Koenker, Roger; Mizera, Ivan

Working Paper

Shape constrained density estimation via penalized Rényi divergence

cemmap working paper, No. CWP54/18

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

cemmap

centre for microdata methods and practice

Shape constrained density estimation via penalized Rényi divergence

Roger Koenker
Ivan Mizera

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP54/18

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.

In contrast to our prior work, Koenker and Mizera (2010), that focused on the log concave ($\alpha = 1$) and Hellinger ($\alpha = 1/2$) constraints, here we describe methods enabling imposition of even weaker, $\alpha \leq 0$ constraints. An alternative formulation of the concavity constraints for densities in dimension $d \geq 2$ also significantly expands the applicability of our proposed methods for multivariate data. Finally, we illustrate the use of the Renyi divergence criterion for norm-constrained estimation of densities in the absence of a shape constraint.

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)

Version: April 20, 2018. The authors would like to express their appreciation to Bodhi Sen for suggesting the reformulation of the multivariate convexity constraint employed here, and to both referees for their constructive comments.

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 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.

The evolutionary development of Koenker and Mizera (2010), no longer followed here, began with the variational formulation of the log-concave MLE problem for given $X = \{X_1, \cdots, X_n\}$, with $X_i \in \mathbb{R}^d$:

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

with $\mathcal{K}(X)$ denoting 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 $\hat{g}(x) = -\infty$, meaning 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) may be interpreted as 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 e.g. 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 classes are provided by s-concave functions. Loosely speaking, a function is called *s-concave*, for $s > 0$, if its s-th power is concave. More precisely, a non-negative, real function f, defined on a convex set $C \subset \mathbb{R}^d$ is s-concave, if there is a convex function g such that

$$f = \begin{cases} (-g)^{1/s} & \text{for } s > 0, \\ \mathrm{e}^{-g} & \text{for } s = 0, \\ g^{1/s} & \text{for } s < 0. \end{cases}$$

This is equivalent to the definition of Avriel (1972) used by Koenker and Mizera (2010), who define f to be s-concave in terms of the means of order s, as defined, by Hardy, Littlewood, and Pólya (1934). Note that log-concave functions are 0-concave, and concave functions are 1-concave; also, if f is s-concave, then f is also s'-concave for any $s' < s$. The limiting class of $-\infty$-concave, the union of all s-concave classes for all $s \in \mathbb{R}$, is the class of *quasi-concave* functions – functions with upper level sets convex. In the one-dimensional case, for $d = 1$, this class is identical with that of unimodal functions.

Once log-concavity is imposed, maximizing log likelihood in (P_1) appears to be especially convenient, as it leads to a convex program with the only nonlinearity arising from the integrability constraint. However, when weaker forms of concavity are considered, it proves more convenient to adapt the fitting criterion – in particular to retain the convexity of the optimization formulation. This was already apparent in an earlier work of Groeneboom, Jongbloed, and Wellner (2001) who employed 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) suggests the possibility of replacing it by one of the abundant assortment of alternative divergences. Koenker and Mizera (2008, 2010) pointed out that for s-concave densities, this turns out to produce a lucky match. They proposed replacing the Shannon entropy in (D_1) by a variationally equivalent form of the Rényi entropy, the move that yielded a family of new dual and primal pairings,

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

and

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

The Rényi exponent α here corresponds to Avriel's $s = \alpha - 1$, and β is conjugate to α in the usual sense: $1/\alpha + 1/\beta = 1$.

Among the Rényi entropies, the ones enjoying particular connections to the existing literature happen to be are those with α being a multiple of 1/2. Koenker and Mizera (2010) 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 describe some further explorations of this approach that take us into the netherworld of $\alpha \leq 0$. Apart from emphasizing computational aspects, we highlight applications from the diverse fields of economics, astronomy and anthropometry where the methods exhibit special salience. The recent work of Han and Wellner (2016), and Laha and Wellner (2017) provides considerable further theoretical development and justification for the pairing of the Rényi criterion with weaker forms of the concavity constraints.

Existence, uniqueness and Fisher consistency results are extended to this broader class of quasi-concave density estimators. An alternative formulation of concavity constraints for densities in dimension $d \geq 2$ is shown to significantly expand the applicability of the methods for multivariate data. Finally, we illustrate the use of the Renyi divergence criterion for norm constrained estimation of densities without a shape constraint. An implementation of all the methods described here is available in the R package MeddeR, Koenker and Mizera (2017), which relies on the convex optimization software Mosek, Andersen (2010), and its R interface Rmosek, Friberg (2012); further details are provided in Section 4.

2. Divergences and Entropies

A natural point of departure for the exploration of weaker concavity constraints is the general form of the dual formulation. We want to adapt the Shannon criterion appearing in (D_1) to the chosen form of the shape constraint, at the same time preserving the convexity of the dual formulation. Maintaining our definitions of $\mathcal{K}(X)^o$, $\mathbb{Q}(X)$, and $\mathcal{H}(X)$, we consider the shape-constrained formulation

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

which has constraints identical to those of (D_1), only the objective function is now open to reconsideration.

Another minor variation is that the dominating measure dx is generalized to dm: this may appear not that essential, but it offers a convenient bridge between the theory, which favors dx, to more pragmatic choices like dx restricted to a bounded domain, or versions of the latter discretized to a fine grid; we should stress that when it comes to instances of dm, we always have in mind those quite close to the original dx. In what follows, we will assume that dm is a regular Borel (nonnegative) measure which is either finite so we may without loss of generality require $\int f dm = 1$, or at least assigns finite values to bounded sets as does dx. The following proposition holds true with any restrictions on dm – just as a consequence of our definitions. Proofs are deferred to the Appendix.

Proposition 1. *Any f satisfying the constraints of (D) – in particular, the optimal $\hat{f}$ – satisfies the following:*

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

In general, the higher-order moments are not preserved; in particular, the variance rendered by $\hat{f}$ is always *less or equal* to the sample variance of X; this underlies an ingenious method of smoothing the shape-constrained MLE for log-concave densities devised by Dümbgen and Rufibach (2009).

Now, a convenient family of objectives for (D) 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 variational equivalence, that is, up to monotone transformations that do not affect the outcome of the optimization problem (D), the entropies to act as objective functions in (D) are obtained from the divergences above by taking $g \equiv 1$ and changing the sign (before finally replacing dx by dm). This yields

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

the expression that assures the correct sign for $\alpha < 0$, and also enables the limit transition to the integrand $\log f$ as $\alpha \to 0$. Another variationally equivalent form is, for $\alpha \neq 0$,

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

in literature often referred to as the Tsallis (1988) entropy; see, however, Perez (1967) and Havrda and Charvát (1967). The latter delivers the correct sign for $\alpha > 0$ and yields Shannon entropy in the limit transition $\alpha \to 1$. It is monotonically related, and hence variationally equivalent to the original Rényi entropy expression, the expression that predates all the others; for $\alpha < 1$, $\alpha \neq 0$, the latter is equal to

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

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

$\alpha \in (0,1)$. Koenker and Mizera (2010) used equivalent integrands

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

It should be stressed that minimum divergence estimation methods remain an active field of study, see e.g. Ghosh (2015) and Basu, Harris, Hjort, and Jones (1998). The survey of Broniatowski and Vajda (2012) lists "four types of point estimators based on minimization of information-theoretic divergences between hypothetical and empirical distributions", other relevant references include Broniatowski and Keziou (2006), Broniatowski and Keziou (2009), and Liese and Vajda (2006). The divergence estimators discussed here differ not only in their focus on nonparametric density estimation rather than parametric models, but more importantly, they do not seek minimum distance of the estimate to the empirical distribution; instead, as already noted above, (D) seeks f minimizing the distance to the uniform distribution, on $\mathcal{H}(X)$, among admissible distributions specified by the constraint that depends on the data through the empirical distribution $\mathbb{Q}(X)$. For further explanation of this aspect in the context of norm-constrained density estimation, see Koenker and Mizera (2006).

3. The Dual of the Dual and Shape Constraints

Koenker and Mizera (2010) derived (D_1) for $\alpha = 1$, the dual of the log-concave MLE, from the primal MLE formulation (P_1); for the other α's they proceeded the other way round. Now, we obtain the primal formulation

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

as the dual of (D), functions conjugate to ψ^* being denoted by ψ. For the particular ψ_α^* from the previous section we have

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

Revisiting the proof of strong duality in Koenker and Mizera (2010) reveals that their Theorem 3.1 can accommodate general dm; the details are given in the Appendix, in the proof of the following proposition. Recall that the standing assumption for dm is that it assigns finite values to bounded sets; as far as ψ is concerned, we assume hereafter that it is a nonincreasing, proper convex function on $\mathbb{R}$ its domain containing $(0, +\infty)$. In fact, all the assumptions we make here and later about ψ are satisfied by all the ψ_α above.

Proposition 2. *Suppose that ψ is differentiable on its domain. The solutions $\hat{f}$ of (D) and $\hat{g}$ of (P) satisfy*

$$\hat{f} = -\psi'(\hat{g}). \tag{2}$$

The proposition reveals how the requirement $G \in \mathcal{K}(X)^o$ stipulated in the constraints of (D) translates to the crucial fact that the solutions $\hat{f}$ are $(\alpha-1)$-concave. For ψ_α listed above, (2) translates to

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

in view of the requirement in (P) that $\hat{g} \in \mathcal{K}(X)$ this means that $\hat{f}$ is $(\alpha-1)$-concave. If ψ is differentiable, the monotonicity of ψ implies the existence of the inverse of ψ', from $(0, +\infty)$ to the domain of ψ, hereafter denoted as φ. The conjugate of ψ can be obtained as its Legendre transformation,

$$\psi^*(y) = -y\varphi(y) - \psi(\varphi(y)), \tag{3}$$

and with $g = \varphi(-f)$ then, (P) can be rewritten in terms of the estimated f as

$$(F) \qquad \min\Big\{\frac{1}{n}\sum \varphi(-f(X_i)) + \int \psi(\varphi(-f))dm \;\Big|\; \varphi(-f) \in \mathcal{K}(X)\Big\}.$$

The formulation (P) also leads to the geometric characterization of the optimal $\hat{g}$. For $dm = dx$, Theorem 2.1 of Koenker and Mizera (2010) asserts that the optimal $\hat{g}$ belongs to $\mathcal{G}(X)$, the collection of all polyhedral convex functions of the form

$$(4) \qquad g_{(X,Y)}(x) = \inf\Big\{\sum_{i=1}^{n} \lambda_i Y_i \mid x = \sum_{i=1}^{n} \lambda_i X_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0\Big\}.$$

where, as before, $X = (X_1, X_2, ..., X_n)$ are datapoints, $X_i \in \mathbb{R}^d$ and $Y = (Y_1, Y_2, ..., Y_n)$ are function values, $Y_i \in \mathbb{R}$, at those – of the function which is the lower convex hull of the points $(X_i, Y_i) \in \mathbb{R}^{d+1}$. The convention $\inf \emptyset = +\infty$ used in (4) means that the domain of $g_{(X,Y)}$ is equal to $\mathcal{H}(X)$; that is, $\hat{g}$ is equal to $+\infty$ outside $\mathcal{H}(X)$, which, in view of the transformations listed above means that $\hat{f}$ is equal to zero outside of $\mathcal{H}(X)$. This fact facilitates an extension of Theorem 4.1 of Koenker and Mizera (2010), which was originally proved under the assumption that ψ is bounded from below, an assumption satisfied for positive values of α, which Koenker and Mizera (2010) focused on. Regarding dm, we again need only that it assigns finite values to bounded sets; for ψ, we have to assume a bit more beyond the standing assumption of monotonicity and convexity.

Proposition 3. *Suppose that the limit of $\psi(y + \tau x)/\tau$, for $\tau \to +\infty$ and every real y, is respectively $+\infty$ and 0, for $x < 0$ and $x > 0$. The solution of (P) then exists in $\mathcal{G}(X)$; it is unique when dm assigns positive measure to every open set within $\mathcal{H}(X)$ – in particular, for $dm = dx$.*

Note that the assumptions are still true for every ψ_α listed above: while the limit $+\infty$ for $x < 0$ has to be explicitly calculated for ψ_α with $\alpha > 0$, it is automatic for those with $\alpha < 0$, as then $\psi_\alpha(y) = +\infty$ for $y < 0$. On the other hand, the limit for $x > 0$ has to be explicitly calculated to be 0 for ψ_α with $\alpha < 0$; for those with $\alpha > 0$, it is automatic by the fact that $\psi_\alpha(y) \to 0$ for $y \to +\infty$.

The polyhedral characterization of $\hat{g}$ for general dm is "non-exclusive", in particular dm arising from discretizations, do not generally assign positive measure to every open set within $\mathcal{H}(X)$. Nonetheless, the proof of Theorem 2.1 of Koenker and Mizera (2010) shows that the optimal $\hat{g}$ still *can be found* in $\mathcal{G}(X)$. We will thus hereafter assume that any solutions $\hat{g}$ of (P) is from $\mathcal{G}(X)$ – this is hardly a restriction, as precisely such solutions are those that are obtained by practical implementations. We will use the characterization of solutions to establish a continuity property that provides a theoretical justification for our "approximate" computational strategies. Note that pointwise convergence, hereafter just "convergence", of polyhedral $\hat{g}_\nu \in \mathcal{G}(X)$ is equivalent to their uniform convergence, and also implies uniform convergence of their corresponding transforms $\hat{f}_\nu$, the solutions of the related versions of (D). The notion of weak convergence of measures we use here is that of Billingsley (1968), in the treatises of functional-analytic flavor often referred to as that of weak* topology. The proof of the following proposition is facilitated by the convexity, not that much that of solutions, but that of the objective functions involved, along the lines of well-known principles exemplified e.g. by Hjort and Pollard (2011). The latter reference indicates that error bounds are also possible; this is left for future work.

Proposition 4. *Suppose that dm_ν is a sequence of measures converging weakly to dm_0. Any accumulation point, for $\nu \to \infty$, of any sequence of solutions $\hat{g}_\nu$ of (P) with $dm = dm_\nu$ is a solution of (P) with $dm = dm_0$. In particular, $\hat{g}_0$ is a limit of any such sequence, if it is a unique solution of (P) with $dm = dm_0$.*

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, $\alpha = 1$, our point of departure, yields the MLE of log-concave densities, with the link to the Kullback-Leibler divergence and Shannon entropy. Koenker and Mizera (2010) 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 (1967) 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 MeddeR, Koenker and Mizera (2017).

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 $\hat{f}$ 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 $1/N$ – 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 N, 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 5. *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, 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 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 MeddeR.

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, $\hat{f}_n$ – in this section, we emphasize the dependence on the sample size, n, in the notation – 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) addressed the problem from the point of the asymptotics for $n = +\infty$, rather than $n \to \infty$, establishing Fisher consistency for the shape constrained Rényi divergence estimators, with $dm = dx$, and $\alpha > 0$. In the latter case, all the relevant integrands are bounded from below by 0; nonetheless, the proof of and the discussion following their Theorem 4.2 indicates that the essential requirement for more general α is the integrability of $\psi^*(-f_0)$. The objective function of (D) has to be finite for the underlying f_0, that is, for the density f_0 governing the stochastic behavior of $X_1, X_2, \ldots, X_n$. Such an assumption does not create a problem for dm with bounded domains – but if we eschew philosophical detours and adhere to the usual mathematical formalism of $dm = dx$, we may have to concede that this condition may be almost necessary. The finiteness of the Shannon entropy, the fact that the integral of $-f_0 \log f_0$ exists and is finite, is pretty much the minimal standard component of the consistency proofs for maximum likelihood estimators – as in Assumption 6 of Wald (1949), or, in a bit stronger version, page 62 of van der Vaart (1998).

For the reader's convenience, we restate the result here, in the strengthened form applicable to all α. Our starting point is the transformed primal formulation (F), with $n^{-1} \sum_{i=1}^{n} \varphi(-f(X_i))$ interpreted as the integral of $\varphi(-f)$ with respect to the empirical probability $\mathbb{Q}(X)$. Fisher consistency then concerns the objective function,

$$\Phi_0(f) = \int \varphi(-f) f_0 + \psi(\varphi(-f))\, dm, \tag{5}$$

arising from (F) by replacing $d\mathbb{Q}(X)$ by $f_0\, dm$. Using the strategy of Huber (1967), we add to the objective function of (F) a term depending only on f_0; this yields an equivalent minimization problem, in terms of f, with the objective function

$$\int \left(\varphi(-f) + \frac{\psi^*(-f_0)}{f_0} \right) d\mathbb{Q}(X) + \int \psi(\varphi(-f)) dm, \tag{6}$$

for which we are able to establish the desired result, with the objective function

$$\tilde{\Phi}_0(f) = \int \varphi(-f) f_0 + \psi^*(-f_0) + \psi(\varphi(-f))\, dm, \tag{7}$$

resulting now from (6) by replacing $d\mathbb{Q}(X)$ by $f_0\, dm$. It is necessary to sort out some subtle issues here: although the relationship between (F) and (6) is clear apart from the possibility that $f_0(X_i) = 0$ for some i, a problem that we consider decidedly minor, we cannot a priori exclude certain other problems arising with (5) and (7). The integrals may not exist, and even when they do, they could be equal to $+\infty$ for all f, making the resulting Fisher consistency result somewhat trivial. Koenker and Mizera (2010) concentrated on the cases when $\alpha > 0$, with terms like $\psi(\phi(-f))$ bounded from below by 0, when such possibilities were excluded. The following result establishes Fisher consistency for the full range of α.

Proposition 6. *Suppose that ψ is differentiable on its domain. For all f, the integral (7) defining the function $\tilde{\Phi}_0(f)$ exists, and $0 = \tilde{\Phi}_0(f_0) \leq \tilde{\Phi}_0(f)$, with the possibility that $\tilde{\Phi}_0(f) = +\infty$ for some f. If the integral*

$$\int \psi^*(-f_0)\, dm$$

exists and is finite, then the integral (5) *defining the function* $\Phi_0(f)$ *exists for all* f*, for* $f = f_0$ *it is finite, and for all* f *we have the inequality* $\Phi_0(f_0) \leq \Phi_0(f)$*, again with the possibility that* $\Phi_0(f) = +\infty$ *for some* f*.*

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.

Permitting Cauchy tail behavior may be regarded as sufficiently indulgent for most statistical purposes, but the next example illustrates

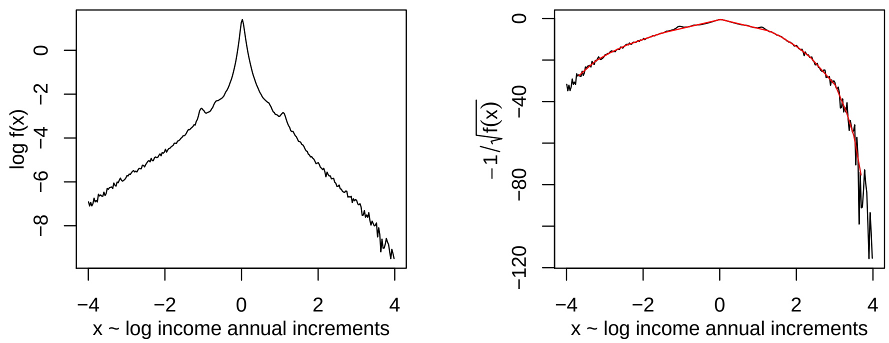

Figure 1. Density estimation of annual increments in log income for U.S. individuals over the period 1994-2013. The left panel of the figure reproduces a plot of the logarithm of a kernel density estimate from Guvenen *et al*, Figure 6, showing that annual income increments are clearly not log concave. 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.

that even weaker concavity constraints paired with Rényi fitting criteria with $\alpha < 1/2$ is sometimes necessary 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 MeddeR. 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 a substantial over-shooting of the modal peak. This example vividly illustrates that the weaker forms of concavity constraints implied by

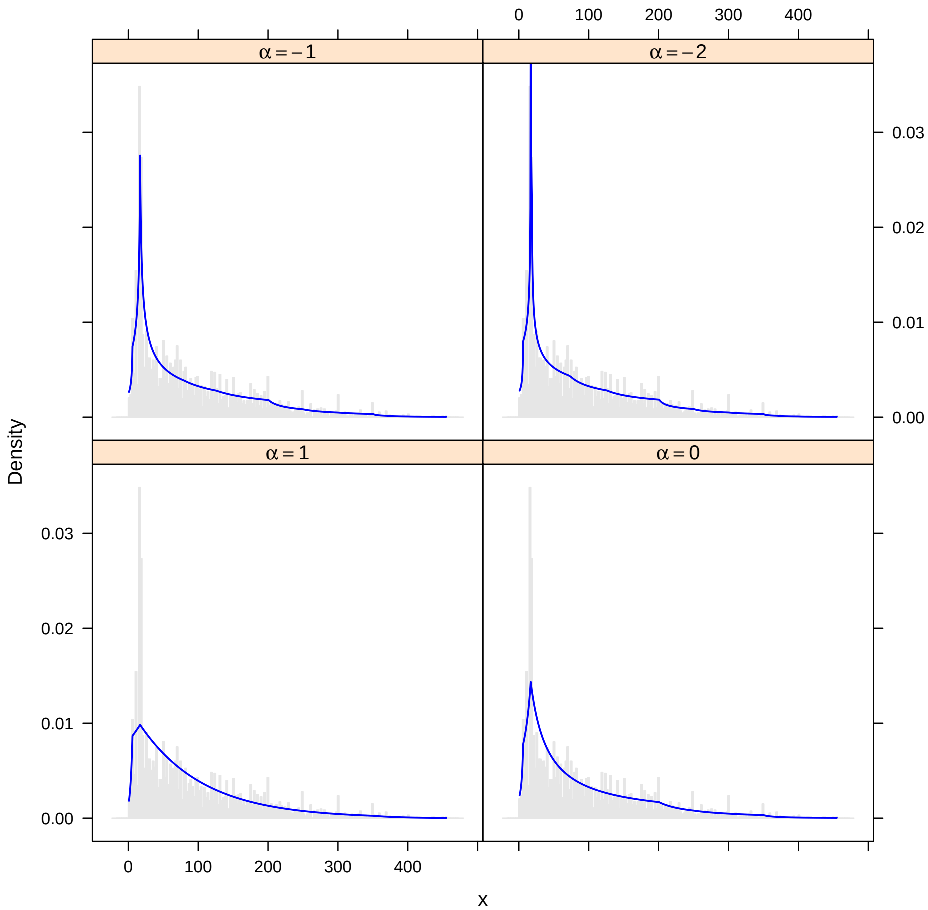

FIGURE 2. Rotational velocity of stars with three quasi concave shape constrained density estimates using the Rényi likelihood.

$\alpha < 0$ 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.

7. Rényi Entropies in 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,

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

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.

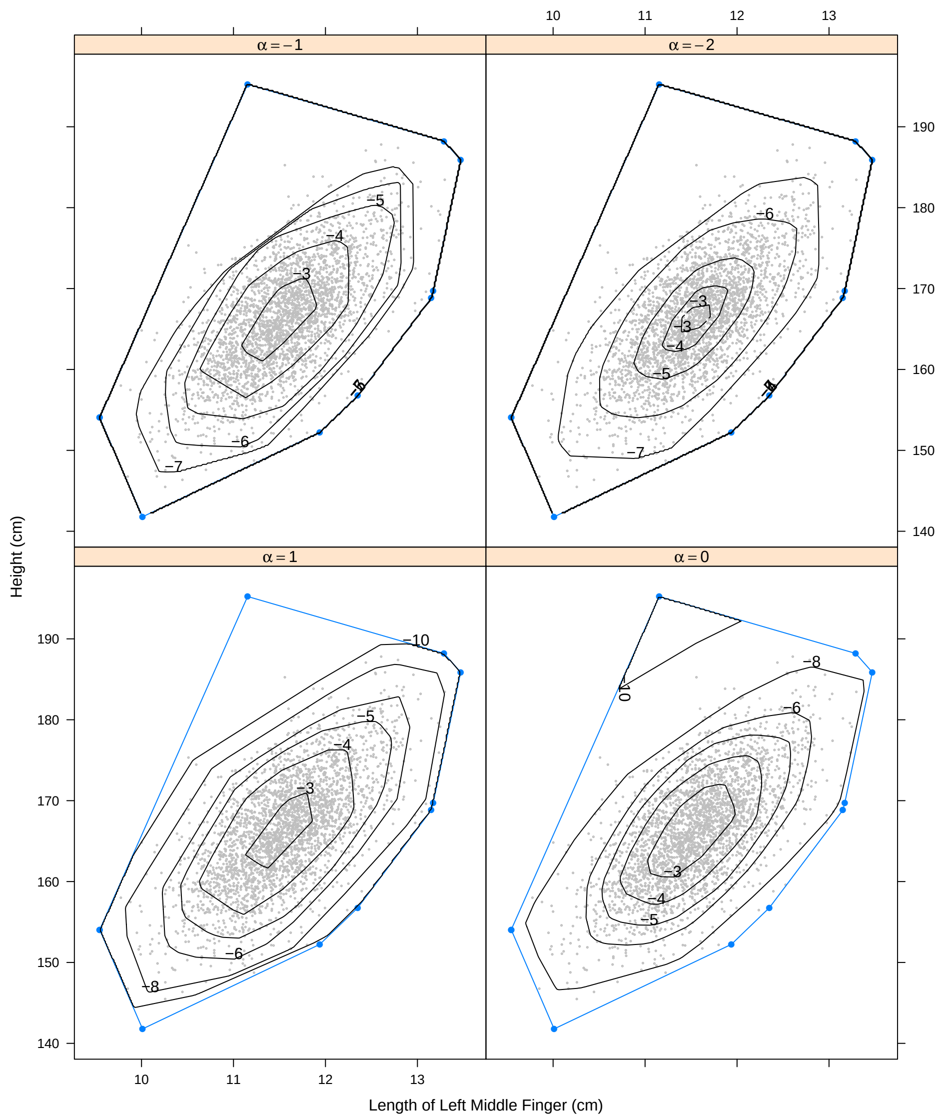

FIGURE 3. Contour Plots of British Criminal Heights and Finger Lengths: Contour estimates are based on four values of the Renyi exponent $\alpha \in \{-2, -1, 0, 1\}$ and are all labeled in units of log density. Note that the tail behavior near the anomalous point is quite different for the two Renyi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

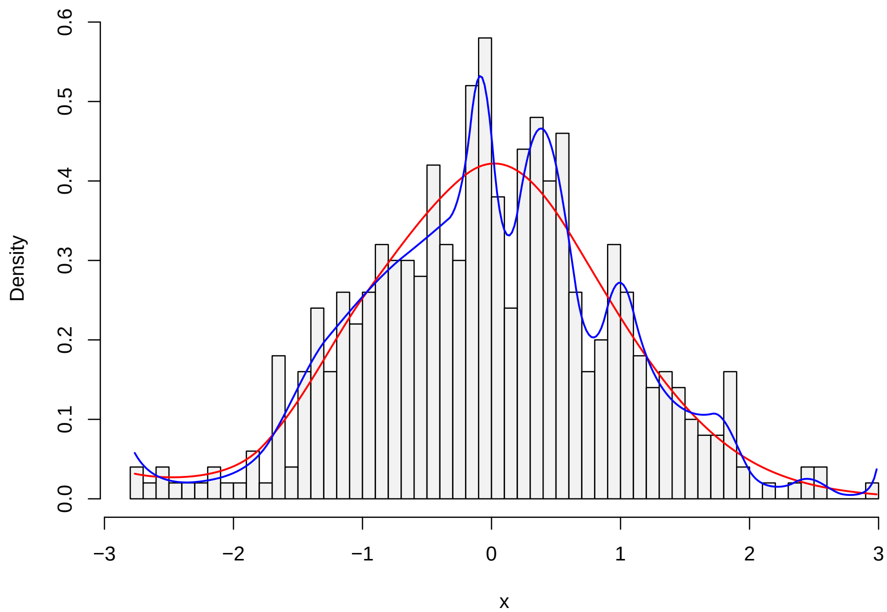

FIGURE 4. Gaussian histogram based on 500 observations and two penalized maximum likelihood estimates with total variation norm penalty and $\lambda \in \{0.5 \times 10^{-4}, 0.5 \times 10^{-6}\}$.

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

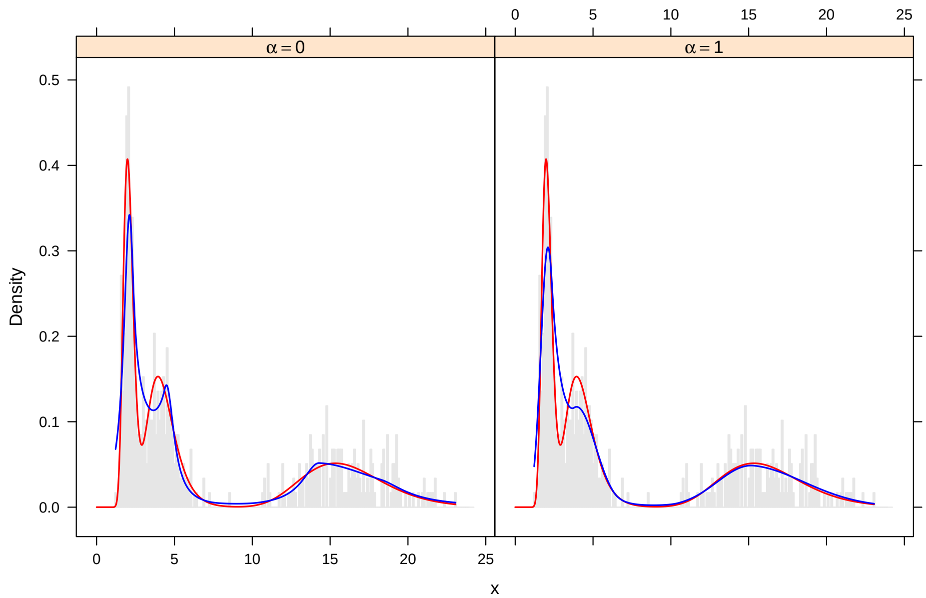

FIGURE 5. Mixture of three 3-parameter lognormals with histogram and two Rényi likelihood estimates with total variation (L_1 norm) penalty with $\alpha \in \{0, 1\}$ based on 500 iid observations and penalty parameter, $\lambda = 9$. The true density is depicted in red and the estimated density is in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

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. While 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 significantly expanded the theoretical underpinnings of this approach providing new existence, uniqueness and continuity results as well as extending its computational tractability a wide domain of α's. The approach has been illustrated with several examples taken from economics, astronomy and anthropometrics. We also briefly discussed related methods that pair norm-based smoothing penalties with the Rényi divergence estimation criterion.

Many problems remain for future research. As already mentioned, the convexity of the problems yielding our estimates entails not only favorable continuity properties, but also facilitates possible error bounds. Adaptive choice of α is undoubtedly an appealing question. However, we do not regard α as a typical tuning parameter, rather its selection is best dictated by close examination of its influence on particular features of the fitted density. This is revealed in our empirical examples: sharpness of the modal peak in the case of the rotational velocity application, and tail behavior in the Gosset anthropometry application. Global measures of fit, while certainly feasible criteria for guiding this choice, seem less well suited. The theoretical properties of the underlying estimators, despite the impressive accomplishments of Han and Wellner (2016), leave much still unknown especially about the limiting asymptotic behavior. The "netherworld" of $\alpha < 0$, in particular, remains to be charted. We look forward to future progress on these and other aspects of such methods.

APPENDIX A. PROOFS

Proof of Proposition 1. The proposition follows from the fact that if $G \in \mathcal{K}(X)^o$, then it annihilates all constant and linear functions – as these are precisely those g that both g and $-g$ are convex. In such a case

$$0 \geq \int g dG = -\int -g dG \geq 0, \quad \text{and thus} \quad \int g dG = 0.$$

Note that the constraint on f in (D) means that the integral with respect to $f dm$ is the same as that with respect to $d(\mathbb{Q}(X) - G)$. Thus, for every feasible f,

$$\int f dm = \int 1\, d(\mathbb{Q}(X) - G) = \int 1 d\mathbb{Q}(X) - \int 1 dG = \int 1 d\mathbb{Q}(X) = 1$$

and

$$\int x f dm = \int x\, d(\mathbb{Q}(X) - G) = \int x d\mathbb{Q}(X) - \int x dG = \int x d\mathbb{Q}(X).$$

□

Proof of Proposition 2. The proposition is the consequence of the duality Theorem 3.1 of Koenker and Mizera (2010) – formulated, however, not for a general dm, but the Lebesgue measure dx. The careful inspection of their proof reveals that dx is specifically involved in the invocation of Corollary 4A of Rockafellar (1970); the careful inspection of the latter reveals that it is in fact formulated for a general Borel regular measure dt – our dm.

Next paragraph of the proof of Koenker and Mizera (2010), devoted to the constraint qualification, makes a substantial use of the fact that the integral of a constant function over $\mathcal{H}(X)$, a bounded set, is finite. This follows from our standing assumption on dm: it assigns finite values to bounded sets.

Finally, the extremal condition follows from the form of the subgradient given by Corollary 4B of Rockafellar (1970) – which is again formulated for general Borel regular dm. □

Proof of Proposition 3. The proof follows from that of Theorem 4.1 of Koenker and Mizera (2010). In spite of the theorem imposing the assumption that ψ be bounded from below by 0, the proof briefly addresses in the last paragraph a potential treatment of ψ not necessarily bounded from below. In such cases, one needs to find an integrable minorant; this is possible here due to the fact that the support of the solutions is restricted to $\mathcal{H}(X)$ – and consequently dm only needs to be considered on that domain as well. The standing assumption that dm assigns a finite value to $\mathcal{H}(X)$ is first required for a constant function to be in the domain of the objective function; without loss of generality, this constant function can be equal to 1, that is, we can set y appearing in the proof of Koenker and Mizera to be equal to 1 – given that a is contained in $(0, +\infty)$ and thus, due to the assumptions of the proposition, is in the domain of ψ as well. The convexity of ψ then entails that the linear function supporting ψ at 1 lies entirely below the graph of ψ,

$$\psi(1 + z) \geq \psi'(1)z + \psi(1).$$

This inequality then yields for $\tau \geq 1$,

$$\frac{\psi(1 + \tau g_{(X,Z)}(x))}{\tau} \geq \psi'(1) g_{(X,Z)}(x) + c$$

where $c = \min\{\psi(1), 0\}$ and $g_{(X,Z)}$ is the function introduced in the proof of Koenker and Mizera. As Z is fixed in the proof, the right-hand side provides the desired minorant: the range of $g_{(X,Z)}(x)$ for $x \in \mathcal{H}(X)$ is bounded and the integrability then follows from the fact that the integration is with respect to a finite measure, dm restricted to $\mathcal{H}(X)$. The assumptions of the proposition regarding the limits of $\psi(y+\tau x)/\tau$ then conclude the proof of existence along the lines of the proof of Theorem 4.1 of Koenker and Mizera (2010); the proof of uniqueness goes exactly along the lines of the same proof – namely, its penultimate paragraph. □

Proof of Proposition 4. Given the form of the objective function in (P) and the fact that all solutions are convex functions supported by $\mathcal{H}(X)$, and thus continuous and bounded, we obtain that the objective functions of $(P)$for $dm = dm_\nu$ converges to the objective function of (P) for $dm = dm_0$, at every $g \in \mathcal{G}(X)$. In view of the finite-dimensional parametrization of $\mathcal{G}(X)$ by values $Y_i = g(X_i)$ this pointwise convergence can be strengthened to uniform convergence on compacts, due to convexity of the objective functions, as in Theorem 10.8 of Rockafellar (1970), see also Pollard (1991) or Hjort and Pollard (2011). This uniform convergence on the compact lower level sets of the objective function of (P) for $dm = dm_0$ containing, as revealed by the proof of Proposition 3, the solution of (P) for $dm = dm_0$ in its interior, entails the proposition. □

Proof of Proposition 5. See Lemma 2.2 of Seijo and Sen (2011). □

Proof of Proposition 6. Under the assumptions on ψ, its conjugate can be obtained as its Legendre transformation,

$$\psi^*(-f) = -\varphi(-f)f - \psi(\varphi(-f)) \tag{8}$$

which means that the integrand of

$$\tilde{\Phi}_0(f) = \int \varphi(-f)f_0 + \psi^*(-f_0) + \psi(\varphi(-f))\,dm \tag{9}$$

is identically equal to 0 for $f = f_0$. The nonnegativity of this integrand for all other f follows from the same inequality argument as in the proof of Theorem 4.2 of Koenker and Mizera (2010), and yields

$$0 = \tilde{\Phi}_0(f_0) \le \tilde{\Phi}_0(f), \tag{10}$$

possibly with the right-hand side equal to $+\infty$.

To obtain the analogous inequality for the objective function Φ_0, we need only to "subtract" the integral of $\psi^*(-f_0)$, heeding the subtleties that may arise when infinities are involved. For $f = f_0$, the equality (8) implies that we can legitimately write

$$0 = \tilde{\Phi}_0(f_0) = \int \psi^*(-f_0)\,dm + \int \varphi(-f_0)f_0 + \psi(\varphi(-f_0))\,dm, \tag{11}$$

as the existence of the finite integral of $\psi^*(f_0)$ on the left-hand side implies the same for the term on the right-hand side of (8) for $f = f_0$.

If $\tilde{\Phi}_0(f) < +\infty$, that is, if the integral of a nonnegative integrand in (9) exists and is finite, then the existence of the finite integral of $-\psi^*(f)$ implies the integrability of the sum, that is the existence of the finite integral in (5); then we can legitimately write

$$\tilde{\Phi}_0(f) = \int \psi^*(-f_0)\, dm + \int \varphi(-f) f_0 + \psi(\varphi(-f))\, dm \tag{12}$$

and combine (10), (11), and (12) to obtain the proposition.

Suppose that $\tilde{\Phi}_0(f) = +\infty$. Since both integrals in (11) are finite, the proposition will follow if we show that the rightmost integral in (12) is also equal to $+\infty$. Note that we cannot use (12) now (which would make the conclusion obvious), as we did not establish it in this case. We can, however, use the following: suppose that p is an integrable function, its integral exists and is finite, and $q \geq 0$ is a nonnegative function such that its integral is $+\infty$. Then the integral of $q - p$ exists and is equal to $+\infty$. To demonstrate this, we define, in a usual manner,

$$(q-p)^+ = \max\{(q-p), 0\}, \qquad (q-p)^- = \max\{-(q-p), 0\},$$

and

$$p^+ = \max\{p, 0\}, \qquad p^- = \max\{-p, 0\};$$

we know that $p = p^+ - p^-$, and that the integrals of both p^+ and p^- are finite. The finiteness of the integral of p^+ implies the same for the integral of

$$(q-p)^- = \max\{p-q, 0\} = \max\{p^+ - p^- - q, 0\} \leq \max\{p^+, 0\} = p^+,$$

due to the nonnegativity of p^- and q. Now, the integral of $(q-p)^+$ can be finite or $+\infty$. If it is finite, then we know that the integral of $(q-p)$ exists and is finite; this means that the integral of $q = (q-p) + p$ also exists and is finite, which contradicts our assumption about q; therefore, the integral of $(q-p)^+$, and consequently that of $q - p$ is equal to $+\infty$. □

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.

BASU, A., I. R. HARRIS, N. L. HJORT, AND M. C. JONES (1998): "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, 85, 549–559.

BILLINGSLEY, P. (1968): *Convergence of probability measures*. Wiley, New York.

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

BRONIATOWSKI, M., AND A. KEZIOU (2006): "Minimization of ϕ-divergences on sets of signed measures," *Studia Scientiarum Mathematicarum Hungarica*, 43(4), 403–442.

——— (2009): "Parametric estimation and tests through divergences and the duality technique," *Journal of Multivariate Analysis*, 100(1), 16–36.

BRONIATOWSKI, M., AND I. VAJDA (2012): "Several applications of divergence criteria in continuous families," *Kybernetika*, 48(4), 600–636.

BURG, J. (1967): "Maximum entropy spectral analysis," in *Proceedings of 37th Annual Meeting of the Society of Exploration Geophysicists*. SEG, Oklahoma City, OK.

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," *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`.

GHOSH, A. (2015): "Influence function analysis of the restricted minimum divergence estimators: A general form," *Electronic Journal of Statistics*, 9, 1017–1040.

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," *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.

Hjort, N. L., and D. Pollard (2011): "Asymptotics for minimisers of convex processes," *Preprint arXiv:1107.3806*.

Hoffleit, D., and W. H. Warren (1991): *The Bright Star Catalog (5th ed.)*. Yale University Observatory, New Haven.

Huber, P. J. (1967): "The Behavior of Maximum Likelihood Estimates Under Nonstandard Conditions," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 221–233.

Kim, A. K. H., A. Guntuboyina, and R. J. Samworth (2016): "Adaptation in log-concave density estimation," arXiv preprint available from `https://arxiv.org/abs/1609.00861`.

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

Koenker, R., and I. Mizera (2006): "The alter egos of the regularized maximum likelihood density estimators: deregularized maximum-entropy, Shannon, Renyi, Simpson, Gini, and stretched strings," in *Prague Stochastics 2006, Proceedings of the joint session of 7th Prague Symposium on Asymptotic Statistics and 15th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, held in Prague from August 21 to 25, 2006*, ed. by M. Hušková, and M. Janžura, pp. 145–157. Matfyzpress, Prague.

——— (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.

——— (2017): "MeddeR: Maximum Entropy Deregularized Density Estimation in R ," R package version 0.51, available from `http://www.econ.uiuc.edu/~roger/research/densiles/quasi.html`.

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

Laha, N., and J. Wellner (2017): "Bi-s^*-concave distributions," available from arXiv:1705.00252.

Liese, F., and I. Vajda (2006): "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, 52(10), 4394–4412.

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.

POLLARD, D. (1991): "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 7(2), 186–199.

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," Available from `https://www.R-project.org/`.

ROCKAFELLAR, R. T. (1970): *Convex analysis*. Princeton University Press, Princeton.

SEIJO, E., AND B. SEN (2011): "Nonparametric least squares estimation of a multivariate convex regression function," *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.

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

WALD, A. (1949): "Note on the consistency of the maximum likelihood estimate," *Annals of Mathematical Statistics*, 20(4), 595–601.

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.