SHAPE CONSTRAINED DENSITY ESTIMATION VIA PENALIZED RÉNYI DIVERGENCE

ROGER KOENKER AND IVAN MIZERA

Abstract. Shape constraints play an increasingly prominent role in nonparametric function estimation. While considerable recent attention has been focused on log concavity as a regularizing device in nonparametric density estimation, weaker forms of concavity constraints encompassing larger classes of densities have received less attention but offer some additional flexibility. Heavier tail behavior and sharper modal peaks are better adapted to such weaker concavity constraints. When paired with appropriate maximal entropy estimation criteria these weaker constraints yield tractable, convex optimization problems that broaden the scope of shape constrained density estimation in a variety of applied subject areas.

1. Introduction

The observation of Grenander (1956) that maximum likelihood, while failing for the general problem of probability density estimation, still delivers a viable result under monotonicity restriction may be considered the genesis of shape constrained nonparametric density estimation. Prakasa Rao (1969) first investigated nonparametric maximum likelihood estimation of a unimodal density assuming a known mode and developing large sample theory for the Grenander (1956) estimator. An extensive literature has followed, including work by Birgé (1997) incorporating estimation of the mode, and work on exploratory diagnostics for unimodality by Cox (1966), Silverman (1981), Hartigan and Hartigan (1985), and others.

As noted by Dümbgen and Rufibach (2009), estimating unimodal densities à la Grenander is not fully satisfactory; even when the mode is known some additional restrictions on the estimated density are needed to achieve global consistency. This may help to explain the recent shift in research focus to surrogates of unimodality. Log-concave densities, or *strongly* unimodal densities constitute a natural alternative since they play an important role in core statistical theory as well as many application areas, and offer some distinct advantages over unimodality *per*

Version: September 7, 2017.

se from both computational and theoretical perspectives as elucidated by early exponents of the approach: Eggermont and LaRiccia (2001), Walther (2002), Dümbgen and Rufibach (2009), Pal, Woodroofe, and Meyer (2007), and Cule, Samworth, and Stewart (2010). See Walther (2009) for a more extensive review, and Eggermont and LaRiccia (2000, 2001) for related discussion from a slightly different perspective.

Shape constraints can be formalized as imposing a "hard" penalty that takes the value 0 if the constraint is satisfied and $+\infty$ otherwise. Such penalties, in constrast to the "soft" norm-type penalties considered in Koenker and Mizera (2007, 2008), have the salient virtue that they require no choice of tuning parameters. Shape constraints are consequently somewhat simpler mathematically, so we will consider them first, returning to norm-type penalties toward the end of our exposition.

Given $X = \{X_1, \cdots, X_n\}$, Koenker and Mizera (2010) started from the variational formulation of the log-concave MLE problem, as the convex program,

$$(P_1) \qquad \min\Big\{ \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int e^{-g(x)} dx \;\Big|\; g \in \mathcal{K}(X) \Big\},$$

where $\mathcal{K}(X)$ denotes the set of closed convex functions on the convex hull, $\mathcal{H}(X)$, of X. A solution $\hat{g} : \mathcal{H}(X) \mapsto \mathbb{R}$ yields a density estimate $\hat{f}(x) = \exp(-\hat{g}(x))$ on $\mathcal{H}(X)$; the fact that this obviously positive quantity is a probability density estimate, that is, its integral is equal to one, is assured by the presence of the integral term in (P_1) Outside $\mathcal{H}(X)$, the solution $\hat{g}(x) = -\infty$, which means that $\hat{f}(x) = 0$.

Interpreting (P_1) as a "primal" formulation in the context of convex programming, Koenker and Mizera (2008, 2010), derived the associated "dual" problem,

$$(D_1) \qquad \max\Big\{ \int -f \log f dx \Big|\; f = \frac{d(\mathbb{Q}(X) - G)}{dx},\; G \in \mathcal{K}(X)^o \Big\},$$

where $\mathbb{Q}(X) = n^{-1}\sum_{i=1}^{n} \delta_{X_i}$ is the empirical probability measure,

$$\mathcal{K}(X)^o = \Big\{ G \in \mathcal{C}^*(X) \;\Big|\; \int g\, dG \le 0,\; g \in \mathcal{K}(X) \Big\}$$

is the polar cone associated with $\mathcal{K}(X)$, and $\mathcal{C}^*(X)$ denotes the set of (signed) Radon measures on $\mathcal{H}(X)$. The appearance of the Shannon entropy in the dual formulation (D_1) can be interpreted as expressing the desire to find $\hat{f}$ closest in the Kullback-Leibler divergence to the uniform distribution on $\mathcal{H}(X)$ subject to the concavity constraint.

For the problem (P_1), the solutions admit further characterization: $\hat{g}$ are piecewise linear on $\mathcal{H}(X)$, so estimated densities are piecewise exponential; see for example Koenker and Mizera (2010), Theorem 2.1. This feature motivated us to look for larger classes of quasi-concave densities that would accommodate heavier tails and more sharply peaked densities than the log concaves. Such a structure can be built upon the foundation of the classical means of order s, as for example in Hardy, Littlewood, and Pólya (1934). For any p in the unit simplex $\mathcal{S}_n \equiv \{p \in \mathbb{R}^d | p \geq 0, \sum p_i = 1\}$, let

$$M_s(a; p) = M_s(a_1, \cdots, a_n; p) = \Big(\sum_{i=1}^{n} p_i a_i^s \Big)^{1/s},$$

for $s \neq 0$; the limiting case for $s = 0$ is

$$M_0(a; p) = M_0(a_1, \cdots, a_n; p) = \prod_{i=1}^{n} a_i^{p_i}.$$

The familiar arithmetic, geometric, and harmonic means correspond to s equal to 1, 0, and -1, respectively. Following Avriel (1972), a non-negative, real function f, defined on a convex set $C \subset \mathbb{R}^d$ is called *s-concave* if for any $x_0, x_1 \in C$, and $p \in \mathcal{S}_2$,

$$f(p_0 x_0 + p_1 x_1) \geq M_s(f(x_0), f(x_1); p).$$

In this terminology log-concave functions are 0-concave, and concave functions are 1-concave. As $M_s(a, p)$ is monotone increasing in s for $a \geq 0$ and any $p \in \mathcal{S}$, it follows that if f is s-concave, then f is also s'-concave for any $s' < s$. Thus, concave functions are log-concave, but not vice-versa. In the limit $-\infty$-concave functions satisfy the condition

$$f(p_0 x_0 + p_1 x_1) \geq \min\{f(x_0), f(x_1)\},$$

so they are *quasi-concave*, as are all other s-concave functions. Note that in the one-dimensional case, for $d = 1$, the class of quasi-concave functions is identical with that of unimodal functions.

Once one decides to impose log-concavity, maximizing log likelihood appears to be especially convenient as seen in (P_1) where the only source of nonlinearity of the convex program arises from the integrability constraint. However, when weaker forms of concavity are considered it proves more convenient to adapt the fitting criterion. This is already apparent in earlier work of Groeneboom, Jongbloed, and Wellner (2001) who employ least squares fitting rather than log-likelihood when imposing the stronger requirement of concavity of the density itself. While it is not really obvious how to adapt (P_1) to obtain a

viable fitting formulation, the appearance of the Kullback-Leibler divergence in (D_1) suggested the possibility of replacing it by one of the abundant assortment of alternative divergences. For s-concave densities, this turned out to produce a lucky match. In Koenker and Mizera (2008, 2010), we proposed replacing the Shannon entropy in (D_1) by a variationally equivalent form of the Rényi entropy, for $\alpha < 1$, $\alpha \neq 0$, equal to

$$\mathcal{E}^\alpha(f) = (1-\alpha)^{-1} \log\Big(\int f^\alpha(x) dx\Big).$$

This move yields a family of new dual and primal pairings,

$$(D_\alpha) \qquad \max\Big\{\frac{1}{\alpha}\int f^\alpha(y)\, dy \;\Big|\; f = d(P_n - G)/dy, \quad G \in \mathcal{K}(X)^o\Big\},$$

and

$$(P_\alpha) \qquad \min\Big\{\sum_{i=1}^n g(X_i) + \frac{|1-\alpha|}{\alpha}\int g^\beta\, dx \;\Big|\; g \in \mathcal{K}(X)\Big\}.$$

Here the Rényi exponent α corresponds to Avriel's $s = \alpha - 1$, and β is conjugate to α in the usual sense that $\alpha^{-1} + \beta^{-1} = 1$.

Among the Rényi entropies, the ones enjoying particular connections to the existing literature are those with α being multiples of $1/2$. In Koenker and Mizera (2010), we focused primarily on the log concave, $\alpha = 1$, case and the Hellinger, $\alpha = 1/2$, case; the latter imposes the weaker constraint that $-1/\sqrt{f}$ be concave. Here we will describe some further explorations of this approach that take us into the netherworld of $\alpha \leq 0$. We will mainly emphasize computational aspects and several illustrative applications, while briefly surveying recent work of Han and Wellner (2016), who provide considerable further theoretical development beyond the basic Fisher consistency results that appeared in Koenker and Mizera (2010).

2. Divergences, Entropies and Shape Constraints

A natural point of departure for the exploration of weaker concavity constraints is the general form of the dual formulation. However, to preserve the convexity of the dual formulation we need to consider adapting the Shannon criterion appearing in (D_1) to the chosen form of the constraint. Thus, we seek a density estimate, $\hat{f}$, with respect to a dominating measure dm (this to accommodate not only dx, but perhaps also dx restricted to some domain, or discretized on a grid, to prove later the convergence under the refinement of the latter), such that $\int f dm = 1$. The family of formulations we are looking into maximize the integral, with respect to dm, of $-\psi^*(-f)$, for some convex

function, ψ^* to be determined. The negative signs are here to conform to the notation of Koenker and Mizera (2010). The shape-constrained formulation,

$$(D) \qquad \max\Big\{-\int \psi^*(-f)dm \;\Big|\; f = \frac{d(\mathbb{Q}(X) - G)}{dm},\ G \in \mathcal{K}(X)^o\Big\},$$

has constraints identical to that of (D_1); only the objective function is open to reconsideration. A convenient family of objectives can be derived from α-divergences as described in Cichocki and Amari (2010),

$$(1) \qquad \mathcal{D}^\alpha(f,g) = \frac{1}{\alpha(\alpha-1)}\Big(\int f^\alpha g^{1-\alpha} dx - 1\Big), \quad \text{for } \alpha \notin \{0,1\}.$$

For $\alpha = 1$ and $\alpha = 0$, we have the limiting values,

$$\mathcal{D}^1(f,g) = \int f \log \frac{f}{g}\, dx$$

and

$$\mathcal{D}^0(f,g) = \int g \log \frac{g}{f}\, dx,$$

respectively. Since $\mathcal{D}^\alpha(f,g) = \mathcal{D}^{1-\alpha}(g,f)$ it follows that $\mathcal{D}^{1/2}$ is the distinguished, symmetric element. Up to various possible additive and multiplicative constants, the entropies to act as objective functions in (D) are obtained from the divergences above by simply plugging in $g \equiv 1$, and replacing dx by dm. The entropies resulting for $\alpha \notin \{0,1\}$ are variationally equivalent to the integrals of $-f^\alpha$ for $\alpha > 1$ and $\alpha < 0$, and to those of f^α for $\alpha \in (0,1)$; signs are chosen to ensure concavity. The equivalent integrands yielding neat expressions for conjugate functions used by Koenker and Mizera (2010) are as follows:

$$\begin{aligned}
\psi^*_\alpha(y) &= \begin{cases} (-y)^\alpha/\alpha & \text{for } y \le 0, \\ +\infty & \text{for } y > 0, \end{cases} && \text{for } \alpha > 1, \\
&= \begin{cases} (-y)\log(-y), & \text{for } x < 0, \\ +\infty & \text{for } x \ge 0 \end{cases} && \text{for } \alpha = 1, \\
&= \begin{cases} -(-y)^\alpha/\alpha & \text{for } y \le 0, \\ +\infty & \text{for } y > 0, \end{cases} && \text{for } \alpha < 1,\ \alpha \ne 0, \\
&= \begin{cases} -\log(-y) & \text{for } y < 0, \\ +\infty & \text{for } y \ge 0, \end{cases} && \text{for } \alpha = 0.
\end{aligned}$$

Although anticipated by earlier work of Perez (1967) and Havrda and Charvát (1967), the expression

$$\bar{\mathcal{E}}^{\alpha}(f) = \frac{1}{1-\alpha}\Big(\int f^{\alpha} dx - 1\Big) \quad \alpha \neq 0,$$

is often referred to as Tsallis entropy, Tsallis (1988). It delivers the correct sign for $\alpha > 0$ and yields Shannon entropy in the limit transition $\alpha \to 1$. It is also monotonically related, and hence variationally equivalent to, the original Rényi entropy expression, used in Koenker and Mizera (2010), for $\alpha > 0$. Another variationally equivalent formulation for $\alpha \notin \{0, 1\}$ is obtained by plugging $g \equiv 1$ into (1) and changing the sign,

$$\mathcal{E}^{\alpha}(f) = \alpha^{-1}\mathcal{E}^{\alpha}(f) = \frac{1}{\alpha(1-\alpha)}\Big(\int f^{\alpha}(x) dx - 1\Big).$$

This expression assures the correct sign for $\alpha < 0$, and also enables the limit transition to the integrand $\log f$ as $\alpha \to 0$.

Just as maximum likelihood can be interpreted as the search for a $\hat{f}_n$ closest to the empirical in Kullback-Leibler divergence subject to the log concavity constraint for $\alpha = 1$, we can interpret the solutions for $\alpha \neq 1$ as finding $\hat{f}_n$ that is closest to the empirical in Rényi divergence subject to the $(\alpha - 1)$-concavity constraint. Despite that being not explicitly expressed in the constraints of (D), $\hat{f}_n$ is automatically a probability density; it also preserves the first moment of X.

Proposition 1. *Any solution, $\hat{f}_n$, of (D) satisfies the following:*

$$\int \hat{f}_n(y) dm(y) = 1, \qquad \int y\hat{f}_n(y) dm(y) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In general, the higher-order moments are not preserved; in particular, the variance rendered by $\hat{f}_n$ is always *less or equal* to the sample variance of X – this forms a basis of an ingenious method of smoothing the shape-constrained MLE estimates of log-concave densities devised by Dümbgen and Rufibach (2009). It is, however, not clear, how the requirement $G \in \mathcal{K}(X)^o$ stipulated in the constraints of (D) translates to the fact that the solution $\hat{f}_n$ is $(\alpha-1)$-concave. To this end, we need to turn to the dual of (D), the primal of Koenker and Mizera (2010), in particular to the extremal conditions holding under strong duality.

3. The Dual of the Dual

While for $\alpha = 1$, it was the dual (D_1) of the log-concave MLE that was derived from the primal MLE formulation (P_1), for other α's it was

the other way round. The primal formulation

$$(P) \qquad \min\Big\{\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g)dm \;\Big|\; g \in \mathcal{K}(X)\Big\}$$

is obtained as the dual of (D). Here ψ are functions conjugate to ψ^*, for the particular ψ^*_α appearing in the previous section we have

$$\begin{aligned}
\psi_\alpha(x) &= \begin{cases} (-x)^\beta/\beta & \text{for } x \le 0, \\ 0 & \text{for } x > 0 \end{cases} && \text{for } \alpha > 1, \\
&= \mathrm{e}^{-x} && \text{for } \alpha = 1, \\
&= \begin{cases} +\infty & \text{for } x \le 0, \\ -x^\beta/\beta & \text{for } x > 0, \end{cases} && \text{for } \alpha < 1,\ \alpha \neq 0, \\
&= \begin{cases} +\infty & \text{for } x \le 0, \\ -\log x & \text{for } x > 0, \end{cases} && \text{for } \alpha = 0.
\end{aligned}$$

The strong duality Theorem 3.1 of Koenker and Mizera (2010) implies that the solutions $\hat{f}_n$ of (D) and $\hat{g}_n$ of (P) satisfy $\hat{f}_n = -\psi'(\hat{g}_n)$, which for ψ_α listed above translates to

$$\begin{aligned}
f(x) &= \max\{(-g(x))^{\frac{1}{\alpha-1}}, 0\} && \text{for } \alpha > 1, \\
&= \mathrm{e}^{-g(x)} && \text{for } \alpha = 1, \\
&= (g(x))^{\frac{1}{\alpha-1}} && \text{for } \alpha < 1,
\end{aligned}$$

and, due to the requirement that $\hat{g}_n \in \mathcal{K}(X)$ in (P) it follows that $\hat{f}_n$ is $(\alpha - 1)$-concave.

The other perceived virtue of the general formulation (P) is that it facilitates proof of the existence and geometric characterization of the optimal $\hat{g}_n$, through Theorems 4.1 and 2.1 of Koenker and Mizera (2010). The optimal $\hat{g}_n$ is a convex polyhedral function, with vertices lying in the datapoints X_i: it is equal to $+\infty$ (thus $\hat{f}_n$ is then equal to zero) outside the convex hull of X. This makes the solution compactly supported and the optimization problem essentially finite-dimensional, which greatly simplifies the mathematical analysis of this density estimation method. It also tempts to lend itself to "exact" methods of numerical computation – however, as we point out below, this avenue, albeit straightorward, is not always easy to follow.

As mentioned above, noteworthy values of α are those that are multiples of $1/2$. In particular, $\alpha = 2$ has a connection to the Pearson χ^2; the solution corresponds to the least-squares estimator of Groeneboom, Jongbloed, and Wellner (2001) and yields a density estimate which is

itself concave. Obviously, $\alpha = 1$ being our point of departure, is the MLE of log-concave densities, with the link to the Kullback-Leibler divergence and Shannon entropy. In Koenker and Mizera (2010), we somewhat championed $\alpha = 1/2$, linked to the Hellinger distance, the only symmetric choice among the α-divergences; the resulting density estimates are those with the convex reciprocal of the square root, the class including, in particular, all t densities with degrees of freedom greater or equal to one (and all log-concave densities as well).

The α-divergences for $\alpha < 1/2$ are reverse versions of their symmetric, about $1/2$, counterparts for $1 - \alpha$. An important instance, for which we in 2010 did not possess a reasonably stable algorithm, is that for $\alpha = 0$, corresponding to the reverse Kullback-Leibler divergence and the entropy that is sometimes called the Burg entropy. The corresponding density estimate can thus be interpreted as an empirical likelihood estimate of a density with convex reciprocal. Another noteworthy instance is that for $\alpha = -1$, corresponding to the reverse χ^2, or the Neyman χ^2. For further discussion and other α, see Koenker and Mizera (2008, 2010).

4. Computational Aspects

In this section we will briefly describe our implementation which relies crucially on the convex optimization software Mosek, Andersen (2010), and its interface Rmosek, Friberg (2012) to the R language, R Core Team (2017). Additional software and data to reproduce the computational results reported here is available in the R package REBayes, Koenker, Gu, and Mizera (2016).

While for theoretical purposes it is useful to replace dx by dm restricted to a compact (albeit large) set, for the purpose of numerical computations we need to make our variational formulation of the Rényi divergence estimator finite-dimensional; that is, to discretize it in some way. This formally corresponds to choosing a dm that approximates dx, the latter restricted to a compact set, and is concentrated on a finite set of atoms – called hereafter *evaluation points*. The finite-dimensional problem then estimates the values of $\hat{f}_n$ at these points. The most straightforward examples arise in the one-dimensional case: we take dm supported on a uniformly spaced fine grid, typically $N = 300$ to 1000 points, starting with the minimum and ending with the maximum of the X_i's, and assigning to each grid point mass $1/m$ – except perhaps for the end points, depending on whether standard rectangular or trapezoidal integration formula is to be applied.

In dimension one implementation poses few problems: the dm grid becomes an input to the estimating function solving (D). The complexity of the algorithm depends only on N, the number of evaluation points, and is independent of n, the size of data. Given the speed of the optimization algorithm, the problem of this algorithm in the one-dimensional case is seldom the size of m, which can be easily increased. When N does become prohibitively large – this situation can occur in one-dimensional problems with extreme outliers, and is almost inevitable in multi-dimensional problems – it is usually more fruitful to turn to the primal formulation (P). Since our variational problem has a solution, g, that is polyhedral, convex and piecewise linear on a triangulation – or for $d > 2$, on simplices spanned by the observed X_i's – the solution is characterized by the n function values, $\gamma_i = g(X_i)$. In fact, this amounts to making the X_i's the evaluation points, although at this point with uncertain masses attached to them; as $N = n$ in such a case, the complexity of the algorithm now depends on n, the number of data points.

There are two important difficulties that have to be tackled in this approach. The first one is enforcing the convexity of the fitted g. In dimension one this is very easy, owing to the fact that the evaluation points either come already ordered, or can be easily sorted. One has then only to make sure that any three adjacent evaluation points satisfy the convexity requirement. The number of required constraints is linear, $\mathcal{O}(N)$, in N. More generally, let V denote a diagonal matrix with diagonal elements consisting of the order statistics of the X_i, and set $A_k = D^{k+1}V$ where D denotes the differencing operator on V, then $A_1\gamma \geq 0$ imposes monotonicity, $A_2\gamma \geq 0$ convexity, and so forth.

In higher dimensions imposing convexity is somewhat more onerous, but conceptually still quite simple. As noted by Seijo and Sen (2011), we need only to impose $n(n-1)$ linear equality constraints in view of the following observation, which goes back at least to Afriat (1967, 1972).

Proposition 2. *Let $v_i \in \mathbb{R}^d$, $\gamma_i \in \mathbb{R}$, for $i = 1, 2, \cdots, n$. There is a convex function, g, such that $g(v_i) = \gamma_i$, if and only if there are $h_i \in \mathbb{R}^d$, such that*

$$(v_i - v_j)^\top h_i \leq \gamma_i - \gamma_j, \quad \textit{for all } i \textit{ and } j \neq i.$$

The geometric interpretation is quite self evident: at each vertex of the triangulation, (v_i, γ_i), there must be a supporting hyperplane in the direction of every other vertex. Order $\mathcal{O}(N^2)$ linear inequality constraints may seem burdensome, but the good news is that their number

does not depend on the dimension, d, any more; only the number of variables grows linearly with d and N, $\mathcal{O}(Nd)$, via the dimension of the subgradients h_i.

Once the mechanism for imposing convexity is in place, the only remaining challenge is to approximate the integrability constraint on the estimated density. Again, in dimension one this is would not be that much a big deal, as the integrals in the segments of adjacent ordered evaluation points can be interpolated via various numerical schemes; for instance, one can take a fine grid of points between the two, interpolate linearly the values of g in between, and use standard rectangular or trapezoidal integration formula for $\psi(g)$, which due to the convexity of ψ preserves the convex character of the optimization task. And, after all, in dimension one we do not have to bother, as we rather use (D) instead of (P) for computing the estimates.

In multi-dimensional problems, $d \geq 2$, this strategy is not so straightforward – already in the two-dimensional case, linear interpolation poses a problem: we know that g is polyhedral, but to determine how to interpolate one needs to know the triangulation. Optimizing over triangulations, however, is challenging.

A way out is to eschew linear interpolation and consider instead the right prism Riemann sums where each point x of the integration domain belongs to the base polygon containing X_i closest to x, and the height of the prism is $g(X_i)$. The polygonal tesselation of the integration domain corresponding to the nearest X_i is the well-known Voronoi tesselation. There are efficient algorithms for its construction, in arbitrary dimension, and also for the calculation of the volumes of the polygons.

In view of the strategy outlined above, with discrete dm approximating dX, this scheme can be seen as selecting the data points X_i as the evaluation points, and assigning them masses in dm equal to the volumes of the Voronoi polygons formed by the evaluation points. Experiments in the one-dimensional case, when comparisons with other methods are easily made, indicate that the approximation is good in the center of the data, as the data points are typically dense there. The polygons become larger in the tails, but this is counterbalanced by the fact that the density is smaller. If necessary, some additional evaluation points ("undata") can be added at the tails. On the other hand, for large datasets, when n is large, we may want to choose a smaller number of evaluation points, that is, we may want $N < n$, as it is N that determines the complexity of the algorithm through the $\mathcal{O}(N^2)$ convexity constraints. We may achieve this by including only some, not all, of the X_i's in the evaluation points. Indeed, we may

even avoid some X_i's completely and choose evaluation points that are somehow uniformly spread over $\mathcal{H}(X)$.

In the case that no evaluation point is equal to a particular data point X_i, a question arises how X_i is expressed in the "likelihood" term $n^{-1}\sum_{i=1}^{n} g(X_i)$. Again, there are several possibilities for such an "evaluation functional" in the one-dimensional case: either X_i is replaced by the nearest neighbor evaluation point, or its contribution is divided to that of the nearest two, with weights equal to the weights linearly interpolating X_i by the nearest two. The evaluation functionals in this fashion enter also the implementation via (D) if the evaluation points do not necessarily contain all the X_i, for instance, if they are uniformly spaced. It should be said that while both approaches return a solution that integrates to one under dm, it is only the evaluation functional via linear interpolation that leads to the estimate preserving the mean of the data, in the sense of Proposition 1.

In the higher dimensions, it is only nearest neighbor interpolation that is practical in this context, due to complications arising from the triangulation for linear interpolations. In such a way, the "likelihood" term seems to be counting the number of X_i falling into the particular polygonal base, so that the discretized computational method can be viewed as a regularization, through shape constraints, of a histogram formed by the resulting right prisms. Further details are available in the documentation and code of the R package REBayes.

5. Prospects in Asymptopia

There has been considerable recent progress in understanding the large sample behavior of shape constrained density estimators. The log concave MLE, $\hat{f}_n$, has been extensively studied with rate results established by Doss and Wellner (2016) and Kim and Samworth (2016), and showing that $\hat{f}_n$ achieves the minimax optimal rate of $\mathcal{O}(n^{-4/5})$ for squared Hellinger distance over the class of log concave densities. Even more recently, Kim, Guntuboyina, and Samworth (2016) have shown that for univariate densities such that $\log f$ is piecewise linear with k distinct segments, $\hat{f}_n$ converges in squared Kullback-Leibler divergence at rate $\mathcal{O}(\frac{k}{n}\log^{5/4} n)$, that is at essentially the parametric rate up to the log factor. This is obviously a substantial improvement over the minimax rate of $\mathcal{O}(n^{-4/5})$ achievable over the entire class of log concaves.

As noted by Han and Wellner (2016), comparatively little is known about the asymptotic behavior of the other shape constrained Rényi divergence estimators. Doss and Wellner (2016) have shown that a

maximum likelihood estimator for the class of s concave densities does not exist for any $s < -1$, i.e. $\alpha < 0$. Thus, abandoning log likelihood in favor of the Rényi entropy criterion is not simply a matter of computational convenience, but may be motivated by more fundamental considerations. Koenker and Mizera (2010) establish a basic form of Fisher consistency for the shape constrained Rényi divergence estimator; Han and Wellner (2016) provide a much more detailed analysis of the large sample behavior of the Rényi estimators with convergence results in weighted L_1 and L_∞ norms. They also provide limiting distribution theory, including results on the asymptotic cost of imposing weaker forms of concavity when stronger forms would have sufficed. A limitation of this theory at this stage is that many results are restricted to the $s > -1$, i.e. $\alpha > 0$, setting. In view of our computational results reported above, we would be eager to learn more about to what extent the theory can be extended into the netherworld of $\alpha < 0$.

6. Some Examples

In this section we present several applications of shape constrained density estimation, in an effort to illustrate the potential advantages of the weaker concavity constraints imposed by the methods we have described above.

6.1. **Annual Log Income Increments.** In an influential recent paper Guvenen, Karahan, Ozkan, and Song (2016) have estimated models of income dynamics using a very large, 10 percent, sample of U.S. Social Security records linking to Internal Revenue Service data. Their work reveals quite surprising features of annual increments in log income. In the left panel of Figure 1 we reproduce Figure 6 of Guvenen *et al.* It depicts a conventional kernel density estimate after log transformation based on their sample. There are two immediately striking features: first, the spread of the density from -4 to 4 documents a surprising volatility for some individuals we see annual changes in (unlogged) income by a factor of more than 50 in both tails; second, the shape of log density estimate is clearly *not* concave. However, when we plot $-1/\sqrt{\hat{f}(x)}$ instead of $\log \hat{f}(x)$ in the right panel of the figure, we obtain a much smoother curve that is fit almost exactly by the Hellinger, $\alpha = 1/2$, concavity constraint. As we have already noted the $\alpha = 1/2$ constraint is special in the sense that linear extrapolation in the tails corresponds to Cauchy, t_1 behavior, and in terms of our estimation criterion corresponds to the symmetric case midway between Kullback-Leibler and reverse Kullback-Leibler divergence.

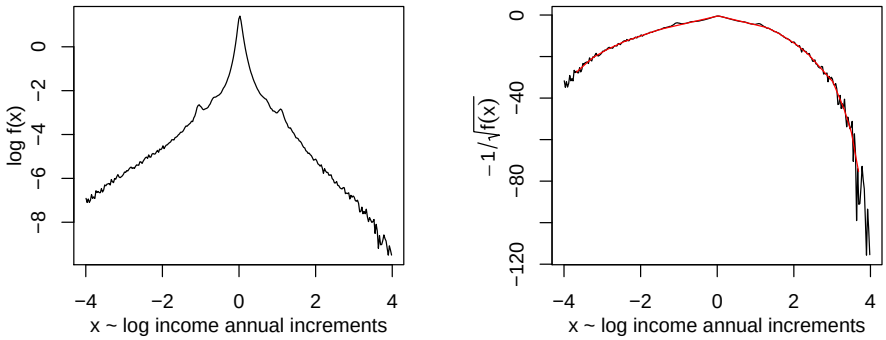

FIGURE 1. Density estimation of annual increments in log income for U.S. individuals over the period 1994-2013. The left panel of the figure reproduces a plot of the logarithm of a kernel density estimate from Guvenen *et al*, Figure 6, showing that annual income increments are clearly not log concave. However the right panel shows that $-1/\sqrt{f}$ does appear to be nicely concave and is fit remarkably well by the Renyi procedure with $\alpha = 1/2$, superimposed in red.

Permitting Cauchy tail behavior may be regarded as sufficiently indulgent for most statistical purposes, but the next example illustrates that even weaker concavity constraints paired with Rényi fitting criteria with $\alpha < 1/2$ is sometimes necessry to accommodate very sharp peaks in the target density.

6.2. **Rotational Velocity of Stars.** We reconsider the rotational velocity of stars data considered previously in Koenker and Mizera (2010). The data was taken originally from Hoffleit and Warren (1991) and is available from the R package REBayes. Figure 2 illustrates a histogram of the 3806 positive rotational velocities from the original sample of 3933. After dropping the 127 zero velocity observations, the histogram looks plausibly unimodal and we compare four distinct Rényi shape constrained estimates. The log concave, $\alpha = 1$, estimate is clearly incapable of capturing the sharp peak around $x = 18$, and even the fit for $\alpha = 0$ fails to do so. But pressing further, we see that setting $\alpha = -1$ provides much better fit by constraining $-1/f^2$ to be concave. The even weaker concavity constraint with $\alpha = -2$ seems too extreme with

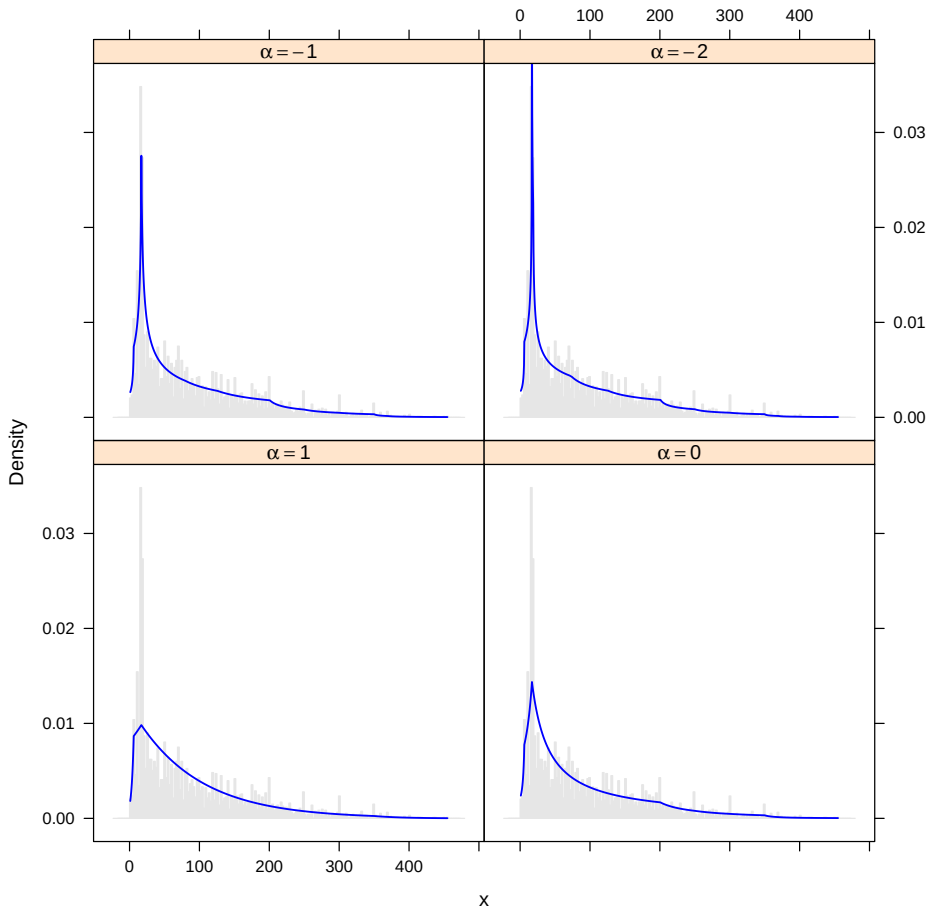

FIGURE 2. Rotational velocity of stars with three quasi concave shape constrained density estimates using the Rényi likelihood.

a substantial over-shooting of the modal peak. This example vividly illustrates that the weaker forms of concavity constraints implied by $\alpha < 0$ can be effective complements to more familiar shape constrained estimation methods when the target densities are sharply peaked or heavy tailed.

6.3. **Gosset's Criminal Anthropometrics.** Shape constraints for multivariate density estimation offers several new challenges, not the least of which is the computational challenge of finding a tractable way to represent the concavity constraints. Further details on computational methods will be provided in the next section, here we will revisit the bivariate problem of estimating the density for the well known

MacDonell (1902) data on the heights and left middle finger lengths of 3000 British criminals. This data is perhaps best known for its role in preliminary simulations reported in "Student" (1908).

Figure 3 illustrates contour plots for four different values of the constraint parameter α, together with the scatter of dithered values of the original data. Contours are labeled in units of log density. A notable feature of the data is the anomalous point at the upper region of the convex hull. This individual is extremely tall, but possesses a rather diminutive left middle finger; a grandfather of the "fanta-faced Falangist" perhaps? Although the central contours appear somewhat similar for the various α's, the labeling of the contours near this extreme point differ dramatically. When $\alpha = 1$ so we are imposing log concavity, such a person is highly anomalous and the nearest contour is labeled $\log f(x) = -20$ in this region, so $f(x) \approx 2 \times 10^{-9}$ there. When $\alpha = 0$, the corresponding contour is labeled -10, so $f(x) \approx 4.5 \times 10^{-5}$ in roughly the same region, making him look far less unusual. [Who knows? Perhaps the fellow lost a finger tip cracking a safe in Glasgow.]

7. Rényi Divergence and Norm Constrained Density Estimation

Although our original intent for using Rényi divergence as an estimation criterion was strictly pragmatic – to maintain the convexity of the optimization problem underlying the estimation while maintaining weaker forms of the concavity constraint – we would now like to briefly consider its use in norm constrained settings where the objective of penalization is smoothness of the estimated density rather than shape constraint.

There is a long tradition of norm penalized nonparametric maximum likelihood estimation of densities. Perhaps the earliest example is Good (1971) who proposed the penalty,

$$J(f) = \int ((\sqrt{f})')^2 dx,$$

which shrinks the estimated density toward densities with smaller Fisher information for location. A deeper rationale for this form of shrinkage remains obscure, and most of the subsequent literature has instead focused on penalizing derivatives of $\log f$, with the familiar cubic smoothing spline penalty,

$$J(f) = \int (\log f'')^2 dx,$$

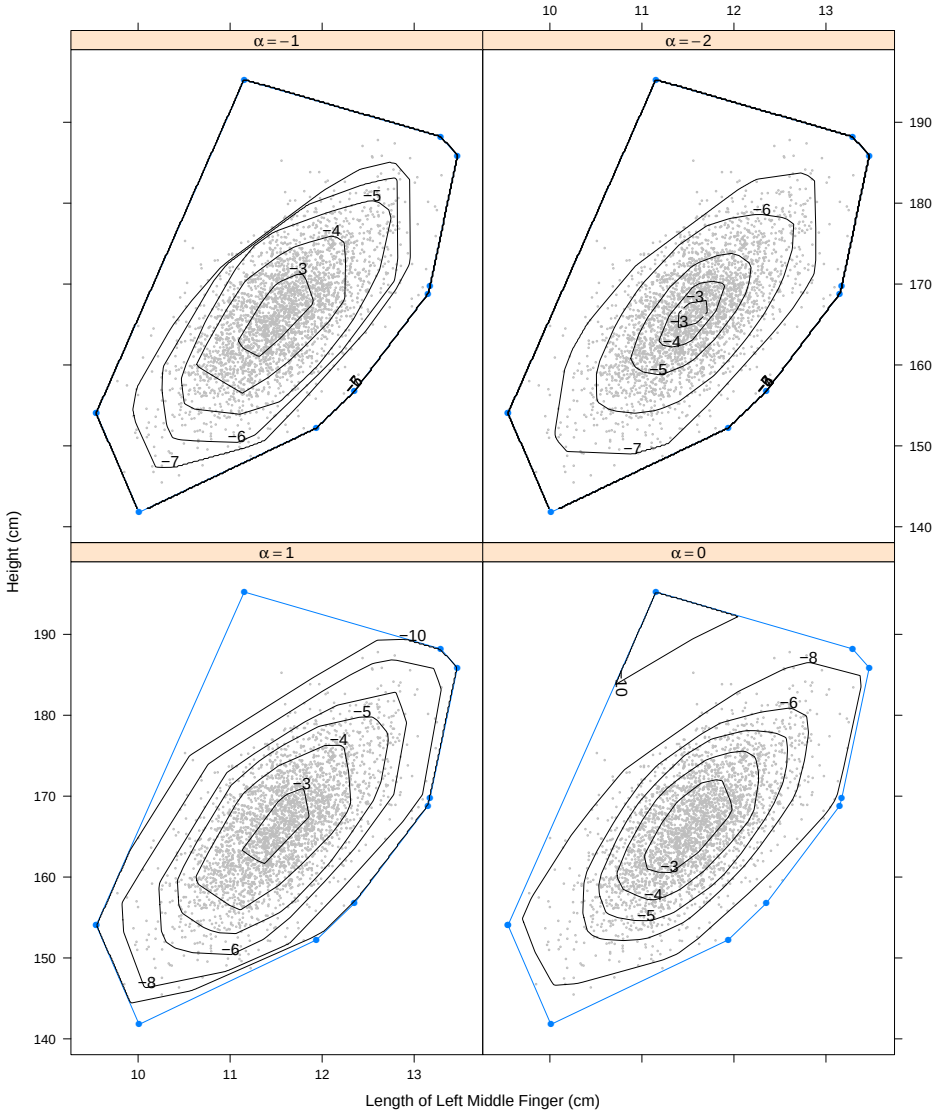

FIGURE 3. Contour Plots of British Criminal Heights and Finger Lengths: Contour estimates are based on four values of the Renyi exponent $\alpha \in \{-2, -1, 0, 1\}$ and are all labeled in units of log density. Note that the tail behavior near the anomalous point is quite different for the two Renyi exponents, and the density is also much more sharply peaked for the smaller α's. [This obviously needs some further fine tuning of the contours and their labels.]

receiving most of the attention. A notable exception is the Silverman (1982) proposal to penalize the squared L_2 norm of the *third* derivative of $\log f$ as a means of shrinking toward the Gaussian density.

Squared L_2 norm penalties are ideal for smoothly varying densities, but they abhor sharp bends and kinks, so there has also been some interest in exploring total variation penalization as a way to expand the scope of penalty methods. The taut-string methods of Davies and Kovac (2001) penalize total variation of the density itself. Koenker and Mizera (2007) describe some experience with penalties of the form,

$$J(f) = \int |(\log f)''| dx,$$

that penalize the total variation of the first derivative of $\log f$. In the spirit of Silverman (1982) the next example illustrates penalization of the total variation of the third derivative of $\log f$, again with the intent of shrinking toward the Gaussian, but in a manner somewhat more tolerant of abrupt changes in the derivatives than with Silverman's squared L_2 norm.

7.1. **Total Variation Shrinkage to the Gaussian.** In Figure 4 we illustrate a histogram based on 500 iid standard Gaussian observations, and superimpose two fitted densities estimated by penalized maximum likelihood as solutions to

$$\min_f \Big\{ -\sum_{i=1}^{n} \log f(X_i) + \lambda \int |(\log f)'''| dx \Big\},$$

for two choices of λ. For λ sufficiently large solutions to this problem conform to the parametric Gaussian MLE since the penalty forces the solution to take a Gaussian shape, but does not constrain the location or scale of the estimated density. For smaller λ we obtain a more oscillatory estimate that conforms more closely to the vagaries of the histogram.

Penalizing total variation of $(\log f)''$ as in Figure 4 raises the question: What about other Rényi exponents for $\alpha \neq 1$? Penalizing $(\log f)''$ is implicitly presuming sub-exponential tail behavior that may be better controlled by weaker Rényi penalties. To explore this conjecture we consider in the next example estimating a mixture of three lognormals.

7.2. **Lognormal Mixtures.** Figure 5 illustrates a histogram based on 500 observations from a mixture of three 3-parameter lognormals with the population density superimposed in red. This density serves

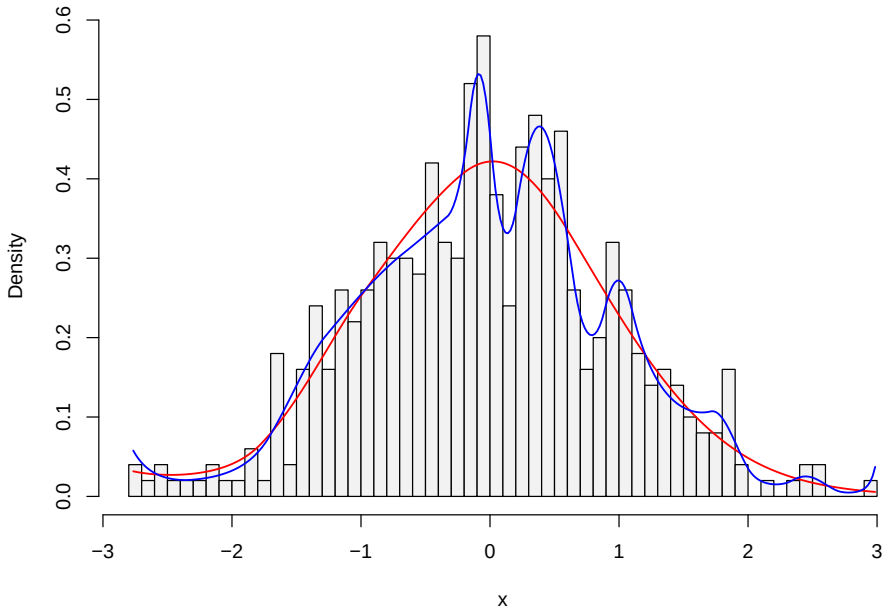

FIGURE 4. Gaussian histogram based on 500 observations and two penalized maximum likelihood estimates with total variation norm penalty and $\lambda \in \{0.5 \times 10^{-4}, 0.5 \times 10^{-6}\}$.

as a cautionary illustration of how difficult it can be to choose an effective bandwidth for conventional fixed bandwidth kernel estimation. A fixed bandwidth sufficiently small to distinguish the two left-most modes is incapable of producing a smooth fit to the upper mode, and this makes adaptive bandwidth kernel methods difficult due to poor performance of the pilot estimate. Logspline methods as proposed by Kooperberg and Stone (1991) perform much better in such cases, but in our experience they can be sensitive to knot selection strategies. The methods under consideration here are allied more closely to the smoothing spline literature, and thereby circumvent the knot selection task, but in so doing introduce new knobs to turn and buttons to push. Not only do we need to choose the familiar λ, there is now a choice of the order of the derivative in the penalty, and the Rényi exponent, α, determining the transformation of the density. We would argue that these choices are more easily adapted to particular applications, but others may feel differently. From a Bayesian perspective, however, it

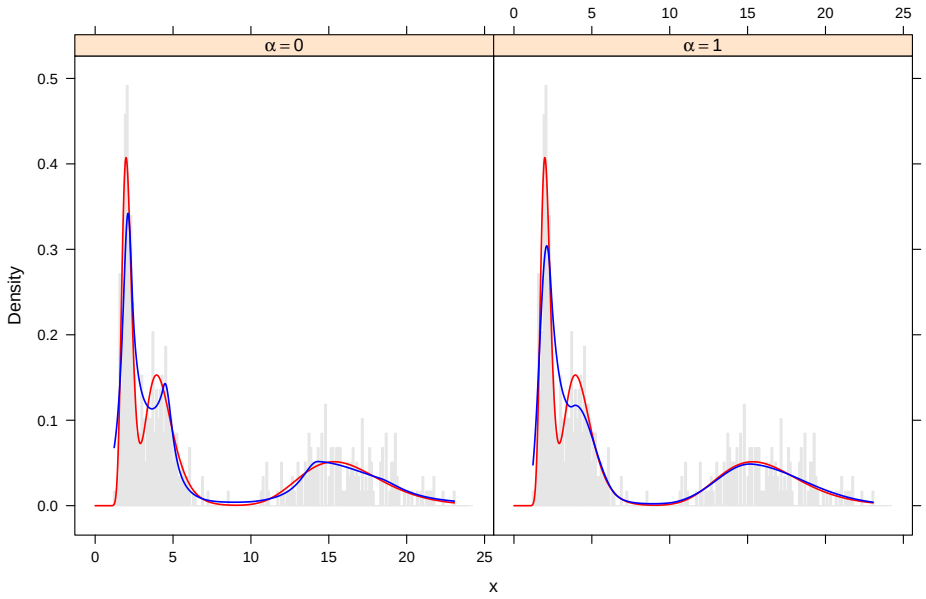

FIGURE 5. Mixture of three 3-parameter lognormals with histogram and two Rényi likelihood estimates with total variation (L_1 norm) penalty with $\alpha \in \{0, 1\}$ based on 500 iid observations and penalty parameter, $\lambda = 9$. The true density is depicted in red and the estimated density is in blue.

seems indisputable that more diversity in the class of computationally tractable prior specifications is desirable.

Examining Figure 5 we see that the $\alpha = 1$ maximum likelihood estimate is a bit too smooth, barely able to find the second mode, whereas the $\alpha = 0$ solution is somewhat better at capturing the first mode, and also better at identifying the second mode. Both methods produce an excellent fit to the third mode, almost indistinguishable from the true density.

8. CONCLUSION

Shape constrained nonparametric density estimation offers a valuable compromise between restrictive parametric methods and conventional smoothing methods. Log concavity is a natural constraint in some applications and can be efficiently implemented by maximum likelihood. In other applications it can be advantageous to impose weaker forms of the concavity constraint, and for this purpose it is convenient

to pair constraints that require that $-1/f^\alpha$ be concave with a Rényi α-divergence criterion for goodness of fit. We have illustrated this approach with several examples taken from economics, astronomy and anthropometrics. We also briefly discuss related methods that pair norm-based smoothing penalties with the Rényi divergence estimation criterion.

Many problems remain for future research. Among these, adaptive choice of α is obviously prominent. Despite the impressive accomplishments of Han and Wellner (2016), much is still unknown about the limiting asymptotic behavior of these estimators. In particular, the region of $\alpha < 0$ remains to be charted. We look forward to future progress on these and other aspects of such methods.

REFERENCES

AFRIAT, S. N. (1967): "The construction of utility functions from expenditure data," *International Economic Review*, 8(1), 67–77.

——— (1972): "Efficiency estimation of production functions," *International Economic Review*, 13(3), 568–598.

ANDERSEN, E. D. (2010): "The Mosek Optimization Tools Manual, Version 6.0," Available from `http://www.mosek.com`.

AVRIEL, M. (1972): "r-Convex Functions," *Math. Programming*, 2, 309–323.

BIRGÉ, L. (1997): "Estimation of unimodal densities without smoothness assumptions," *The Annals of Statistics*, 25, 970–981.

CICHOCKI, A., AND S. AMARI (2010): "Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities," *Entropy*, 12, 30–35.

COX, D. (1966): "Notes on the analysis of mixed frequency distributions," *The British Journal of Mathematical and Statistical Psychology*, 19, 39–47.

CULE, M., R. SAMWORTH, AND M. STEWART (2010): "Maximum likelihood estimation of a multi-dimensional log-concave density," *Journal of the Royal Statistical Society (B)*, 72, 545–607.

DAVIES, P. L., AND A. KOVAC (2001): "Local extremes, runs, strings and multiresolution," *The Annals of Statistics*, 29, 1–65.

DOSS, C. R., AND J. A. WELLNER (2016): "Global rates of convergence of the MLEs of log-concave and s-concave densities," *Annals of Statistics*, 44, 954–981.

DÜMBGEN, L., AND K. RUFIBACH (2009): "Maximum likelihood estimation of a log-concave density: Basic properties and uniform consistency," *Bernoulli*, 15, 40–68.

EGGERMONT, P. P. B., AND V. N. LARICCIA (2001): *Maximum penalized likelihood estimation, Vol. I: Density estimation*. Springer, New York.

FRIBERG, H. A. (2012): "Users Guide to the R-to-Mosek Interface," Available from `http://rmosek.r-forge.r-project.org`.

GOOD, I. J. (1971): "A nonparametric roughness penalty for probability densities," *Nature*, 229, 29–30.

GRENANDER, U. (1956): "On the theory of mortality measurement, part II.," *Skandinavisk Aktuarietidskrift*, 39, 125–153.

GROENEBOOM, P., G. JONGBLOED, AND J. A. WELLNER (2001): "Estimation of a Convex Function: Characterizations and Asymptotic Theory," *The Annals of Statistics*, 29(6), 1653–1698.

GUVENEN, F., F. KARAHAN, S. OZKAN, AND J. SONG (2016): "What Do Data on Millions of U.S. Workers Reveal about Life-Cycle Earnings Dynamics?," Federal Reserve Bank of New York Staff Reports.

HAN, Q., AND J. A. WELLNER (2016): "Approximation and estimation of s-concave densities via Rényi divergences," *Annals of Statistics*, 44, 1332–1359.

HARDY, G. H., J. E. LITTLEWOOD, AND G. PÓLYA (1934): *Inequalities*. Cambridge U. Press, London.

HARTIGAN, J., AND P. HARTIGAN (1985): "The dip test of unimodality," *Annals of Statistics*, 13, 70–84.

HAVRDA, J., AND F. CHARVÁT (1967): "Quantification method of classification processes: Concept of structural a-entropy," *Kybernetika*, 3, 30–35.

HOFFLEIT, D., AND W. H. WARREN (1991): *The Bright Star Catalog (5th ed.)*. Yale University Observatory, New Haven.

KIM, A. K. H., A. GUNTUBOYINA, AND R. J. SAMWORTH (2016): "Adaptation in log-concave density estimation," Arxiv preprint.

KIM, A. K. H., AND R. J. SAMWORTH (2016): "Global rates of convergence in log-concave density estimation," *The Annals of Statistics*, 44, 2756–2779.

KOENKER, R., J. GU, AND I. MIZERA (2016): *REBayes: Empirical Bayes Estimation and Inference in R*R package version 0.85.

KOENKER, R., AND I. MIZERA (2007): "Density estimation by total variation regularization," in *Advances in statistical modeling and inference, Essays in honor of Kjell A. Doksum*, ed. by V. Nair, pp. 613–633. World Scientific, Singapore.

——— (2008): "Primal and dual formulations relevant for the numerical estimation of a probability density via regularization," in *Tatra Mountains Mathematical Publications*, ed. by A. Pázman, J. Volaufová, and V. Witkovský, vol. 39, pp. 255–264. Slovak Academy of Sciences, Proceedings of the conference ProbaStat '06 held in Smolenice, Slovakia, June 5-9, 2006.

——— (2010): "Quasi-concave density estimation," *Annals of Statistics*, 38(5), 2998–3027.

KOOPERBERG, C., AND C. J. STONE (1991): "A Study of Logspline Density Estimation," *Computational Statistics and Data Analysis*, 12, 327–347.

MACDONELL, W. (1902): "On Criminal Anthropometry and the Identification of Criminals," *Biometrika*, 1, 177–227.

PAL, J. K., M. WOODROOFE, AND M. MEYER (2007): "Estimating a Polya frequency function," in *Complex datasets and inverse problems: tomography, networks and beyond*, ed. by R. Liu, W. Strawderman, and C.-H. Zhang, vol. 54 of *IMS Lecture Notes-Monograph Series*. Institute of Mathematical Statistics.

PEREZ, A. (1967): "Information-theoretic risk estimates in statistical decision," *Kybernetika*, 3, 1–21.

PRAKASA RAO, B. (1969): "Estimation of a Unimodal Density," *Sankhyā (A)*, 31, 23–36.

R CORE TEAM (2017): *R: A Language and Environment for Statistical Computing*R Foundation for Statistical Computing, Vienna, Austria.

SEIJO, E., AND B. SEN (2011): "Nonparametric least squares estimation of a multivariate convex regression function," *The Annals of Statistics*, 39, 1633–1657.

SILVERMAN, B. (1981): "Using kernel density estimates to investigate multimodality," *Journal of the Royal Statistical Society (B)*, 43, 97–99.

SILVERMAN, B. W. (1982): "On the estimation of a probability density function by the maximum penalized likelihood method," *Ann. Statist.*, 10, 795–810.

"STUDENT" (1908): "The Probable Error of the Mean," *Biometrika*, 6, 1–23.

TSALLIS, C. (1988): "Possible generalizations of Boltzmann-Gibbs statistics," *Journal of Statistical Physics*, 52, 479–487.

WALTHER, G. (2002): "Detecting the presence of mixing with multiscale maximum likelihood," *Journal of the American Statistical Association*, 97, 508–513.

——— (2009): "Inference and modeling with log-concave distributions," *Statistical Science*, 24(3), 319–327.