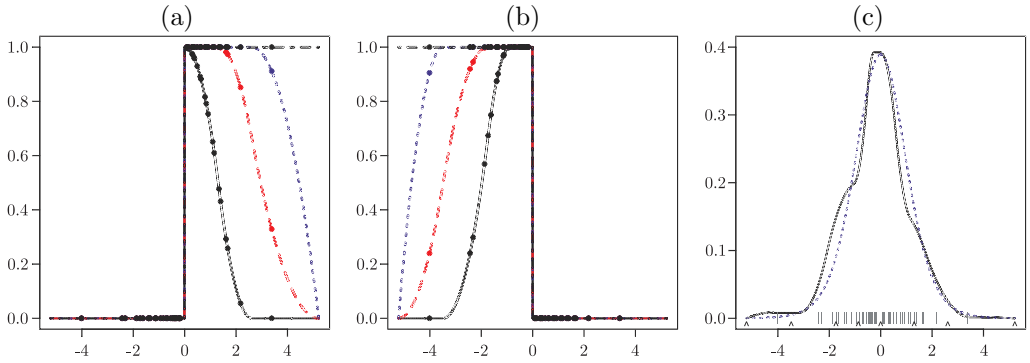

Figure 3. Plots (a) and (b) show the basis functions for unimodal density estimation. The estimated density in (c) uses the sample indicated by the tick marks, generated from the density shown as the dashed curve. The knots are marked with "∧."

is a polyhedral convex cone in $\mathbb{R}^{m-1}$, and can be written as

$$\Big\{\boldsymbol{\phi} \in \mathbb{R}^{m-1} : \boldsymbol{\phi} = \sum_{j=1}^{M} c_l \boldsymbol{\sigma}_j, \ \text{ for } \ c_j \geq 0, \ \ j = 1, \ldots, M\Big\},$$

where $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_M$ are the "edges" of the cone. These vectors can be found using Proposition 1 of Meyer (1999), where it is shown that the number M is typically larger than m, for the case of more constraints than dimensions. Here $M = (k_1 + 2)(k_2 + 2)$ and the cone of allowed coefficient vectors for the basis functions has edges $\boldsymbol{b}_l = \boldsymbol{W}\boldsymbol{\sigma}_l$, $l = 1, \ldots, M$. The linear combination $\sum_{j=1}^{m} b_j \delta_j$ is unimodal and continuous with continuous first derivative if and only if the coefficients are in the cone

$$\mathcal{C} = \Big\{\boldsymbol{b} \in \mathbb{R}^m : \boldsymbol{b} = \sum_{i=1}^{M} a_l \boldsymbol{b}_l, \ \text{ for } \ a_1, \ldots, a_M \geq 0\Big\}. \tag{2.1}$$

3. Least-squares Criterion for Spline Estimator

Groeneboom, Jongbloed, and Wellner (2001) used the following criterion to estimate a convex density based on a random sample. The functional to minimize is

$$\psi(f) = \int_0^\infty f(x)^2 dx - 2\int_0^\infty f(x) dF_n(x), \tag{3.1}$$

where F_n is the empirical cumulative distribution function. The intuition for this criterion function comes from its equivalence to the expression

$$\int_0^\infty (f(x) - f_n(x))^2 dx,$$

where f_n is the "empirical density function" whose integral is F_n. If we minimize (3.1) over decreasing densities with no smoothness assumptions, we get the Grenander estimator (Groeneboom, Jongbloed, and Wellner (2001)).

Using the appropriate basis functions defined in the last section, consider estimators in the form $\hat{f}(x) = \sum_{i=1}^{m} b_j \delta_j(x)$, noting that $\hat{f}$ will satisfy the shape assumption if and only if the coefficient vector is contained in a polyhedral convex cone. For the decreasing cases, the cone is simply $\boldsymbol{b} \geq \boldsymbol{0}$, but in the unimodal case it is given by (2.1). Let $\boldsymbol{\Delta}$ be the $n \times m$ matrix with columns containing the basis vectors defined by the knots and the observed x-values, so that $\Delta_{ij} = \delta_j(x_i)$. Then (3.1) can be written as

$$\boldsymbol{b}'\boldsymbol{H}\boldsymbol{b} - 2\boldsymbol{c}'\boldsymbol{b}, \tag{3.2}$$

where

$$H_{jl} = \int_0^\infty \delta_j(x)\delta_l(x)dx$$

and $\boldsymbol{c} = \boldsymbol{\Delta}'\boldsymbol{1}/n$. A cone projection algorithm is used to minimize (3.2) over $\mathcal{C}$, where $\mathcal{C}$ is the non-negative orthant of $\mathbb{R}^m$ for the decreasing and decreasing convex cases, and (2.1) for the unimodal case. Code for cone projection and for estimation methods in this paper, written in the `R` programming language, can be found at `www.stat.colostate.edu/~meyer/denspline.htm`.

If $\int_0^\infty \delta_j(x)dx = r_j$ for $j = 1, \ldots, m$, then the area constraint is satisfied if $\boldsymbol{r}'\boldsymbol{b} = 1$. Using a Lagrange multiplier, we have the criterion function

$$\psi_\lambda(\boldsymbol{b}) = \boldsymbol{b}'\boldsymbol{H}\boldsymbol{b} - 2(\boldsymbol{c} + \frac{\lambda \boldsymbol{r}}{2})'\boldsymbol{b}.$$

For the unconstrained solution $\tilde{\boldsymbol{b}} = \boldsymbol{H}^{-1}(\boldsymbol{c} + \lambda\boldsymbol{r}/2)$, it is easy to see that $\lambda = 0$, because $\boldsymbol{r}$ coincides with the last row of $\boldsymbol{H}$, and hence $\boldsymbol{a}'\boldsymbol{H}^{-1}\boldsymbol{c} = c_m = 1$. For the constrained solution, the value of λ may be adjusted so that $\boldsymbol{r}'\hat{\boldsymbol{b}} = 1$.

An example of a decreasing density estimate is shown as the dashed curve in Figure 2(c), using a sample of size $n = 50$ from an exponential density, shown as the dotted curve. The data are indicated by the tick marks along the bottom of the plot, and the knots are marked with "X." The solid curve is the decreasing and convex density estimate for the same data and the same knots. An example of a unimodal estimate is shown in Figure 3(c), using a sample of size $n = 50$ from the density shown as the dotted curve. The knots and basis functions used to construct the estimator are shown in plots (a) and (b).

4. Rates of Convergence

For functions on an interval $\mathcal{I}$, define the norms

$$\| g \|^2 = \int_{\mathcal{I}} g(x)^2 dx; \text{ and } \quad \| g \|_\infty = \max_{x \in \mathcal{I}} |g(x)|. \tag{4.1}$$

Let $\mathcal{G}$ be the linear space spanned by spline functions $\delta_1(x), \ldots, \delta_m(x)$ defined on $\mathcal{I}$, and let $x_1, \ldots, x_n$ be a random sample from a density f that satisfies a shape assumption and has support in $\mathcal{I}$. Let $\tilde{f}$ be the unconstrained least-squares spline density estimator, and let $\hat{f}$ be the corresponding constrained estimator. We obtain the rate of convergence of $\| \hat{f} - f \|^2$ by determining the rate of $\| \tilde{f} - f \|^2$ and showing that the former is at least as fast. We give the details for the decreasing density estimator with mode at the origin; the convex and unimodal with known mode cases follow similarly. The necessary assumptions for the decreasing case are the following.

A1. There is an $0 < M_0 < \infty$ such that $f(x) = 0$ for $x \notin [0, M_0]$.

A2. The true density f is twice continuously differentiable on $(0, M_0)$.

A3. There is an $0 < M_1 < \infty$ such that $-M_1 \leq f'(x) \leq 0$ on $(0, M_0)$.

A4. Knots are defined in $[0, M_0]$ according to a scheme that has "bounded mesh ratio," i.e., the ratio of the largest inter-knot interval to the smallest is bounded for diverging n.

Theorem 1. *Under* A1–A4*, if the number of knots is* $O(n^{1/(2p+3)})$*, then* $\| \hat{f} - f \| = O_P(n^{-(p+1)/(2p+3)})$*, where* p *is the degree of the spline.*

The proof is given in the Appendix. Some results from Huang (1998) and Huang and Stone (2002) are used. Huang (1998) derived rates of convergence for estimators that are least-squares projections onto a linear space $\mathcal{G}$, with application to the functional ANOVA model. For data $\{(X_i, Y_i)\}, i = 1, \ldots, n$ and bounded functions $\mu(x) = E(Y|X = x)$ and $\sigma^2(x) = var(Y|X = x)$, Huang defined $\hat{\mu}(x)$ to be the least-squares projection of the vector $\boldsymbol{Y}$ onto $\mathcal{G}$. The function $\bar{\mu}$ is the the projection of μ onto $\mathcal{G}$, the approximation error is $\| \bar{\mu} - \mu \|$, and the estimation error is $\| \hat{\mu} - \bar{\mu} \|$. Under some conditions on the smoothness of μ, the distribution of the X_i, and knots having bounded mesh ratio, the rate is $\| \hat{\mu} - \mu \| = O_P(n^{-(p+1)/(2p+3)})$ if the number of knots is $O(n^{1/(2p+3)})$. Huang and Stone (2002) considered extended linear modeling with polynomial splines for models in which the log-likelihood is concave. The function η is to be estimated based on a random variable with density depending on η. The maximum likelihood spline estimator is shown to have the convergence rate $O_P(n^{-(p+1)/(2p+3)})$ in the L_2 norm, with the same assumptions on the knots. Details about how these results are used to obtain the convergence rates for the least-squares spline density estimates are given in the Appendix.

4.1. Unimodal estimator with unknown mode

Suppose the true mode m is unknown, but we estimate the mode using $\hat{m}$ where $|\hat{m} - m| = O(n^{-3/7})$. Without loss of generality, suppose $m = 0$, and let

$\mathcal{G}$ be the space of splines defined for the known-mode case. For clarity we label the knots $t_{-k_1-1}, \ldots, t_{-1}, 0, t_1, \ldots, t_{k_2+1}$. As n increases, we assume that the knots have bounded mesh ratio with $k_1, k_2 = O(n^{1/7})$, so that with probability approaching one, $\hat{m} \in (t_{-1}, t_1)$. Let $\mathcal{G}_m$ be the space of splines defined on knots $t_{-k_1-1}, \ldots, t_{-1}, \hat{m}, t_1, \ldots, t_{k_2+1}$, and let $\hat{f}_m$ be the minimizer of $\psi(g; \boldsymbol{x})$ on $\mathcal{G}_m$.

Suppose δ_j^*, $j = 1, \ldots, m$ are the basis functions for $\mathcal{G}_m$, while δ_j, $j = 1, \ldots, m$ are the basis functions for $\mathcal{G}$. Define $f^*(x) = \sum_{j=1}^m \hat{b}_j \delta^*$, where $\hat{f}(x) = \sum_{j=1}^m \hat{b}_j \delta_j$; that is, f^* is to have the same coefficients as $\hat{f}$. Then $f^*(x)$ is identical to $\hat{f}(x)$ except on (t_{-2}, t_2). Using the definitions of the basis functions defined in Section 2 and the convergence rate of $\hat{m}$, it is straightforward to show that $\| \hat{f} - f^* \| = O(n^{-3/7})$. This gives $\| \bar{f} - f^* \| = O(n^{-3/7})$, so using the convexity argument at the end of the approximation error rate derivation (see Appendix) and $\psi(\hat{f}_m; \boldsymbol{x}) \leq \psi(f^*, \boldsymbol{x})$, we have $\| \hat{f}_m - f \| = O(n^{-3/7})$. Therefore, the plug-in estimator has the same convergence rate as the estimator with known mode m, if the mode estimator converges at a rate such that $\hat{m} - m = O(n^{-3/7})$.

If the unimodal density can be assumed to be symmetric, then the median as an estimator of the mode has $\hat{m} - m = O(n^{-1/2})$. Otherwise, Eddy (1980) gives mode estimates using polynomial kernel density estimation that have sufficiently fast convergence.

4.2. Case of infinite support

Suppose the support of the decreasing density f is $[0, \infty)$. The support for the constrained spline estimator $\hat{f}$ must span the data and may be taken to be $[0, a_n)$; for example, $a_n = x_{(n)}(1 + \xi)$, where $x_{(n)}$ is the largest observation and $\xi > 0$. Then the length of the support for the estimator is random and increasing to infinity. If the knots are chosen to be at equal data quantiles, the bounded mesh ratio is attained. However, for increasing support, there is no constant c such that $\| g - f \| \leq c \| g - f \|_\infty$, even if g and f are restricted to be decreasing densities. The convergence rate for the finite support case is not attained, but if the support does not increase too quickly, a slightly slower rate may be attained. For example, suppose $a_n = O_p(\log(n))$. Then it can be shown that $\rho_n = O_p(k^{-(p+1)} \log(n))$, and the approximation error is $O_p(n^{-(p+1)/(2p+3)} \log(n)^{3/2})$.

5. Simulations

5.1. Knot choices

The piecewise-constant monotone density estimator and the piecewise-linear convex or log-concave estimators do not require user-defined parameters, but smooth non-parametric density estimation typically requires a choice of bandwidth, or number and position of knots. These choices can be user-defined or

data-driven, "automatic" choices. Imposing shape restrictions disallows the "wiggling" associated with over-fitting, and in general constrained estimators are more robust to these parameters than unconstrained estimators. Hence the bandwidth can be chosen to be smaller for decreasing or unimodal kernel density estimation (Fougères (1997)), and a larger number of knots can be chosen for the spline estimator, without introducing extra modes. However, for an excessive number of knots, the decreasing density estimator approaches that of the unsmoothed monotone MLE, and the convex decreasing estimator begins to look like the piecewise linear solution. The asymptotically optimal number of knots, on the order of $n^{1/7}$, does not help much in choosing the number and placement of knots for small to moderate-sized samples.

For density estimation with known finite support, the knots should be chosen to span the support in some "even" manner. If the knots are placed at equal data quantiles, there may be large gaps if the density is large over some range and small elsewhere. If the knots are evenly spaced, there may be knot intervals containing too large a fraction of the observations. Data-driven compromises that seems to work nicely in practice are as follows. The first method is to choose an initial number of knots k_0 spaced equally in the support. Then add more knots where the data are "thickest," until there are at most $2n/(k_0 - 1)$ observations in any knot interval. If k_0 is of the optimal order, then the final number of knots is as well. Alternatively, the knots can initially be spaced in equal data quantiles, with large gaps filled in with extra knots. For estimation of a density with unknown support, the exterior knots should be chosen to span the data, or to span an *a-priori* range of "reasonable" values for the phenomenon. Then the interior knots can be filled in according to one of these schemes.

5.2. Comparison with other methods

We compare the shape-restricted regression spline estimators with other estimators in the literature, for a grid of evenly spaced points $g_1, \ldots, g_{n_g}$ spanning the knot range, using

$$SMSE = \Big[\frac{1}{Nn_g}\sum_{j=1}^{N}\sum_{i=1}^{n_g}\Big(\hat{f}_j(g_i) - f(g_i)\Big)^2\Big]^{1/2}$$

and $N = 10,000$. We first compare the decreasing and decreasing-convex regression spline estimator with the Fougères kernel rearrangement, the decreasing linear spline estimator of Meyer and Woodroofe (2004) with concave interval at the mode, and an unsmoothed, "constraints-only" decreasing convex least-squares estimator similar to the Groeneboom, Jongbloed, and Wellner (2001) estimator.

This constraints-only estimator differs slightly from theirs because the changes in slope are forced to be at sample points instead of between sample points; it can be obtained using the methods proposed in this paper, where the n basis functions are $\delta_1(x) = 1$, $\delta_i(x) = (1 - x/x_i)_+$, $i = 2, \ldots, n$, where $(\cdot)_+ = \max(\cdot, 0)$, with coefficients constrained to be non-negative. For the Fougères rearrangement we start with a standard kernel estimator (given by the `R` function `density`), truncated to the positive reals and scaled to have unit area. As recommended by that author, the bandwidth is chosen to be $0.8h$, where $h = 1.06Sn^{-1/5}$ for sample standard deviation S. For the Meyer-Woodroofe estimator, the length of the concave interval is the recommended $n^{-3/5}$. For the regression spline estimators we chose the initial numbers of knots to be $k_0 = 5$, 6, and 7 for $n = 50$, 100, and 200, respectively. The grid used for the SMSE calculation was the same for each estimator, $n_g = 1,000$ equally spaced points in the interval $[0, t_{k+1}]$, where t_{k+1} was chosen so that the support of the spline estimator was the same as that for the default kernel used to construct the Fougères estimator.

The four underlying densities are the linear decreasing density $f(x) = 2(1-x)$ on $(0, 1)$, the standard exponential, the half-normal, and a mixture of half-normals that is three-quarters standard normal and one-quarter mean-zero normal with standard deviation 10. Results in Table 1 show that for the half-normal, the Fougères estimator has the smallest SMSE values, with the difference decreasing with increasing n. For the mixture of half-normals, the decreasing spline has the smallest SMSE values, and for the exponential density, the convex decreasing spline estimator "wins." For the linear density, the Fougères rearrangement does better for the smallest sample size, and the convex decreasing does best for the largest size. The constraints-only convex estimator tends to have large SMSE because of the spiking problem at the mode, where the estimator is too large. The number in parentheses is the SMSE for the portion of the interval starting at $x_{(4)}$, the fourth-smallest observed value.

The unimodal regression spline density estimator is compared with the standard kernel obtained through the `R` function `density` with default bandwidth h, the Fougères unimodal kernel rearrangement with bandwidth $0.8h$, and the log-concave density estimate as coded in `logConDens`, with default parameters. For the regression spline estimators we chose the initial numbers of knots to be $k_0 = 6$, 6, and 8 for $n = 50$, 100, and 200, respectively. The support for the estimate is chosen to be the range of x-values provided by the `R` function `density`, and the supports for the four estimators are identical. The three underlying densities are standard normal, a normal mixture with 75% $N(0, 1)$ and 25% $N(0, 10)$, and a beta density with parameters $\alpha = 6$ and $\beta = 2$. Results in Table 2 show that the regression spline estimator (UMRS) has smallest SMSE for the normal

Table 1. Comparison of SMSE in decreasing density estimation methods. The decreasing regression spline (DRS) and the convex decreasing (CDRS) are compared with the decreasing kernel of Fougères (FDK), the linear spline estimate with concave interval at the mode (CONC), and the least-squares decreasing convex "constraints only" density estimator (CLIN).

n	f	DRS	CDRS	FDK	CONC	CLIN
50	linear	0.220	0.202	0.173	0.271	0.600 (0.268)
100	linear	0.175	0.150	0.149	0.213	0.540 (0.208)
200	linear	0.139	0.108	0.130	0.171	0.322 (0.164)
50	exp	0.069	0.064	0.078	0.099	0.546 (0.120)
100	exp	0.051	0.046	0.066	0.069	0.224 (0.098)
200	exp	0.038	0.033	0.055	0.051	0.220 (0.078)
50	halfnorm	0.084	—	0.072	0.131	—
100	halfnorm	0.063	—	0.058	0.089	—
200	halfnorm	0.047	—	0.048	0.064	—
50	normmix	0.027	—	0.038	0.045	—
100	normmix	0.020	—	0.029	0.030	—
200	normmix	0.015	—	0.022	0.022	—

mixture, and does not do as well for the standard normal or the beta density for the smaller sample sizes. The log-concave density is best when the true density is log-concave, but for the normal mixture (which is not log-concave), it is substantially worse than its competitors. Examples of the estimates are shown in Figure 4 for a sample of size $n = 100$ from the Beta(6,2) density, where the tick marks along the bottom indicate the sample and the dotted curves show the true density. A clump of observations to the left of the mode is reflected in the standard kernel (solid curve, plot (a)); the Fougères rearrangement (dot-dash) is not smooth. The regression spline estimate (solid curve) and the log-concave estimate (dashed) are shown in plot (b).

6. Discussion

The constrained spline density estimator is a straightforward method for obtaining smooth density estimates that satisfy *a-priori* assumptions about shape. Convergence rates in the L_2 norm of $O_P(n^{-3/7})$ for the decreasing and unimodal cases and $O_P(n^{-4/9})$ for the decreasing and convex case are attained. These are superior to rates attained by competing estimators. The unsmoothed decreasing estimator attains the rate $n^{-1/3}$ and the unsmoothed convex decreasing and log-concave estimators attain the rate $n^{-2/5}$. The various modifications of the standard kernel estimator attain the rate $n^{-2/5}$ of the unmodified kernel estimator.

The estimate is obtained through a single projection onto a polyhedral convex cone of relatively small dimension. The 10,000 decreasing density estimates

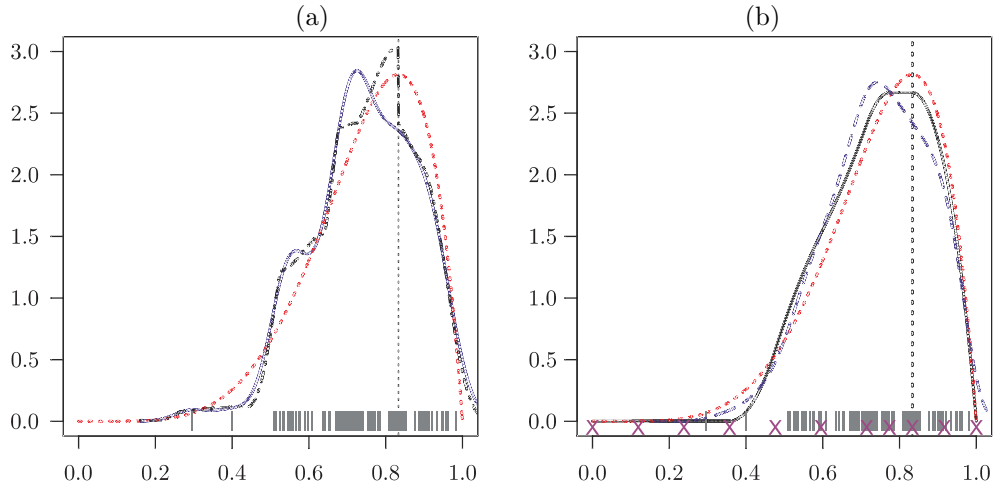

Figure 4. Density estimates for $n = 100$ independent observations (shown as ticks) from a Beta(6,2) density (dotted curves). (a) Solid is unconstrained kernel, dot-dash is Fougères rearrangement. (b) Dashed is log-concave estimate, solid is unimodal spline estimate with knots marked as "X."

Table 2. Comparison of SMSE in unimodal density estimation with known mode. The unimodal spline (UMRS) is compared with the standard kernel (KERN), the unimodal kernel of Fougères (FUM), and the smoothed log-concave density estimator.

n	f	UMRS	KERN	FUM	log-conc
50	N(0,1)	0.044	0.041	0.041	0.038
100	N(0,1)	0.032	0.031	0.031	0.028
200	N(0,1)	0.024	0.023	0.024	0.022
50	normmix	0.017	0.017	0.017	0.049
100	normmix	0.011	0.012	0.012	0.046
200	normmix	0.0078	0.0086	0.0088	0.044
50	Beta(6,2)	0.293	0.285	0.275	0.250
100	Beta(6,2)	0.225	0.221	0.216	0.187
200	Beta(6,2)	0.168	0.171	0.168	0.156

with $n = 200$ for the simulations were accomplished in just over four minutes on a Mac Powerbook with 3.06 GHz Intel processor and 4 GB of RAM. The computational speed makes it suitable for inclusion in iterative computations, such as in problems involving estimation of an error density in regression.

Acknowledgement

This work was funded by NSF DMS-0905656. The author wishes to acknowledge useful discussions with Jayanta Pal, who suggested the least-squares

estimation, as well as very helpful comments from two referees and an associate editor.

Appendix: Convergence Rate Proofs

The criterion function for a candidate density h defined on $[0, M_0]$ is

$$\psi(h; \boldsymbol{x}) = \int_0^{M_0} h(x)^2 dx - \frac{2}{n} \sum_{i=1}^{n} h(x_i). \tag{A.1}$$

It is easy to see that ψ is strictly convex in h in the sense that, for any $\boldsymbol{x}$,

$$\psi[\alpha h_1 + (1-\alpha)h_2; \boldsymbol{x}] < \alpha\psi(h_1; \boldsymbol{x}) + (1-\alpha)\psi(h_2; \boldsymbol{x})$$

for $\alpha \in (0,1)$ and distinct h_1 and h_2. Here, we consider h_1 and h_2 to be distinct if $\int_0^{M_0} [h_1(x) - h_2(x)]^2 dx > 0$.

For the space $\mathcal{G}$ of spline functions on $[0, M_0]$, let $\rho_n = \inf_{g \in \mathcal{G}} \parallel g - f \parallel_\infty$. Huang (1998) showed that by properties of polynomial splines and under assumptions (A2) and (A4), $\rho_n = O(k_n^{-(p+1)})$, where p is the degree of the spline, k_n is the number of knots. Let $\tilde{f}$ minimize $\psi(\cdot; \boldsymbol{x})$ over $\mathcal{G}$. We derive the convergence rate for $\tilde{f}$ to f by considering approximation error and estimation error separately. Consider $q_1 < \cdots < q_n$, the quantiles of the density f so that the *cdf* $F(q_i) = i/n$, and take $q_0 = 0$. Let $\bar{f}$ minimize $\psi(\cdot; \boldsymbol{q})$ over $\mathcal{G}$, then take the approximation error to be $\parallel \bar{f} - f \parallel$, and the estimation error to be $\parallel \tilde{f} - \bar{f} \parallel$. Clearly, the error $\parallel \tilde{f} - f \parallel$ is not larger than the sum of the estimation and approximation errors.

Let f_s minimize $\psi(\cdot; \boldsymbol{q})$ over the set of decreasing densities. Then for $x \in (q_{i-1}, q_i]$,

$$f_s(x) = \frac{1}{n(q_i - q_{i-1})}. \tag{A.2}$$

To see this, consider any decreasing density h, and

$$\begin{aligned}
&\frac{d}{d\alpha}\psi[\alpha h + (1-\alpha)f_s; \boldsymbol{q}] \\
&\quad = \frac{d}{d\alpha}\left\{ \int_0^{M_0} [\alpha h(x) + (1-\alpha)f_s(x)]^2 \, dx - \frac{2}{n}\sum_{i=1}^{n} [\alpha h(q_i) + (1-\alpha)f_s(q_i)] \right\} \\
&\quad = 2\int_0^{M_0} [\alpha h(x) + (1-\alpha)f_s(x)] \, [h(x) - f_s(x)] \, dx - \frac{2}{n}\sum_{i=1}^{n} [h(q_i) - f_s(q_i)] \, .
\end{aligned}$$

Setting $\alpha = 0$ and plugging in (A.2) for f_s, we have

$$\begin{aligned}
&\frac{1}{2}\frac{d}{d\alpha}\psi[\alpha h + (1-\alpha)f_s; \boldsymbol{q}]|_{\alpha=0} \\
&\quad = \sum_{i=1}^{n} \int_{q_{i-1}}^{q_i} \frac{1}{n(q_i - q_{i-1})} \left[h(x) - \frac{1}{n(q_i - q_{i-1})} \right] dx - \frac{1}{n} \sum_{i=1}^{n} \left[h(q_i) - \frac{1}{n(q_i - q_{i-1})} \right] \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} [h(c_i) - h(q_i)] \qquad \text{(A.3)}
\end{aligned}$$

for some c_i where $q_{i-1} \leq c_i \leq q_i$, $i = 1, \ldots, n$, by the Mean Value Theorem. If h is decreasing, this is positive for $h \neq f_s$, showing that the criterion function increases from f_s along any direction toward another decreasing density. To see that $f_s(x) - f(x) = O(1/n)$ for $x \in [0, M_0]$, note that by the continuity properties of f (A2), $f(q_{i-1}) - f(q_i) = O(1/n)$ and for $x \in (q_{i-1}, q_i]$, $f(x), f_s(x) \in [f(q_i), f(q_{i-1})]$ (A3). Then $\| f_s - f \| = O(1/n)$. Now we show that $\bar{f}$ must be close to f_s.

The second derivative of $\psi[\alpha h_1 + (1-\alpha)h_2; \boldsymbol{x}]$ with respect to α does not depend on the values of $\boldsymbol{x}$:

$$\frac{d^2}{d\alpha^2}\psi[\alpha h_1 + (1-\alpha)h_2; \boldsymbol{x}] = 2\int_0^{M_0} [h_1(x) - h_2(x)]^2 \, dx = 2 \| h_1 - h_2 \|^2 . \quad \text{(A.4)}$$

Integration by parts gives an expression for the difference in the criterion function for f_s and an arbitrary function g, in terms of the distance between the f_s and g:

$$\begin{aligned}
\psi(g; \boldsymbol{q}) - \psi(f_s; \boldsymbol{q}) &= \int_0^1 \frac{d}{d\alpha}\psi[\alpha g + (1-\alpha)f_s; \boldsymbol{q}] d\alpha \\
&= \frac{d}{d\alpha}\psi[\alpha g + (1-\alpha)f_s; \boldsymbol{q}]|_{\alpha=0} \\
&\quad + \int_0^1 (1-\alpha)\frac{d^2}{d\alpha^2}\psi[\alpha g + (1-\alpha)f_s; \boldsymbol{q}] d\alpha \\
&= \frac{2}{n}\sum_{i=1}^{n} [g(c_i) - g(q_i)] + \| g - f_s \|^2
\end{aligned}$$

for some $c_i \in [q_{i-1}, q_i]$, $i = 1, \ldots, n$, by (A.3) and (A.4).

The expression $\sum_{i=1}^{n} [g(c_i) - g(q_i)]$ is less than the total variation of g, so that

$$\| g - f_s \|^2 - \frac{2M_3}{n} \leq \psi(g; \boldsymbol{q}) - \psi(f_s; \boldsymbol{q}) \leq \| g - f_s \|^2 + \frac{2M_3}{n}$$

for g with total variation at most M_3. Also, $\| g - f_s \| \leq \| g - f \| + \| f - f_s \|$ so that, for large enough n, there is an M_2 such that $\| g - f_s \|^2 \leq \| g - f \|^2 + M_2/n$.

By compactness of $\mathcal{G}$, there is a $g^* \in \mathcal{G}$ such that $\| g^* - f \|_\infty = \rho_n$; then

$$\psi(g^*; \boldsymbol{q}) - \psi(f_s; \boldsymbol{q}) \leq M_0 \rho_n^2 + \frac{M_2}{n} + \frac{2M_3}{n}$$

because A1 implies that $\| \cdot \| \leq M_0^{1/2} \| \cdot \|_\infty$. Fix $\eta > 0$ and consider $g \in \mathcal{G}$ such that $\| g - f_s \| = (M_0^{1/2} + \eta)\rho_n$. Then

$$\psi(g; \boldsymbol{q}) - \psi(f_s; \boldsymbol{q}) \geq (M_0^{1/2} + \eta)^2 \rho_n^2 - \frac{2M_3}{n}$$

so for large enough n, $\psi(g; \boldsymbol{q}) > \psi(g^*; \boldsymbol{q})$ when $\| g - f \| = (M_0^{1/2} + \eta)\rho_n$. If $\bar{f}$ minimizes $\psi(\cdot; \boldsymbol{q})$ over $\mathcal{G}$, then $\psi(\bar{f}; \boldsymbol{q}) < \psi(g^*; \boldsymbol{q})$ and by convexity of ψ we must have $\| \bar{f} - f_s \| \leq (M_0^{1/2} + \eta)\rho_n$. Using $\| \bar{f} - f \| \leq \| \bar{f} - f_s \| + \| f - f_s \|$ we have that $\| \bar{f} - f \| = O(\rho_n)$. This is the approximation error.

For the estimation error, first we need to show

$$\sup_{g \in \mathcal{G}} \frac{\left| \frac{d}{d\alpha} \psi(\bar{f} + \alpha g; \boldsymbol{x})|_{\alpha=0} \right|}{\| g \|} = O_p\left(\left(\frac{m}{n} \right)^{1/2} \right). \tag{A.5}$$

For $g \in \mathcal{G}$,

$$\begin{aligned}
\frac{d}{d\alpha} \psi(\bar{f} + \alpha g; \boldsymbol{x})|_{\alpha=0} &= \frac{d}{d\alpha} \left\{ \int_0^{M_0} \left[\bar{f}(x) + \alpha g(x) \right]^2 dx - \frac{2}{n} \sum_{i=1}^n \left[\bar{f}(x_i) + \alpha g(x_i) \right] \right\}_{\alpha=0} \\
&= 2 \left\{ \int_0^{M_0} \bar{f}(x) g(x) dx - \frac{1}{n} \sum_{i=1}^n g(x_i) \right\} \\
&= 2 \sum_{i=1}^m b_j \left\{ \int_0^{M_0} \bar{f}(x) \delta_j(x) dx - \frac{1}{n} \sum_{i=1}^n \delta_j(x_i) \right\} \\
&= 2 \sum_{i=1}^m b_j a_j,
\end{aligned}$$

where the last equality defines a_j. Then

$$\sup_{g \in \mathcal{G}} \frac{\left| \frac{d}{d\alpha} \psi(\bar{f} + \alpha g)|_{\alpha=0} \right|}{\| g \|} = \sup_{b \in \mathbb{R}^m} \frac{2 \boldsymbol{a}' \boldsymbol{b}}{[\boldsymbol{b}' \boldsymbol{H} \boldsymbol{b}]^{1/2}} = 2 \left[\boldsymbol{a}' \boldsymbol{H}^{-1} \boldsymbol{a} \right]^{1/2}.$$

If we show that $a_j = O_P(n^{-1/2})$ for each $j = 1, \ldots, m$, then (A.5) follows. Starting with

$$a_j = \int_0^{M_0} \left[\bar{f}(x) - f(x) \right] \delta_j(x) dx + \int_0^{M_0} f(x) \delta_j(x) dx - \frac{1}{n} \sum_{i=1}^n \delta_j(x_i),$$

we have that the second two terms are $O_p(n^{-1/2})$ by the Central Limit Theorem. To show that the first term is $O(n^{-1/2})$, let $\psi_1(g) = \psi(g;\boldsymbol{q})$ and let $\psi_2(g) =\| g \|^2 -2\int_0^{M_0} f(x)g(x)dx$. Let $\bar{g}$ minimize ψ_2 (and hence $\| g-f \|^2$) over $\mathcal{G}$, and recall that $\bar{f}$ minimizes ψ_1 over $\mathcal{G}$. Integration by parts gives

$$\begin{aligned}\psi(g;\boldsymbol{x}) - \psi(\bar{f};\boldsymbol{x}) &= \frac{d}{d\alpha}\psi\left(\bar{f}+\alpha(g-\bar{f});\boldsymbol{x}\right)|_{\alpha=0} \\ &\quad + \int_0^1 (1-\alpha)\frac{d^2}{d\alpha^2}\psi\left(\bar{f}+\alpha(g-\bar{f});\boldsymbol{x}\right)d\alpha \\ &= \frac{d}{d\alpha}\psi\left(\bar{f}+\alpha(g-\bar{f});\boldsymbol{x}\right)|_{\alpha=0} + \frac{1}{2}\| \bar{f}-g \|^2, \qquad \text{(A.6)}\end{aligned}$$

so that (plugging in $\boldsymbol{q}$ for $\boldsymbol{x}$ and noting $\bar{f}$ minimizes $\psi(\cdot;\boldsymbol{q})$ in $\mathcal{G}$),

$$\psi_1(\bar{g}) - \psi_1(\bar{f}) = \frac{1}{2}\| \bar{f}-\bar{g} \|^2,$$

and $\psi_2(\bar{g}) - \psi_2(\bar{f})$ is negative. For any $g \in \mathcal{G}$,

$$\frac{1}{2}\left(\psi_2(g) - \psi_1(g)\right) = \int_0^{M_0} f(x)g(x)dx - \frac{1}{n}\sum_{i=1}^n g(q_i),$$

and it is straight-forward to show that $\psi_2(g) - \psi_1(g) = O(1/n)$ for g with total variation less than M_3. Writing

$$[\psi_1(\bar{g}) - \psi_2(\bar{g})] - \left[\psi_1(\bar{f}) - \psi_2(\bar{f})\right] = \frac{1}{2}\| \bar{f}-\bar{g} \|^2 + \left[\psi_2(\bar{f}) - \psi_2(\bar{g})\right],$$

note that the left hand side is $O(1/n)$ and the second term on the right hand side is positive. Therefore, $\| \bar{f}-\bar{g} \|^2 = O(1/n)$. The first term in a_j is

$$\begin{aligned}&\int_0^{M_0} \left[\bar{f}(x) - f(x)\right]\delta_j(x)dx \\ &\quad = \int_0^{M_0} \left[\bar{f}(x) - \bar{g}(x)\right]\delta_j(x)dx + \int_0^{M_0} \left[\bar{g}(x) - f(x)\right]\delta_j(x)dx.\end{aligned}$$

The second term is zero by definition of $\bar{g}$ and the first term is $O(n^{-1/2})$ by the boundedness of δ_j. Then (A.5) follows.

By (A.5), we can choose a constant K large enough so that with probability approaching one,

$$\left|\frac{d}{d\alpha}\psi\left(\bar{f}+\alpha(g-\bar{f});\boldsymbol{x}\right)|_{\alpha=0}\right| \leq K^{1/2}\left(\frac{m}{n}\right)^{1/2}\| \bar{f}-g \|$$

and, by (A.6),

$$\frac{1}{2} \parallel \bar{f} - g \parallel^2 - K^{1/2} \left(\frac{m}{n}\right)^{1/2} \parallel \bar{f} - g \parallel \leq \psi(g; \boldsymbol{x}) - \psi(\bar{f}; \boldsymbol{x}) \leq \frac{1}{2} \parallel \bar{f} - g \parallel^2 + K^{1/2} \left(\frac{m}{n}\right)^{1/2} \parallel \bar{f} - g \parallel .$$

Consider $g \in \mathcal{G}$ where $\parallel \bar{f} - g \parallel^2 = 4Km/n$. For such a g, we have $\psi(g; \boldsymbol{x}) > \psi(\bar{f}; \boldsymbol{x})$ so, by convexity of ψ, we must have $\parallel \bar{f} - g \parallel^2 \leq 4Km/n$ whenever $\psi(g; \boldsymbol{x}) < \psi(\bar{f}; \boldsymbol{x})$. Because $\tilde{f}$ minimizes $\psi(g; \boldsymbol{x})$ over $\mathcal{G}$, we have that the estimation error is $\parallel \bar{f} - \tilde{f} \parallel^2 = O_p(m/n)$. To minimize the error $\parallel \tilde{f} - f \parallel$, we set the estimation and approximation errors equal and recall that m is k plus a small integer, to get the optimal $k = O(n^{1/(2p+3)})$ and $\parallel \tilde{f} - f \parallel = O_p(n^{-(p+1)/(2p+3)})$.

Finally, we consider $\hat{f}(x) = \sum_{i=1}^m \hat{b}_j \delta_j(x)$, which minimizes $\psi(\cdot; \boldsymbol{x})$ over $\boldsymbol{b} \in \mathcal{C}$, where $\mathcal{C} = \{\boldsymbol{b} \in I\!R^m : \boldsymbol{b} \geq \boldsymbol{0}\}$ for the decreasing cases and $\mathcal{C}$ is defined as in (2.1). Define $\bar{\boldsymbol{b}}$ by $\bar{f} = \sum_{i=1}^m \bar{b}_j \delta_j(x)$ where $\bar{f}$ minimizes $\psi(\cdot, \boldsymbol{q})$ over $\mathcal{G}$. We assume that f satisfies the shape assumption and $\bar{\boldsymbol{b}} \in \mathcal{C}$. The necessary and sufficient conditions for $\hat{\boldsymbol{b}}$ to minimize $\psi(\boldsymbol{b}) = \boldsymbol{b}'\boldsymbol{H}\boldsymbol{b} - 2\boldsymbol{c}'\boldsymbol{b}$ over $\boldsymbol{b} \in \mathcal{C}$ are

$$(\boldsymbol{H}\hat{\boldsymbol{b}} - \boldsymbol{c})'\hat{\boldsymbol{b}} = 0 \ \text{ and } \ (\boldsymbol{H}\hat{\boldsymbol{b}} - \boldsymbol{c})'\boldsymbol{b} \geq 0, \ \text{ for all } \ \boldsymbol{b} \in \mathcal{C};$$

the conditions for the unconstrained $\tilde{\boldsymbol{b}}$ are simply $(\boldsymbol{H}\tilde{\boldsymbol{b}} - \boldsymbol{c})'\boldsymbol{b} = 0$ for all $\boldsymbol{b} \in I\!R^m$. Then we have

$$\begin{aligned}(\tilde{\boldsymbol{b}} - \bar{\boldsymbol{b}})'\boldsymbol{H}(\tilde{\boldsymbol{b}} - \bar{\boldsymbol{b}}) &= (\tilde{\boldsymbol{b}} - \hat{\boldsymbol{b}})'\boldsymbol{H}(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{b}}) + (\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}})'\boldsymbol{H}(\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}) + 2(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{b}})'\boldsymbol{H}(\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}),\end{aligned}$$

so that

$$\begin{aligned}(\tilde{\boldsymbol{b}} - \bar{\boldsymbol{b}})'\boldsymbol{H}(\tilde{\boldsymbol{b}} - \bar{\boldsymbol{b}}) - (\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}})'\boldsymbol{H}(\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}) &= (\tilde{\boldsymbol{b}} - \hat{\boldsymbol{b}})'\boldsymbol{H}(\tilde{\boldsymbol{b}} - \hat{\boldsymbol{b}}) + 2(\boldsymbol{H}\tilde{\boldsymbol{b}} - \boldsymbol{c})'(\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}) - 2(\boldsymbol{H}\hat{\boldsymbol{b}} - \boldsymbol{c})'(\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}}).\end{aligned}$$

According to the conditions on $\tilde{\boldsymbol{b}}$ and $\hat{\boldsymbol{b}}$, the second term on the right is zero and the third term is positive. The first term is positive because H is positive-definite, so

$$(\tilde{\boldsymbol{b}} - \bar{\boldsymbol{b}})'\boldsymbol{H}(\tilde{\boldsymbol{b}} - \bar{\boldsymbol{b}}) \geq (\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}})'\boldsymbol{H}(\hat{\boldsymbol{b}} - \bar{\boldsymbol{b}})$$

or $\parallel \hat{f} - \bar{f} \parallel^2 \leq \parallel \tilde{f} - \bar{f} \parallel^2$. Hence the estimation error is smaller for the constrained estimator, and the approximation errors are the same. Thus, the convergence rate for the constrained spline density estimator is at most that for the unconstrained spline density estimator.

References

Balabdaoui, F. (2007). Consistent estimation of a convex density at the origin. *Math. Methods Statist.* **16**, 77-95.

Balabdaoui, F. and Wellner, J. (2010). Estimation of a k-monotone density: characterizations, consistency and minimax lower bounds. *Statist. Neerlandica* **65**, 45-70.

Bickel, P. and Fan, J. (1996). Some problems of the estimation of unimodal densities. *Statist. Sinica* **6**, 23-45.

Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function. *Bernoulli* **15**, 40-68.

Dümbgen, L. and Rufibach, K. (2011). **logcondens:** Computations related to univariate log-concave density estimation. *J. Statist. Soft.* **39**, 1-28.

Eddy, W. F. (1980). Optimum kernel estimators of the mode. *Ann. Statist.* **8**, 870-882.

Eggermont, P. P. B. and LaRiccia, V. N. (2000). Maximum likelihood estimation of smooth monotone and unimodal densities. *Ann. Statist.* **28**, 922-947.

Fougères, A. L. (1997). Estimation of unimodal densities. *Canad. J. Statist.* **25**, 375-387.

Grenander, P. (1956). On the theory of mortality measurement, part II. *Skand. Akt.* **39**, 125-153.

Groeneboom, P., Jongbloed, G. and Wellner(2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29**, 1653-1698.

Hall, P. and Huang, L. S. (2002). Unimodal density estimation using kernel methods. *Statist. Sinica* **12**, 965-990.

Hall, P. and Kang, K. H. (2005). Unimodal kernel density estimation by data sharpening. *Statist. Sinica* **15** , 73-98.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *Ann. Statist.* **26**, 242-272.

Huang, J. Z. and Stone, C. J. (2002). Extended linear modeling with splines. In *Nonlinear Estimation and Classification*. Dension, D. D., Hansen, M. H., Holmes, C. C., Malick, B. and Yu, B. (Eds.), 213-234. Springer, New York.

Mammen, E., Marron, J. S., Turlach, B. A. and Wand, M. P. (2001). A general projection framework for constrained smoothing. *Statist. Sci.* **16**, 232-248.

Meyer, M. C. (1999). An extension of the mixed primal-dual bases algorithm to the case of more constraints than dimensions. *J. Statist. Plann. Inference* **81**, 13-31.

Meyer, M. C. (2001). An alternative unimodal density estimator with a consistent estimate of the mode. *Statist. Sinica* **11**, 1159-1174.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *Ann. Appl. Statist.* **2**, 1013-1033.

Meyer, M. C. and Woodroofe, M. (2004). Estimation of a unimodal density using shape restrictions. *Canad. J. Statist.* **32**, 85-100.

Pal, J. K., Woodroofe, M. and Meyer, M. C. (2007). Estimating a Polya frequency function. In *Complex Datasets and Inverse Problems: Tomography, Networks, and Beyond*. IMS Lecture Notes- Monograph Series. IMS.

Prakasa Rao, B. L. S. (1969). Estimation of unimodal densities. *Sankhyā Ser. A* **31**, 23-36.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**, 425-461.

Rufibach, K. (2007). Computing maximum likelihood estimators of a long-concave density function. *J. Statist. Comput. Simulation* **77**, 561-574.

Woodroofe, M. and Sun, J. (1993). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. *Statist. Sinica* **3**, 501-515.

Department of Statistics, Colorado State University, 212 Statistics Building, Fort Collins, CO 80523-1877, USA.

E-mail: meyer@stat.colostate.edu

(Received December 2010; accepted June 2011)