The Annals of Statistics
2010, Vol. 38, No. 5, 2998–3027
DOI: 10.1214/10-AOS814

QUASI-CONCAVE DENSITY ESTIMATION

BY ROGER KOENKER[1] AND IVAN MIZERA[2]

University of Illinois and University of Alberta

Maximum likelihood estimation of a log-concave probability density is formulated as a convex optimization problem and shown to have an equivalent dual formulation as a constrained maximum Shannon entropy problem. Closely related maximum Renyi entropy estimators that impose weaker concavity restrictions on the fitted density are also considered, notably a minimum Hellinger discrepancy estimator that constrains the reciprocal of the square-root of the density to be concave. A limiting form of these estimators constrains solutions to the class of quasi-concave densities.

1. Introduction. Our objective is to introduce a general class of shape constraints applicable to the estimation of probability densities, multivariate as well as univariate. Elements of the class are represented by restricting certain monotone functions of the density to lie in convex cones. Maximum likelihood estimation of log-concave densities constitutes an important special case; however, the wider class allows us to include a variety of other shapes. A one parameter subclass modeled on the means of order ρ studied by Hardy, Littlewood and Pólya (1934) incorporates all the quasi-concave densities, that is, all densities with convex upper contour sets. Estimation methods for these densities, as described below, bring new opportunities for statistical data analysis.

Log-concave densities play a crucial role in a wide variety of probabilistic models: in reliability theory, search models, social choice and a broad range of other contexts it has proven convenient to assume densities whose logarithm is concave. Recognition of the importance of log-concavity was already apparent in the work of Schoenberg and Karlin on total positivity beginning in the late 1940s. Karlin (1968) forged a link between log-concavity and classical statistical properties such as the monotone likelihood ratio property, the theory of sufficient statistics and uniformly most powerful tests. Maximum likelihood estimation of densities constrained to be log-concave has recently enjoyed a considerable vogue with important contributions of Walther (2001, 2002, 2009), Pal, Woodroofe and Meyer (2007), Rufibach (2007), Dümbgen and Rufibach (2009), Balabdaoui, Rufibach

Received August 2009; revised December 2009.

[1]Supported in part by NSF Grant SES-08-50060.

[2]Supported by the NSERC of Canada.

AMS 2000 subject classifications. Primary 62G07, 62H12; secondary 62G05, 62B10, 90C25, 94A17.

Key words and phrases. Density estimation, unimodal, strongly unimodal, shape constraints, convex optimization, duality, entropy, semidefinite programming.

and Wellner (2009), Chang and Walther (2007) and Cule, Samworth and Stewart (2010), among others.

Log-concave densities are constrained to exhibit exponential tail behavior. This restriction motivates a search for weaker forms of the concavity constraint capable of admitting common densities with algebraic tails like the t and F families. The ρ-concave densities introduced in Section 2 constitute a rich source of candidates. While it would be possible, in principle, to consider maximum likelihood estimation of such densities, duality considerations lead us to consider a more general class of maximum entropy criteria. Maximizing Shannon entropy in the dual is equivalent to maximum likelihood for the leading log-concave case, but other entropies are also of interest. Section 3 describes several examples arising in the dual from the class of Rényi entropies, each corresponding to a distinct specification of the concavity constraint, and each corresponding to a distinct fidelity criterion in the primal. The crucial advantage of adapting the fidelity criterion to the form of the concavity constraint is that it assures a convex optimization problem with a tractable computational strategy.

2. Quasi-concave probability densities and their estimation. A probability density function, f, is called *log-concave* if $-\log f$ is a (proper) convex function on the support of f. We adhere to the usual conventions of Rockafellar (1970), which allow convex functions to take infinite values—although we will allow only $+\infty$, because all our convex functions will be *proper*. The *domain* of a convex (concave) function, $\operatorname{dom} g$, is then the set of x such that $g(x)$ is finite. We adopt the convention $-\log 0 = +\infty$.

Unimodality of concave functions implies that log-concave densities are unimodal. An interesting connection in the multivariate case was pointed out by Silverman (1981): the number of modes of a kernel density estimate is monotone in the bandwidth when the kernel is log-concave. However, as illustrated by the Student t family, not every unimodal density is log-concave. Laplace densities, with their exponential tail behavior, are; but heavier, algebraic tails are ruled out. This prohibition motivates a relaxation of the log-concavity requirement.

2.1. *A hierarchy of ρ-concave functions.* A natural hierarchy of concave functions can be built on the foundation of the weighted means of order ρ studied by Hardy, Littlewood and Pólya (1934): for any p in the unit simplex, $\mathcal{S} = \{p \in \mathbb{R}^n | p \geq 0, \sum p_i = 1\}$, let

$$M_\rho(a; p) = M_\rho(a_1, \dots, a_n; p) = \left(\sum_{i=1}^n p_i a_i^\rho\right)^{1/\rho},$$

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

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

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

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

In this terminology, log-concave functions are 0-concave and concave functions are 1-concave. As $M_\rho(a, p)$ is monotone, increasing in ρ for $a \geq 0$ and any $p \in \mathcal{S}$, it follows that if f is ρ-concave, then f is also ρ'-concave for any $\rho' < \rho$. 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 (and consequently for all ρ-concave functions) *quasi-concave*.

The hierarchy of ρ-concave density functions was considered in the economics literature by Caplin and Nalebuff (1991) in spatial models of voting and imperfect competition; their results reveal some intriguing connections to Tukey's half-space depth in multivariate statistics; see Mizera (2002). Curiously, it appears that the first thorough investigation of the mathematical concept of quasi-concavity was carried out by de Finetti (1949). Further details and motivation for ρ-concave densities can be found in Prékopa (1973), Borell (1975) and Dharmadhikari and Joag-Dev (1988).

2.2. *Maximum likelihood estimation of log-concave densities.* Suppose that $X = \{X_1, \dots, X_n\}$ is a collection of data points in $\mathbb{R}^d$ such that the convex hull of X, $\mathcal{H}(X)$, has a nonempty interior in $\mathbb{R}^d$; such a configuration occurs with probability 1 if $n \geq d$ and the X_i behave like a random sample from f_0, a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$. Viewing the X_i's as a random sample from an unknown, log-concave density f_0, we can find the *maximum likelihood estimate* of f_0 by solving

$$\prod_{i=1}^n f(X_i) = \max_f! \qquad \text{such that } f \text{ is a log-concave density.} \tag{2.1}$$

It is convenient to recast (2.1) in terms of $g = -\log f$, the estimate becoming $f = e^{-g}$,

$$\sum_{i=1}^n g(X_i) = \min_g! \qquad \text{such that } g \text{ is convex} \quad \text{and} \quad \int e^{-g(x)}\,dx = 1. \tag{2.2}$$

The objective function of (2.2) is equal to $+\infty$, given the convention adopted above, unless all X_i are in the domain of g. As in Silverman (1982), it proves convenient to move the integral constraint into the objective function,

$$\frac{1}{n}\sum_{i=1}^n g(X_i) + \int e^{-g(x)}\,dx = \min_g! \qquad \text{such that } g \text{ is convex,} \tag{2.3}$$

a device that ensures that the solution integrates to one without enforcing this condition explicitly. Apart from the multiplier $1/n$, the crucial difference between (2.2) and (2.3) is that the latter is a convex problem, while the former not.

It is well known that naïve, unrestricted maximum likelihood estimation is doomed to fail when applied in the general density estimation context: once "log-concave" is dropped from the formulation of (2.1), any sequence of putative maximizers is attracted to the the linear combination of point masses situated at the data points. One escape from this "Dirac catastrophe" involves regularization by introducing a roughness, or complexity, penalty; various proposals in this vein can be found in Good (1971), Silverman (1982), Gu (2002) and Koenker and Mizera (2008).

Another way to obtain a well-posed problem is by imposing shape constraints, a line of development dating back to the celebrated Grenander (1956) nonparametric maximum likelihood estimator for monotone densities. While monotonicity regularizes the maximum likelihood estimator, unimodality *per se*—somewhat surprisingly—does not. The desired effect is achieved only by enforcing somewhat more stringent shape constraints—for instance log-concavity, sometimes also called "strong unimodality." An advantage of shape constraints over regularization based on norm penalties is that it is not encumbered by the need to select additional tuning parameters; on the other hand, it is limited in scope—applicable only when the shape constraint is plausible for the unknown density.

2.3. *Quasi-concave density estimation.* Expanding the scope of our investigation, we now replace e^{-g} in the integral of the objective function by a generic function $\psi(g)$ and define

$$\Phi(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g(x))\,dx. \tag{2.4}$$

The following conditions on the form of ψ will be imposed:

(A1) ψ is a nonincreasing, proper convex function on $\mathbb{R}$.
(A2) The domain of ψ is an open interval containing $(0, +\infty)$.
(A3) The limit, as $\tau \to +\infty$, of $\psi(y + \tau x)/\tau$ is $+\infty$ for every real y and any $x < 0$.
(A4) ψ is differentiable on the interior of its domain.
(A5) ψ is bounded from below by 0, with $\psi(x) \to 0$ when $x \to +\infty$.

The most crucial condition is (A1) ensuring the convexity of Φ. Condition (A2) assures that $\psi(x)$ is finite for all $x > 0$, while (A3) is required in the proof of the existence of the estimates. The relationship between primal and dual formulations of the estimation problem is facilitated by (A4), and (A5) rules out possible complications regarding the existence of the integral $\int \psi(g)\,dx$ in (2.4), allowing for the convention $\psi(+\infty) = 0$. In the spirit of the Lebesgue integration theory, the integral then *exists*, although $\psi(g)$ does not have to be *summable*: it is either finite [which is automatically true for any g convex and $\psi(g) = e^{-g}$] or $+\infty$. In the latter case, the objective function $\Phi(g)$ is considered to be equal to $+\infty$; $\Phi(g)$ is also $+\infty$ if $g(X_i) = +\infty$ for some X_i, which occurs unless all X_i lie in the domain of g. On the other hand, any g equal to some positive constant on $\mathcal{H}(X)$ and $+\infty$ elsewhere yields $\Phi(g) < \infty$.

A rigorous treatment without assumption (A5), that is, for functions ψ not bounded below, would introduce technicalities involving handling of the integrals in the spirit of singular integrals of calculus of variations, a strategy resembling the contrivance of Huber (1967) of subtracting a fixed quantity from the objective function to ensure finiteness of the integral. Although we do not believe that such formal complications are unsurmountable, we do not pursue such a development.

Careful deliberation reveals that replacing g by its closure (lower semicontinuous hull) does not change the integral term in (2.4), and potentially only decreases the first term; this means that without any restriction of its scope, we may reformulate the estimation problem as

$$\text{(2.5)} \qquad \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g(x))\,dx = \min_{g}! \qquad \text{subject to } g \in \mathcal{K},$$

where $\mathcal{K}$ stands for the class of closed (lower semicontinuous) convex functions on $\mathbb{R}^d$.

Unlike in (2.3), $\psi(g)$ is not necessarily the estimated density f; the relationship of g to f will be revealed in Section 3, together with the motivation leading to concrete instances of some possible functions ψ.

2.4. *Characterization of estimates.* We now establish that the estimates, the solutions of (2.5), admit a finite-dimensional characterization, which is a key to many of their theoretical properties. For every collection (X, Y) of points $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$, we define a function

$$\text{(2.6)} \qquad g_{(X,Y)}(x) = \inf\left\{\sum_{i=1}^{n} \lambda_i Y_i \;\middle|\; x = \sum_{i=1}^{n} \lambda_i X_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0\right\}.$$

Any function of this type is finitely generated in the sense of Rockafellar (1970), whose Corollary 19.1.2 asserts that it is polyhedral, being the maximum of finitely many affine functions, and therefore convex. The convention $\inf \varnothing = +\infty$ used in (2.6) means that the domain of $g_{(X,Y)}$ is equal to $\mathcal{H}(X)$. If h is a convex function such that $h(X_i) \leq Y_i$, for all i, then $h(x) \leq g_{(X,Y)}(x)$ for all x; the function $g_{(X,Y)}$ is thus the maximum of convex functions with this property—the lower convex hull of points (X_i, Y_i).

For fixed X, we will denote the collection of all functions $g_{(X,Y)}$ of the form (2.6) by $\mathcal{G}(X)$. The collection (X, Y) determines $g_{(X,Y)}$ uniquely, by virtue of its definition (2.6). Given X, we call a vector Y with components $Y_i \in \mathbb{R}$ *discretely* *convex relative to* X, if there exists a convex function h defined on $\mathcal{H}(X)$ such that $h(X_i) = Y_i$. Any function g from $\mathcal{G}(X)$ determines a unique discretely convex vector $Y_i = g(X_i)$. The converse is also true: there is a one–one correspondence between $\mathcal{G}(X)$ and $\mathcal{D}(X) \subseteq \mathbb{R}^n$, the set of all vectors discretely convex relative to X.

THEOREM 2.1. *Suppose that assumption* (A1) *holds true. For every convex function h on $\mathbb{R}^d$, there is a function $g \in \mathcal{G}(X)$ such that $\Phi(g) \leq \Phi(h)$; the strict inequality holds whenever $h \notin \mathcal{G}(X)$ and $\mathcal{H}(X)$ has nonempty interior.*

The theorem shows that it is sufficient to seek potential solutions of (2.5) in $\mathcal{G}(X)$; this means, due to the one–one correspondence of the latter to $\mathcal{D}(X)$, that the optimization task (2.5) is essentially finite dimensional. The theorem also justifies the transition to a more convenient optimization domain in the primal formulation appearing in the next section.

3. Duality, entropy and divergences. The conjugate dual formulation of the primal estimation problem (2.5) conveys a maximum entropy interpretation and leads us to several concrete proposals for ψ. To conform to existing mathematical apparatus, we begin by further clarifying the optimization and constraint functional classes of our primal formulation. For definitions and general background on convex analysis, our primary references are Rockafellar (1970, 1974) and Zeidler (1985); we may also mention Hiriart-Urruty and Lemaréchal (1993) and Borwein and Lewis (2006).

3.1. *The primal formulation.* Hereafter, $\mathcal{K}(X)$ will denote the cone of closed (lower semicontinuous) convex functions on $\mathcal{H}(X)$, the convex hull of X. This cone is a subset of $\mathcal{C}(X)$, the collection of functions continuous on $\mathcal{H}(X)$; it is important that $\mathcal{C}(X)$ is a linear topological space, with respect to the topology of uniform convergence. Note that $\mathcal{G}(X) \subset \mathcal{K}(X) \subset \mathcal{C}(X)$. In view of Theorem 2.1, any solution of (2.5) is also the solution of

$$(3.1) \qquad \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g(x))\,dx = \min_{g \in \mathcal{C}(X)}! \qquad \text{subject to } g \in \mathcal{K}(X),$$

and conversely; thus, we will refer to (3.1) as our *primal formulation*.

3.2. *The dual formulation.* The conjugate of ψ is

$$\psi^*(y) = \sup_{x \in \operatorname{dom}\psi} \bigl(yx - \psi(x)\bigr).$$

Since ψ is nonincreasing, there are no affine functions with positive slope that minorize the graph of ψ, hence $\psi^*(y) = +\infty$ for all $y > 0$. If ψ is differentiable on the (nonempty) interior of its domain, then ψ^* can be obtained using differential calculus—as the *Legendre transformation* of ψ; denoting the derivative ψ' by χ, we have

$$\psi^*(y) = y\chi^{-1}(y) - \psi(\chi^{-1}(y)), \tag{3.2}$$

where $\chi^{-1}(y)$ is any solution, z, of the equation $\chi(z) = y$. The (topological) dual of $\mathcal{C}(X)$ is $\mathcal{C}^*(X)$, the space of (signed) Radon measures on $\mathcal{H}(X)$; its distinguished element is P_n, the empirical measure supported by the data points X_i. The *polar cone* to $\mathcal{K}(X)$ is

$$\mathcal{K}(X)^- = \left\{ G \in \mathcal{C}^*(X) \,\middle|\, \int g\, dG \le 0 \text{ for all } g \in \mathcal{K}(X) \right\}.$$

THEOREM 3.1. *Suppose that assumptions* (A1) *and* (A2) *hold. The strong* (*Fenchel*) *dual of the primal formulation* (3.1) *is*

$$\begin{aligned} -\int \psi^*(-f(y))\, dy = \max_f! \qquad \textit{subject to } f = \frac{d(P_n - G)}{dy}&, \\ G \in \mathcal{K}(X)^-&, \end{aligned} \tag{3.3}$$

in the sense that the value, $\Phi(g)$, *of the primal objective for any* g *satisfying the constraints of* (3.1), *dominates the value*, *for any* f *satisfying the constraints of* (3.3), *of the objective function in* (3.3); *the minimal value of* (3.1) *and maximal value of* (3.3) *coincide*. *Moreover*, *there exists* f *attaining the maximal value of* (3.3). *Any dual feasible function* f, *that is*, *any* f *satisfying the constraints of* (3.3) *and yielding finite objective function of* (3.3), *is a probability density with respect to the Lebesgue measure*: $f \ge 0$ *and* $\int f\, dx = 1$. *If condition* (A4) *is also satisfied*, *then the dual and primal optimal solutions satisfy the relationship* $f = -\psi'(g)$.

It should be emphasized that the expression of absolute continuity in (3.3) is a requirement on $F = P_n - G$; the dual objective function is defined as the conjugate to the primal objective function Φ, and is equal to $-\infty$ for any Radon measure that is *not* absolutely continuous with respect to the Lebesgue measure. This is how regularization operates here: only those F qualify for which P_n gets canceled with the discrete component of G. Once F satisfies this requirement, its density integrates to 1, as shown in the proof of Theorem 3.1. The nonnegativity for f yielding finite dual objective function is the consequence of $\psi^*(-y)$ being infinite for $y < 0$. In practical implementations, it may be prudent to enforce $f \ge 0$ in the dual explicitly as a feasibility constraint.

3.3. *The interpretation of the dual.* An immediate consequence of Theorem 3.1 is that we can reformulate the maximum likelihood problem posed in (2.3) as an equivalent maximum (Shannon) entropy problem.

COROLLARY 3.1. *Maximum likelihood estimation of a log-concave density as posed in* (2.3) *has an equivalent dual formulation*

$$-\int f(y)\log f(y)\,dy = \max_f! \qquad \textit{subject to } f = \frac{d(P_n - G)}{dy}, \tag{3.4}$$
$$G \in \mathcal{K}(X)^-,$$

whose solution satisfies the relationship $f = e^{-g}$, *where* g *is the solution of* (2.3). *In particular, the solution of* (2.3) *satisfies* $\int e^{-g(x)}\,dx = 1$, *therefore problems* (2.2) *and* (2.3) *are equivalent.*

The emergence of the Shannon entropy is hardly surprising—in view of the well-established connections of maximum likelihood estimation to the Kullback–Leibler divergence and maximum entropy. Note that the dual criterion can be also interpreted as choosing the f closest in the Kullback–Leibler divergence to the uniform distribution on $\mathcal{H}(X)$, from all f satisfying the dual constraints.

3.4. *Rényi entropies.* While the outcome of Corollary 3.1, the equivalence of (2.2) and (2.3), could be also shown by elementary means, it is important to emphasize that the real value of the dual connection lies in the vista of new possibilities it opens. To explore the link to potential alternatives, we consider the family of entropies originally introduced for $\alpha > 0$ by Rényi (1961, 1965),

$$(1-\alpha)^{-1}\log\left(\int f^\alpha(x)\,dx\right), \qquad \alpha \neq 1, \tag{3.5}$$

as an extension of the limiting case for $\alpha = 1$, the Shannon entropy. For $\alpha \neq 1$, maximizing (3.5) over f is equivalent to the maximization of

$$\frac{\operatorname{sgn}(1-\alpha)}{\alpha}\int f^\alpha(x)\,dx = -\operatorname{sgn}(\alpha-1)\int \frac{f^\alpha(x)}{\alpha}\,dx. \tag{3.6}$$

The dependence of convexity/concavity properties of y^α necessitates a separate treatment of the cases with $\alpha > 1$, when the conjugate pair is

$$\psi(x) = \begin{cases} (-x)^\beta/\beta, & \text{for } x \le 0,\\ 0, & \text{for } x > 0,\end{cases} \qquad \psi^*(y) = \begin{cases} (-y)^\alpha/\alpha, & \text{for } y \le 0,\\ +\infty, & \text{for } y > 0,\end{cases}$$

and the cases with $\alpha < 1$, where

$$\psi(x) = \begin{cases} +\infty, & \text{for } x \le 0,\\ -x^\beta/\beta, & \text{for } x > 0,\end{cases} \qquad \psi^*(y) = \begin{cases} -(-y)^\alpha/\alpha, & \text{for } y \le 0,\\ +\infty, & \text{for } y > 0,\end{cases}$$

where β and α are conjugates in the usual sense that $1/\beta + 1/\alpha = 1$. See Figure 1.

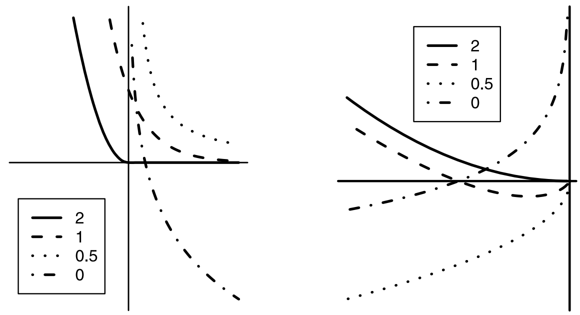


FIG. 1. *Primal ψ (left) and dual ψ^* (right) for selected $\alpha \geq 0$ from the Rényi family of entropies.*

The general form of the primal formulations (3.1) corresponding to (3.5) can be written, for $\alpha \neq 1$, in a unified way as

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{1}{|\beta|}\int |g(x)|^{\beta}\,dx = \min_{g\in\mathcal{C}(X)}! \tag{3.7}$$

together with the relation between the dual and primal solutions, $f = |g|^{\beta-1}$. Several particular instances merit special attention.

3.5. *Power divergences.* For $\alpha > 1$, we may write $(-g)$ instead of $|g|$, and then introduce $h = -g$. The resulting primal formulation is

$$-\frac{1}{n}\sum_{i=1}^{n} h(X_i) + \frac{1}{\beta}\int h^{\beta}(x)\,dx = \min_{h\in\mathcal{C}(X)}! \qquad \text{subject to } h \in \mathcal{K}(X). \tag{3.8}$$

By Theorem 2.1, this formulation is equivalent to

$$-\frac{1}{n}\sum_{i=1}^{n} h(X_i) + \frac{1}{\beta}\int h^{\beta}(x)\,dx = \min_{h}! \qquad \text{subject to } h \in \mathcal{K}. \tag{3.9}$$

After substituting $f^{1/(\beta-1)}$ for h, multiplying by β, and rewriting in terms of α we obtain a new objective function

$$-\left(\frac{\alpha}{\alpha-1}\right)\frac{1}{n}\sum_{i=1}^{n} f^{\alpha-1}(X_i) + \int f^{\alpha}(x)\,dx, \tag{3.10}$$

which recalls the "minimum density power divergence estimators," proposed, for $\alpha \geq 1$, by Basu et al. (1998) in the context of estimation in parametric families.

3.6. *Pearson χ^2.* Although $\alpha = 2$ is a special case of the power divergence family mentioned above, it deserves a special mention. The choice of $\alpha = 2$ in the Rényi family leads to the dual formulation

$$-\int f^2(y)\,dy = \max_{f}! \qquad \text{subject to } f = \frac{d(P_n - G)}{dy}, \qquad G \in \mathcal{K}(X)^-. \tag{3.11}$$

The primal formulation can be written, after the application of Theorem 2.1, in a particularly simple form

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{1}{2}\int g^2(x)\,dx = \min_g! \qquad \text{subject to } g \in \mathcal{K}, \tag{3.12}$$

which can be interpreted as a variant of the minimum Pearson χ^2 criterion. A similar theme can be found in the dual, which can be interpreted as returning among all densities satisfying its constraints the one with minimal Pearson χ^2 distance to the uniform density on $\mathcal{C}(X)$.

The relation between primal and dual optimal solutions is $f = -g$; the convexity constraint on g therefore implies that f must be concave. Replacing g in (3.12) by $-f$ and appropriately modifying the cone constraint gives a variant of the "least-squares estimator," studied by Groeneboom, Jongbloed and Wellner (2001) and going back at least to Birgé and Massart (1993); the estimate was defined to estimate a *convex* (and decreasing) density on $\mathbb{R}^+$, a domain that is apparently still under the scope of Theorem 2.1.

3.7. *Hellinger.* While the form of the objective function for $\alpha = 2$ has some computational advantages, its secondary consequence—constraining the density itself to be concave rather than its logarithm—is not at all appealing. Indeed, all Rényi choices with $\alpha > 1$ impose a more restrictive form of concavity than log-concavity. From our perspective, it seems more reasonable to focus attention on weaker forms of concavity, corresponding to $\alpha \le 1$. Apart from the celebrated log-concave case $\alpha = 1$, a promising alternative would seem to be Rényi entropy with $\alpha = 1/2$. This choice in the Rényi system leads to the dual

$$\int \sqrt{f(y)}\,dy = \max_f! \qquad \text{subject to } f = \frac{d(P_n - G)}{dy}, \qquad G \in \mathcal{K}(X)^-, \tag{3.13}$$

and primal, again after the application of Theorem 2.1,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \frac{1}{g(x)}\,dx = \min_g! \qquad \text{subject to } g \in \mathcal{K}. \tag{3.14}$$

The estimated density satisfies $f = 1/g^2$, which means that the primal constraint, $g \in \mathcal{K}$, enforces the convexity of $g = 1/\sqrt{f}$. In the terminology of Section 2, the estimated density is now required to be only $-1/2$-concave, a significant relaxation of the log-concavity constraint; in addition to all log-concave densities, all the Student t_ν densities with $\nu \ge 1$ satisfy this requirement. The dual problem (3.13) can be interpreted as a Hellinger fidelity criterion, selecting from the cone of dual feasible densities the one closest in Hellinger distance to the uniform distribution on $\mathcal{H}(X)$.

3.8. *The frontier and beyond*? Although the original Rényi system was confined to $\alpha > 0$, a limiting form for $\alpha = 0$ can be obtained similarly to the $\alpha = 1$ case. It yields the conjugate pair

$$\psi(x) = \begin{cases} +\infty, & \text{for } x \le 0, \\ -1/2 - \log x, & \text{for } x > 0, \end{cases}$$

$$\psi^*(y) = \begin{cases} -1/2 - \log(-y), & \text{for } y < 0, \\ +\infty, & \text{for } y \ge 0. \end{cases}$$

As is apparent from Figure 1, this ψ violates our condition (A5), but may nevertheless deserve a brief consideration. Note first that the possible complications with existence of integrals may occur only in the formulation (2.5) with unbounded domain—not in (3.1), where all integrals are of bounded functions over a compact domain. The major technical complications with ψ violating (A5) concern theorems in Section 4, and are briefly discussed there. Here we mention only that the resulting dual, adapted directly from (3.3), is

$$\int \log f(y)\, dy = \max_f ! \qquad \text{subject to } f = \frac{d(P_n - G)}{dy}, \qquad G \in \mathcal{K}(X)^-,$$

and the primal becomes

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \int \log g(x)\, dx = \min_{g \in \mathcal{C}(X)} ! \qquad \text{subject to } g \in \mathcal{K}(X).$$

In this case $g = 1/f$, and the estimate is constrained to be -1-concave, a yet still weaker requirement that admits all of the Student t_ν densities for $\nu > 0$.

If we interpret the dual problem (3.4), for $\alpha = 1$, as choosing a constrained f to minimize the Kullback–Leibler divergence of f from the uniform distribution on $\mathcal{H}(X)$, we can similarly interpret the $\alpha = 0$ dual as minimizing the reversed Kullback–Leibler divergence. In parametric estimation, the latter objective is sometimes associated with empirical likelihood, while the former is associated with exponentially tilted empirical likelihood. See, for example, Hall and Presnell (1999) for related discussion in the context of kernel density estimation, and Schennach (2007).

One might try to continue in this fashion marching inexorably toward weaker and weaker concavity requirements. There appears to be no obstacle in considering $\alpha < 0$; the general form (3.7) of the primal is still applicable. The shape constraints corresponding to negative α encompass a wider and wider class of quasi-concave densities, eventually arriving at the $-\infty$-concave constraint, at which point we would have sanctioned all of the quasi-concave densities. But formal complications, as well as computational difficulties dictate the more prudent strategy of restricting attention to $\alpha > 0$ cases.

4. Existence and Fisher consistency of estimates. Returning to our general setting, existence, uniqueness and Fisher consistency are established under mild conditions on the function ψ.

4.1. *Existence of estimates.* Theorem 2.1 not only justifies the choice of the optimization domain in (3.1), but also shows, due to the one–one correspondence between $\mathcal{G}(X)$ and $\mathcal{D}(X)$, that the optimization task (3.1) is essentially finite dimensional, parametrized by the values $Y_i = g(X_i)$. This facilitates the proof of the following existence result.

THEOREM 4.1. *Suppose that assumptions* (A1), (A2), (A3) *and* (A5) *hold, and that $\mathcal{H}(X)$ has a nonempty interior. Then the formulation* (2.5) *has a solution $g \in \mathcal{C}(X)$; if ψ is strictly convex, then this solution is unique.*

4.2. *Fisher consistency.* In our general setting, a comprehensive asymptotic theory for the proposed estimators remains a formidable objective. Considerable recent progress has been made on theory for the univariate log-concave ($\alpha = 1$) maximum likelihood estimator: Pal, Woodroofe and Meyer (2007) proved consistency in the Hellinger metric, Dümbgen and Rufibach (2009) prove consistency in the supremum norm on compact intervals, and Balabdaoui, Rufibach and Wellner (2009) derive asymptotic distributions. For maximum likelihood estimators in $\mathbb{R}^d$, Cule and Samworth (2010) establish consistency for estimators of a log-concave density, and Seregin and Wellner (2009) for estimators of convex-transformed densities. These results are surely suggestive of the plausibility of analogous results for other α and dimensions greater than one. However, the highly technical nature of the proofs, and their strong reliance on special features of the univariate setting indicate that such a development may not be immediate.

While anything else in this direction may be viewed as speculative, Fisher consistency, a crucial prerequisite for a more detailed asymptotic theory, can be verified in a quite straightforward manner and essentially complete generality. For differentiable ψ, Theorem 3.1 gives the relationship between the solution g of the optimization task (3.1) and the density estimate: $f = -\psi'(g)$. Using the notation χ for ψ', and χ^{-1} for its inverse, as in Section 3, we may write $g = \chi^{-1}(-f)$, and subsequently rewrite the formulation (2.5) in terms of the estimated density f (omitting, for brevity, the integration variables)

$$\int \chi^{-1}(-f)\, dP_n + \int \psi(\chi^{-1}(-f))\, dx = \min_f ! \quad \text{subject to } \chi^{-1}(-f) \in \mathcal{K}. \tag{4.1}$$

This yields a new objective function—which we nevertheless denote, slightly abusing the notation, also Φ. The population version of this Φ is obtained by replacing dP_n by $f_0\, dx$:

$$\Phi_0(f) = \int \chi^{-1}(-f) f_0 + \psi(\chi^{-1}(-f))\, dx. \tag{4.2}$$

The Fisher consistency for an estimator defined by solving (4.1) requires that $\Phi_0(f_0) \le \Phi_0(f)$, for every f; however, there may be a formal problem now with the existence of the integral in (4.2), as $\chi^{-1}(f)$ may take both positive or negative values. A possible way of handling this obstacle is the strategy of Huber (1967), briefly mentioned in Section 2: instead of Φ, we consider a modified objective function

$$\tilde{\Phi}(f) = \int \left(\chi^{-1}(-f) + \frac{\psi^*(-f_0)}{f_0} \right) dP_n + \int \psi(\chi^{-1}(-f))\,dx, \tag{4.3}$$

which, when minimized over f satisfying the constraint of (4.1), yields an optimization problem equivalent with (4.1), since the difference of Φ and $\tilde{\Phi}$ is constant in f. However, the population version of $\tilde{\Phi}$

$$\tilde{\Phi}_0(f) = \int \chi^{-1}(-f) f_0 + \psi^*(-f_0) + \psi(\chi^{-1}(-f))\,dx \tag{4.4}$$

is now better suited for the ensuing version of the Fisher consistency theorem.

THEOREM 4.2. *Suppose that ψ satisfies assumptions* (A1), (A2), (A4) *and* (A5). *The integrand in* (4.4) *is then nonnegative for any probability density f such that $\chi^{-1}(-f) \in \mathcal{K}$, and identically equal to* 0 *for $f = f_0$; therefore, $0 = \tilde{\Phi}_0(f_0) \le \tilde{\Phi}_0(f)$, where $\tilde{\Phi}_0(f)$ is well defined for every f, possibly equal to $+\infty$.*

In fact, Theorem 4.2 can be proved in the same manner for the unmodified Φ, if $\operatorname{dom}\psi = (\omega, +\infty)$ with $\omega > -\infty$. Then the inverse of $\chi = \psi'$, and hence the range of χ^{-1} is bounded from below by ω. In such a case, $\chi(f) f_0 \ge \omega f_0$, so the first term in (4.2) is minorized by an integrable function ωf_0; the second term is bounded from below by 0 by assumption (A5), so the whole integral then exists in the Lebesgue sense, being either finite or equal to $+\infty$.

If, however, assumption (A5) is not satisfied, then the existence of the integral should be assumed explicitly; we return to this point briefly at the end of the proof of Theorem 4.2. Note that, by comparing (3.2) and (4.2), existence of the integral is equivalent to assuming the integrability (summability) of

$$f_0 \chi^{-1}(f_0) + \psi(\chi^{-1}(-f_0)) = -\psi^*(-f_0), \tag{4.5}$$

that is, the existence and finiteness of the entropy term in the dual (3.3).

5. Examples of practical use. We employed two independent algorithms for solving the convex programming problems posed above: `mskscopt` from the MOSEK software package of Andersen (2006), and the `PDCO` MATLAB procedure of Saunders (2003). Both algorithms are coded in MATLAB and employ similar primal-dual, log-barrier methods. Further details regarding numerical implementation appear in Appendix B. The crux of both algorithms is a sequence of Newton-type steps that involve solving large, very sparse least squares problems, a task that is very efficiently carried out by modern variants of Cholesky decomposition. Several other approaches have been explored for computing quasi-concave density estimators that are log-concave. An active set algorithm for univariate log-concave density estimation was described in Dümbgen, Hüsler and Rufibach (2007) and implemented in the R package `logcondens` of Rufibach and Dümbgen (2009). Cule, Samworth and Stewart (2010) have recently implemented a promising steepest descent algorithm for multivariate log-concave estimation that may be adaptable to other quasi-concave density estimation problems.

5.1. *Univariate example*: *Velocities of bright stars*. To illustrate the application of the foregoing methods, we briefly consider some realistic examples. Our first example features data similar to those considered by Pal, Woodroofe and Meyer (2007), the type of data where shape constraints sometimes arise in a natural manner. The two samples consist of 9092 measurements of radial and 3933 of rotational velocity for the stars from Bright Star Catalog, Hoffleit and Warren (1991). The left and right panels of Figure 2 show the results for the radial and rotational velocity samples, respectively.

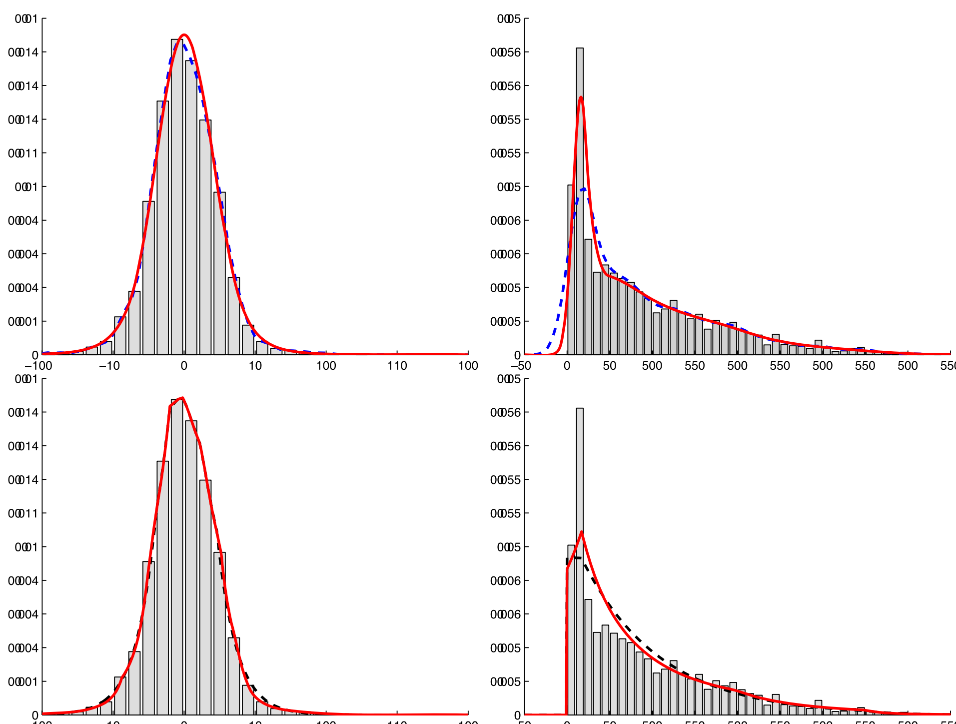


FIG. 2. *The estimates of the densities of radial* (*left*) *and rotational* (*right*) *velocities of the stars from the Bright Star Catalog. Broken lines are kernel density estimates in the upper two panels, and the solid lines are total variation penalized estimates. In the lower two panels the broken lines are the log-concave estimates and the solid lines represent the Hellinger* ($-1/2$-*concave*) *estimates.*

The broken line in the upper panels shows kernel density estimates, each time with default MATLAB bandwith selection; the solid lines correspond to one of the norm penalized estimates proposed in Koenker and Mizera (2008): maximum likelihood penalized by the total variation of the second derivative of the logarithm of the estimated density. This is the L^1 version of the Silverman's (1982) estimator penalizing the squared L^2 norm of the third derivative. The smoothing parameter for the latter estimate was set quite arbitrarily at 1; it seems that this arbitrary choice works quite satisfactorily here, providing—somewhat surprisingly, for both samples—about the same level of smoothing as the kernel estimator. For the radial velocity sample, the two estimates are essentially the same. For the rotational velocity sample, however, the right upper panel shows that the kernel density estimate differs substantially from the penalized one. Both estimators have the unfortunate effect of assigning considerable mass to negative values despite the fact that there are no negative observations. This effect is somewhat more pronounced for the kernel estimate.

Since the preliminary analyses of the upper panels indicates that the hypothesis of unimodality is plausible for both of the datasets, a natural next step is the application of a shape-constrained estimator—a move that, among other things, may relieve us of insecurities related to the arbitrary choice of smoothing parameters. The broken line in the lower panels of Figure 2 shows the log-concave maximum likelihood ($\alpha = 1$), and the solid line the Hellinger ($-1/2$-concave) estimate ($\alpha = 1/2$). While, as expected, there is almost no difference between the two (and in fact, among all four) estimates for the radial velocity dataset, the right lower panel reveals that the log-concave estimate yields for the rotational velocity sample a density that is monotonically decreasing—which contradicts the evidence suggested by all other methods. The Hellinger estimate, on the other hand, exhibits a subtle, but visible bump at the location of the plausible mode, thus turning out to be visually more informative about the center of the data than the tails. This is somewhat paradoxical given its original heavy-tail motivation, confirming that the real universe of data analysis can be much more subtle than that of the surrounding theoretical constructs.

5.2. *Bivariate example*: *Criminal fingers*. To illustrate our approach in a simple bivariate setting, we reconsidered the well-known MacDonell (1902) data on the heights and left middle finger lengths of 3000 British criminals. This data was employed by Gosset in preliminary simulation work described in Student (1908).

Figure 3 illustrates the Hellinger ($-1/2$-concave) fit of this data. Contours are labeled in units of log-density. A notable feature of the data is the unusual observation in the middle of the upper edge. This point is highly anomalous, at least for any density with exponential tail behavior. The maximum likelihood estimate of the log-concave model in Figure 4 has very similar central contours, but the outer contours fall off much more rapidly implying that the log-concave estimate assigns much smaller probability to the region near the unusual point.

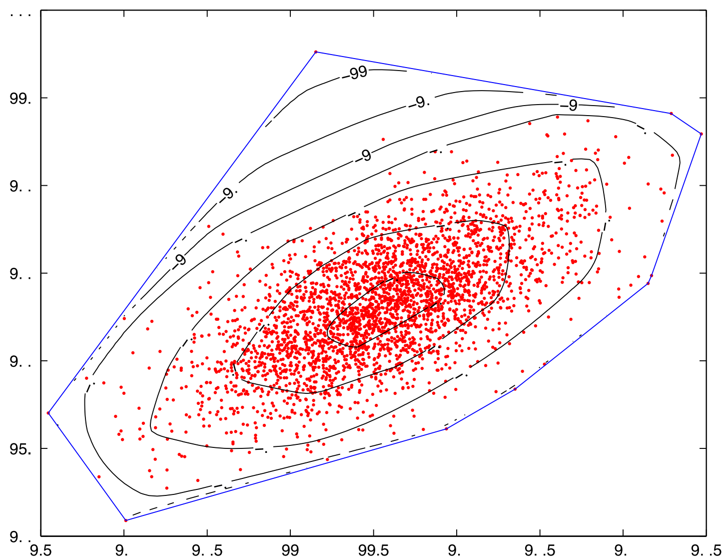


FIG. 3. *Hellinger* ($-1/2$-*concave*) *estimate of the density of Student's criminals. Contours are labeled in units of log-density.*

5.3. *Some simulation evidence.* Motivated by a suggestion of one of the referees, we undertook some numerical experiments to explore performance of our

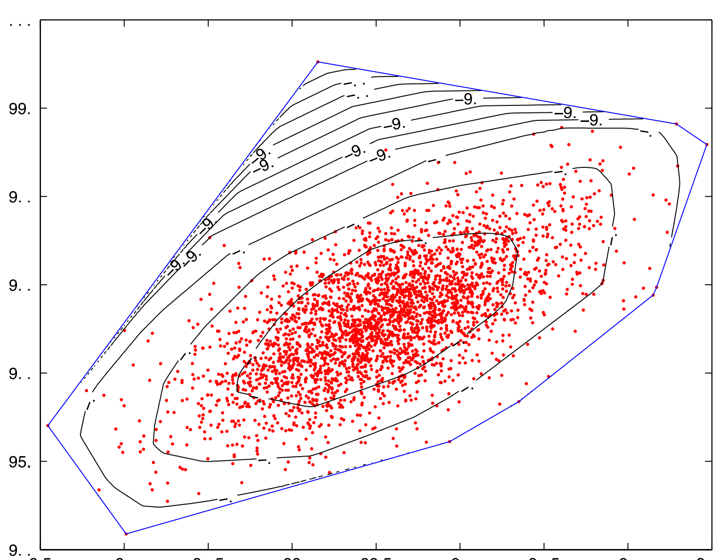
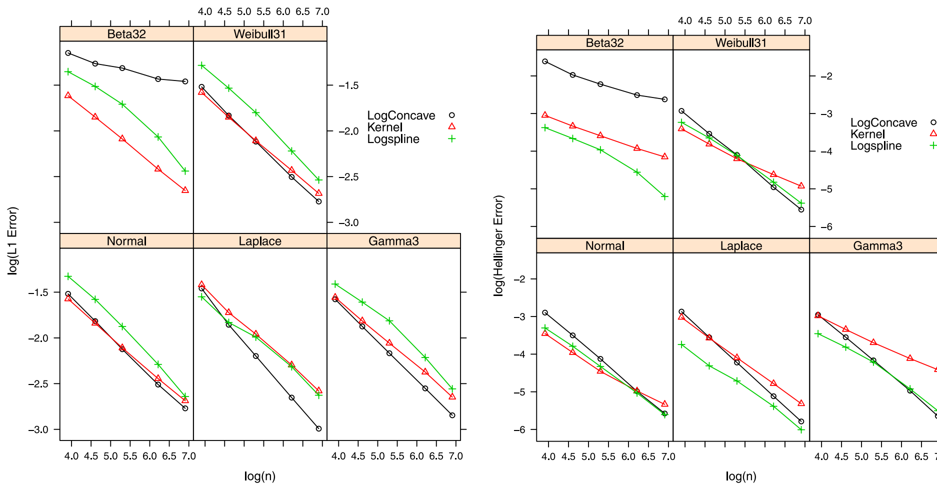


FIG. 4. *Log-concave estimate of the density of Student's criminals. Contours are labeled in units of log-density.*



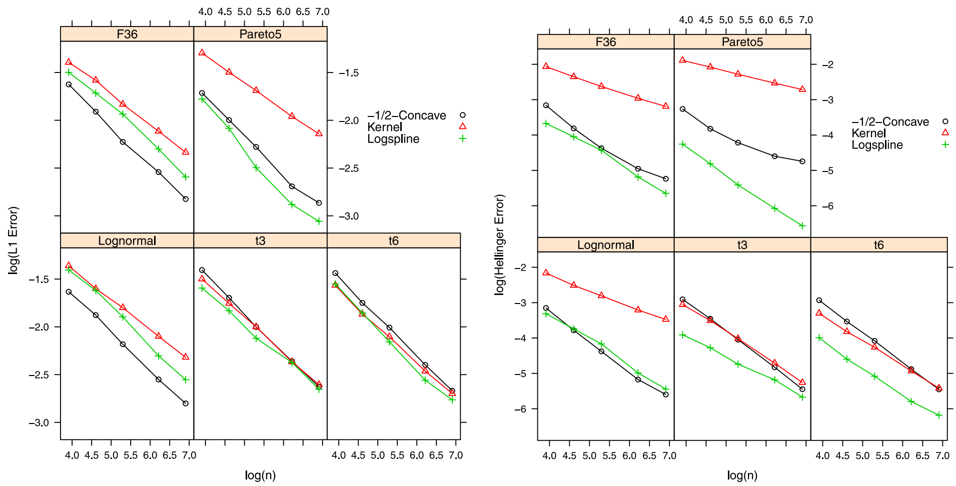
(a) L1 Error (b) Hellinger Error

FIG. 5. *Comparison of estimators of several log-concave densities.*

shape constrained estimators and evaluate whether consistency appeared to be a plausible conjecture. For the log-concave estimator Pal, Woodroofe and Meyer (2007) report "Hellinger error" for a fully crossed design involving five target densities and five sample sizes with 500 replications per cell.

In Figure 5, we report results of our attempt to reproduce the PWM experiment expanded somewhat to consider two competing estimators: the adaptive kernel estimator of Silverman (1986) using a Gaussian kernel, and the logspline estimator of Kooperberg and Stone (1991) as implemented in the `logspline` R package of Kooperberg. Five target densities are considered: (standard) normal, Laplace, Gamma(3), Beta(3, 2) and Weibull(3, 1) as in PWM. Five sample sizes are studied: 50, 100, 200, 500, 1000. And two measures of performance are considered: squared Hellinger distance as in PWM in the left panel and L_1 distance in the right panel. Plotted points in these figures represent cell means. Both figures support the contention that the rates of convergence are comparable for all three estimators.

Figure 6 reports results from a similar experiment for the the $-1/2$-concave estimator described in Section 3.6. We consider five new target densities: lognormal, t_3, t_6, $F_{3,6}$ and Pareto(5), all of which fall into the $-1/2$-concave class. The same competing estimators and sample sizes are used. In a small fraction of cases for the second group of densities, less than 0.2 percent, there were problems either with the convergence of the logspline or shape-constrained estimator, or with the numerical integration required to evaluate the performance measures, so Figure 6 plots cell medians rather than cell means. Again, the figures support the conjecture that the rates of convergence for the shape-constrained estimator are competitive with those of the adaptive kernel and logspline estimators.



(a) L1 Error (b) Hellinger Error

FIG. 6. *Comparison of estimators of several* $-1/2$*-concave densities.*

A concise way to summarize results from these experiments is to estimate the simple linear model

$$\log(y_{ij}) = \alpha_i + \beta \log(n_j) + u_{ij},$$

where y_{ij} denotes a cell average of our two error criteria for one of our three estimators, for target density i and sample size n_j. In this rather naïve framework, $\hat{\beta}$ can be interpreted as an empirical rate of convergence for the estimator. In Tables 1 and 2, we report these estimates suppressing the estimated target density specific α_i's. In this comparison too, the shape constrained estimators perform quite well.

6. Extensions and conclusions. We have described a rather general approach to qualitatively constrained density estimation. Log-concave densities are an important target class, but other, weaker, concavity requirements that permit algebraic tail behavior are also of considerable practical interest. Ultimately, the approach accommodates all quasi-concave densities as a limit of the Rényi entropy family.

TABLE 1
Estimated convergence rates for log-concave target densities

Criterion	Log-concave	Kernel	Logspline
L1 error	−0.417	−0.366	−0.393
	(0.018)	(0.003)	(0.012)
Hellinger	−0.875	−0.498	−0.698
	(0.032)	(0.031)	(0.021)

TABLE 2
Estimated convergence rates for $-1/2$-concave target densities

Criterion	**$-1/2$-concave**	**Kernel**	**Logspline**
L1 error	-0.405	-0.324	-0.386
	(0.004)	(0.008)	(0.01)
Hellinger	-0.751	-0.355	-0.672
	(0.034)	(0.023)	(0.019)

There are many unexplored directions for future research. As we have seen, a consequence of the variational formulation of our concavity constraints is that the estimated densities vanish off the convex hull of the data. Various treatments for this malady may be suggested. Müller and Rufibach (2009) have recently suggested applying one of several estimators of the Pareto tail index to the smoothed ordinates from the log-concave preliminary density estimator. Our inclination would be to prefer solutions that impose further regularization on the initial problem. Thus, for example, we can add a new penalty term to the primal problem, penalizing the total variation of the derivative (gradient) of $\log f$, and choosing a suitable value of the associated Lagrange multiplier to smooth the tail behavior at the boundary.

We have adhered, thus far, to the principle that the entropy choice in the fidelity criterion of the dual problem should dictate the form of the convexity constraint: likelihood thus implies log-concavity, Hellinger fidelity implies $1/\sqrt{f}$ concavity, etc. One may wish to break this linkage and consider maximum likelihood estimation of general ρ-concave densities. This may have some advantages from an inference viewpoint, at the cost of complicating the numerical implementation.

Embedding shape constrained density estimation of the type considered here into semiparametric methods would seem to be an attractive option in many settings. And, it would obviously be useful to consider inference for the validity of shape constraints in the larger context of penalized density estimation. We hope to pursue some of these issues in future work.

APPENDIX A: PROOFS

PROOF OF THEOREM 2.1. Given h convex, put $Y_i = h(X_i)$ and take $g = g_{(X,Y)}$, the function defined by (2.6). The convexity of h implies that $h(x) \le g(x)$ for every x; since ψ is nonincreasing, we have

$$c = \int \psi(h(x))\,dx - \int \psi(g(x))\,dx \ge 0. \tag{A.1}$$

The definition of $g_{(X,Y)}$ implies that $h(X_i) = g_{(X,Y)}(X_i) = Y_i$; therefore, the rest of Φ remains unchanged, and (A.1) implies that $\Phi(g) \le \Phi(h)$.

Suppose that $h \notin \mathcal{G}(X)$. Then $h \neq g$ and the inequality $h \leq g$ implies that $\operatorname{dom} g \subseteq \operatorname{dom} h$. If $\operatorname{dom} h \neq \operatorname{dom} g$, then there is a point $x_0 \notin \operatorname{dom} g$ from the interior of $\operatorname{dom} h$, because $\operatorname{dom} g$ is closed. The continuity of h at x_0 implies that $h(x) \leq K < +\infty = g(x)$ for all x from an open neighborhood of x_0; this proves that $c > 0$. If $\operatorname{dom} h = \operatorname{dom} g = \mathcal{H}(X)$, then the polyhedral character of $\mathcal{H}(X)$ implies through Theorem 10.2 of Rockafellar (1970) that h is upper semicontinuous relative to $\mathcal{H}(X)$ at any $x \in \mathcal{H}(X)$; that is, if $h(x_0) < g(x_0)$ for some $x_0 \in \operatorname{dom} h$, then the inequality holds for all x in an open, relatively to $\mathcal{H}(X)$, neighborhood of x_0. Such a relative neighborhood has positive Lebesgue measure, due to the fact that the interior of $\mathcal{H}(X)$ is nonempty. Hence, we have $c > 0$ also in this case and the strict inequality $\Phi(g) < \Phi(h)$ follows. $\square$

PROOF OF THEOREM 3.1. We use the conventional notation $\langle \ell, x\rangle$ to denote $\ell(x)$, the result of the application of a linear functional to x. The definition of the conjugate of a convex function H in this notation is

$$H^*(y) = \sup_{x \in \operatorname{dom} F} \bigl(\langle y, x\rangle - F(x)\bigr);$$

the resulting function is convex itself, being a sup of affine functions. For any $f \in \mathcal{C}(X)$ and any Radon measure G, a linear functional from $\mathcal{C}(X)^*$, we have

$$\langle G, f\rangle = \int f\, dG.$$

We start the proof by rewriting (3.1) as

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g(x))\, dx + \iota_{\mathcal{K}(X)}(g) = \Phi(g) + \Upsilon(g) = \inf_g !,$$

where Φ is the original objective function of (3.1) and $\Upsilon = \iota_{\mathcal{K}(X)}$ is the indicator function of $\mathcal{K}(X)$. The expression for the Fenchel dual of this type of problem follows from Rockafellar (1966); see also Rockafellar (1974), Section 5, Example 11:

$$-\Phi^*(G) - \Upsilon^*(-G) = \max_G !$$

(note that one of the conjugates, in both cites sources, is in the "concave" sense, which explains the negative sign of the argument in the second term, but not in the first). The conjugate of the indicator of a convex cone $\mathcal{K}(X)$ is the indicator of $-\mathcal{K}(X)^-$ [Rockafellar (1974), Section 3, equation (3.14)]. The term $-\Upsilon^*(-G)$ in the objective can be therefore interpreted as a constraint $-G \in -\mathcal{K}(X)^-$, that is, $G \in \mathcal{K}(X)^-$. The definition of the conjugate of Φ gives

$$\begin{aligned}\Phi^*(G) &= \sup_g\left(\langle G, g\rangle - \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \int \psi(g(x))\, dx\right)\\ &= \sup_g\left(\langle G - P_n, g\rangle - \int \psi(g(x))\, dx\right) = \Psi^*(G - P_n);\end{aligned}\tag{A.2}$$

the sup is taken over all g from

$$\operatorname{dom}\Phi = \left\{ g \in \mathcal{C}(X) \,\Big|\, \int \psi(g(x))\,dx < +\infty \right\} = \operatorname{dom}\Psi,$$

where Ψ is the functional given by

$$\Psi(g) = \int \psi(g(x))\,dx, \tag{A.3}$$

and Ψ^* is its conjugate. The form of the latter is given by Rockafellar (1971), Corollary 4A: if G is absolutely continuous with respect to the Lebesgue measure, then

$$\Psi^*(G) = \int \psi^*\left(\frac{dG}{dx}\right) dx, \tag{A.4}$$

otherwise $\Psi^*(G) = +\infty$. These facts, and expressions (A.2) and (A.4), yield (3.3).

Rockafellar [(1966), Theorem 1; see also Rockafellar (1974), Section 8, Example 11′] gives also a constraint qualification for this type of problem: to prove strong duality, we need to find some g where both Φ and Υ are finite and one of them is continuous. Such a g is provided by a function constant on $\mathcal{H}(X)$, say $g(x) = 1$ for all $x \in \mathcal{H}(X)$. It is convex, thus $\Upsilon(g) = 0$ is finite. So is $\Phi(g)$; the topology on $\mathcal{C}(X)$ is that of uniform convergence, and ψ is continuous at 1, hence there is a neighborhood of g containing only functions for which Φ is finite and Φ is continuous at g.

Once the constraint qualification is verified, we know that the primal and dual optimal values coincide (zero duality gap), and that the dual is attained—there is an optimal solution to the dual; see Theorem 52.B(3) of Zeidler (1985). Due to the fact that ψ is decreasing, $\psi^*(-f) = +\infty$ whenever $f < 0$; thus, if f yields a finite dual objective function, then f is nonnegative. If $G \in \mathcal{K}(X)^-$, then $\langle G, f\rangle \le 0$ for every $f \in \mathcal{K}(X)$; consequently,

$$0 \ge \langle G, 1\rangle = -\langle G, -1\rangle \ge 0.$$

Therefore, $\langle G, 1\rangle = 0$ and for every dual feasible f,

$$\int f(x)\,dx = \langle P_n - G, 1\rangle = \langle P_n, 1\rangle - \langle G, 1\rangle = \int 1\,dP_n = 1.$$

That is, every dual feasible f is a probability density with respect to the Lebesgue measure.

If a primal solution, g, exists—the fact that is established by Theorem 4.1, but here we are exploring only the consequences of such a premise—it is related to the dual solution, f, via extremality (Karush–Kuhn–Tucker) conditions. The form of this relationship asserted by the theorem follows from the second condition of (8.24) in Rockafellar [(1974), Section 8, Example 11′], together with the form of the subgradient of Ψ given by Rockafellar [(1971), Corollary 4B], combined with the fact established above that the estimated density f corresponds to $F = P_n - G$. $\square$

PROOF OF THEOREM 4.1. By Theorem 2.1, any potential solution of (2.5) lies within the class $\mathcal{G}(X)$ of polyhedral functions; due to the one–one correspondence between $\mathcal{G}(X)$ and $\mathcal{D}(X)$, the set of vectors discretely convex relative to X, the existence proof can be carried for (3.1) reparametrized by Y_i, the putative values of $g(X_i)$. In what follows, X remains fixed, and α, β will denote generic coefficients of convex combinations: any real numbers satisfying $\alpha, \beta \geq 0$, $\alpha + \beta = 1$.

As the correspondence between $\mathcal{G}(X)$ and $\mathcal{D}(X)$ is not a linear mapping (except for $d = 1$), the first thing to be shown is that (3.1) remains a convex problem when reparametrized in terms of a vector $Y \in \mathbb{R}^n$, with components $Y_1, \dots, Y_n$. The resulting problem minimizes, over $Y \in \mathbb{R}^n$, the objective function

$$\begin{aligned}\Phi_{\mathcal{D}}(Y) &= \frac{1}{n}\sum_{i=1}^{n} Y_i + \int \psi\bigl(g_{(X,Y)}(x)\bigr)\,dx \qquad \text{if } Y \in \mathcal{D}(X)\\ &= +\infty \qquad \text{otherwise.}\end{aligned}$$

Note first that $\mathcal{D}(X)$ is a convex subset of $\mathbb{R}^n$: if $Y, Z \in \mathcal{D}(X)$, then there exist convex functions g, h satisfying $Y_i = g(X_i)$ and $Z_i = h(X_i)$; subsequently, $\alpha Y_i + \beta Z_i = \alpha g(X_i) + \beta h(X_i)$, and as $\alpha g + \beta h$ is also a convex function, we obtain that $\alpha Y + \beta Z \in \mathcal{D}(X)$ for any convex combination of Y and Z. Thus, it is sufficient to show that $\Phi_{\mathcal{D}}$ is convex on $\mathcal{D}$, which amounts to demonstrating the convexity of the function $Y \mapsto \int \psi(g_{(X,Y)}(x))\,dx$. Let $Y, Z \in \mathcal{D}(X)$; the definition (2.6) gives for their convex combination

$$\begin{aligned}g_{(X,\alpha Y+\beta Z)}(x) &= \inf\left\{\sum_{i=1}^{n} \lambda_i(\alpha Y_i + \beta Z_i) \,\middle|\, x = \sum_{i=1}^{n} \lambda_i X_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0\right\}\\ &\geq \alpha \inf\left\{\sum_{i=1}^{n} \lambda_i Y_i \,\middle|\, x = \sum_{i=1}^{n} \lambda_i X_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0\right\}\\ &\quad + \beta \inf\left\{\sum_{i=1}^{n} \lambda_i Z_i \,\middle|\, x = \sum_{i=1}^{n} \lambda_i X_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0\right\}\\ &= \alpha g_{(X,Y)}(x) + \beta g_{(X,Z)}(x).\end{aligned}$$

As ψ is nonincreasing and convex, we obtain

$$\begin{aligned}&\int \psi\bigl(g_{(X,\alpha Y+\beta Z)}(x)\bigr)\,dx\\ &\quad \leq \int \psi\bigl(\alpha g_{(X,Y)}(x) + \beta g_{(X,Z)}(x)\bigr)\,dx\\ &\quad \leq \int \alpha\psi\bigl(g_{(X,Y)}(x)\bigr) + \beta\psi\bigl(g_{(X,Z)}(x)\bigr)\,dx\\ &\quad = \alpha \int \psi\bigl(g_{(X,Y)}(x)\bigr)\,dx + \beta \int \psi\bigl(g_{(X,Z)}(x)\bigr)\,dx\end{aligned} \tag{A.5}$$

as was required. Note that the integral is also finite whenever Y has all components in the domain of ψ, due to the polyhedral character of $g_{(X,Y)}(x)$ and the fact that ψ is nonincreasing and $\mathcal{H}(X)$ is bounded. Otherwise, it may be equal only to $+\infty$; hence $\Psi_{\mathcal{D}}$ is a proper convex function.

LEMMA A.1. *Suppose that Y, Z are vectors in $\mathbb{R}^d$ such that $Y = (y, \dots, y)$ has constant components, and Z is arbitrary. For any $\tau > 0$,*

$$g_{(X,Y+\tau Z)}(x) = g_{(X,Y)}(x) + \tau g_{(X,Z)}(x) = y + \tau g_{(X,Z)}(x). \tag{A.6}$$

PROOF. Note first that by the definition, $g_{(X,Y)}(x) = y$ identically on $\mathcal{H}(X)$ for constant Y; likewise, for every $x \in \mathcal{H}(X)$,

$$\begin{aligned} g_{(X,Y+\tau Z)}(x) &= \inf\left\{\sum_{i=1}^n \lambda_i (Y_i + \tau Z_i) \,\Big|\, x = \sum_{i=1}^n \lambda_i X_i, \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0\right\} \\ &= \inf\left\{y + \tau \sum_{i=1}^n \lambda_i Z_i \,\Big|\, x = \sum_{i=1}^n \lambda_i X_i, \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0\right\} \\ &= y + \tau \inf\left\{\sum_{i=1}^n \lambda_i Z_i \,\Big|\, x = \sum_{i=1}^n \lambda_i X_i, \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0\right\} \\ &= g_{(X,Y)}(x) + \tau g_{(X,Z)}(x), \end{aligned}$$

proving the lemma. $\square$

Choose a real number y lying in the domain of ψ, and set $Y = (y, \dots, y)$. According to Lemma A.1, $g_{(X,Y)}(x) = y$ for every $x \in \mathcal{H}(X)$; the function constant on $\mathcal{H}(X)$ is convex, hence $Y \in \mathcal{D}(X)$. Then

$$\Phi_{\mathcal{D}}(Y) = \frac{1}{n}\sum_{i=1}^n Y_i + \int \psi\big(g_{(X,Y)}(x)\big)\,dx = y + \psi(y)\operatorname{Vol}(\mathcal{H}(X)) < +\infty;$$

therefore, Y lies in the domain of $\Phi_{\mathcal{D}}$. For arbitrary $Z \in \mathbb{R}^n$, not identically 0, and $\tau > 0$, we have

$$\begin{aligned} \Phi_{\mathcal{D}}(Y + \tau Z) &= \frac{1}{n}\sum_{i=1}^n (Y_i + \tau Z_i) + \int \psi\big(g_{(X,Y+\tau Z)}(x)\big)\,dx \\ &= y + \frac{\tau}{n}\sum_{i=1}^n Z_i + \int \psi\big(y + \tau g_{(X,Z)}(x)\big)\,dx \\ &= y + \tau\left(\frac{1}{n}\sum_{i=1}^n Z_i + \int \frac{\psi(y + \tau g_{(X,Z)}(x))}{\tau}\,dx\right). \end{aligned}$$

We know that $\Phi_{\mathcal{D}}$ is a convex function on a finite-dimensional linear space $\mathbb{R}^n$; to establish the existence of its minimizer, it is sufficient to show that $\Phi_{\mathcal{D}}(Y + \tau Z) \to +\infty$ for $\tau \to +\infty$, which means that we need to verify that the limit of the expression in the parentheses is positive (possibly $+\infty$); see also Hiriart-Urruty and Lemaréchal (1993), Remark 3.2.8. Let E^-, E^0, E^+ be sets in $\mathcal{H}(X)$ where $g(x) = g_{(X,Z)}(x)$ is, respectively, negative, zero or positive; we are to examine the limit behavior of

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_i + \int_{E^-} \frac{\psi(y+\tau g(x))}{\tau}\,dx \\ + \int_{E^0} \frac{\psi(y)}{\tau}\,dx + \int_{E^+} \frac{\psi(y+\tau g(x))}{\tau}\,dx,\end{aligned} \tag{A.7}$$

when $\tau \to +\infty$. For the integral over E^0, the limit is obviously zero. If ψ satisfies assumptions (A1) and (A5), then ψ is nonincreasing and converging to 0 when $\tau \to +\infty$; for every $x \in E^+$ then $\psi(y + \tau g(x))/\tau$ monotonically decreases with increasing τ, hence the limit of the integral over E^+ is zero as well. Finally, if ψ satisfies also (A3), then for every $x \in E^-$ the limit of $\psi(y + \tau g(x))/\tau$ is $+\infty$; at the same time, the expression is bounded from below by 0, due to (A4). The application of the Fatou lemma then gives that the limit of the integral over E^- is $+\infty$, whenever E^+ has positive Lebesgue measure.

The proof of the theorem is then finished by the examination of possible alternatives. If the first term in (A.7), the mean of the Z_i's, is positive, then the theorem is proved, as all other terms in (A.7) converge either to 0 or $+\infty$. If the first term of (A.7) is negative or equal to zero, then there must be some $Z_i < 0$ (the case with all $Z_i = 0$ is excluded). That means that $g(x) = g_{(X,Z)}(x)$ is negative for some x; this implies that E^- has positive Lebesgue measure, and then the limit of the second term in (A.7), the integral over E^-, and thus of the whole expression (A.7) is $+\infty$. This proves the theorem.

Under the strict convexity of ψ, the strict convexity of $\Phi_{\mathcal{D}}$ follows (for appropriate α, β) from the second inequality in (A.5), which becomes sharp—this is due to the fact that the sharp inequality holds pointwise for all x, and thus for the integrals as well. The strict convexity of $\Phi_{\mathcal{D}}$ then implies the uniqueness of the solution.

Finally, functions ψ satisfying (A1)–(A3), but not necessarily (A5) require some special considerations. For the integral over E^-, we have to observe that for every $\tau > 0$ we have $\psi(y + \tau g_{(X,Z)}(x)) \geq \psi$; if $\psi(y) < 0$, then $\psi(y)/\tau \geq \psi(y)$ for every $\tau \geq 1$, if $\psi(y) \geq 0$ then $\psi(y)/\tau \geq 0$ for every $\tau > 0$. Hence we have also in this case an integrable constant minorant [due to the fact that $\mathcal{H}(X)$ has finite Lebesgue measure]; this justifies the limit transition via the Fatou lemma. Finally, for the integral over E^+, we need to assume that the limit of the integrand is 0 for every x, and find an integrable minorant; this may be related to the existence of the integral of the second term in (2.5). $\square$

PROOF OF THEOREM 4.2. The proof relies on the application of what is called Fenchel's inequality by Rockafellar [(1970), page 105] or the (generalized) Young inequality by Hardy, Littlewood and Pólya [(1934), Section 4.8], or Zeidler [(1985), Section 51.1]. The inequality says says that for arbitrary x, y and convex function ψ

$$xy \le \psi(x) + \psi^*(y).$$

Applied pointwise to $x = -f_0$ and $y = \chi^{-1}(-f)$, the inequality yields

$$(-f_0)\chi^{-1}(-f) \le \psi^*(-f_0) + \psi(\chi^{-1}(-f)),$$

which is equivalent to the nonnegativity of the integrand in (4.4). For $f = f_0$, the equality (4.5) implies that

$$f_0\chi^{-1}(-f_0) + \psi^*(-f_0) + \psi(\chi^{-1}(-f_0)) = -\psi^*(-f_0) + \psi^*(-f_0) = 0,$$

which proves the theorem.

For functions not satisfying (A5), integrability of $f_0\chi^{-1}(-f_0)$ is no longer equivalent to that of $-\psi^*(-f_0)$. However, if we assume the integrability of the latter, then the proof can be carried through in the same way. □

APPENDIX B: COMPUTATIONAL DETAILS

Our computational objective is to provide a unified algorithmic strategy for solving the entire class of problems described above. Interior point methods designed for general convex programming and capable of exploiting the inherently sparse structure of the resulting linear algebra offer a powerful, general approach. We have employed two such implementations throughout our development process: the PDCO algorithm of Saunders (2003), and the MOSEK implementation of Andersen (2006).

Our generic primal problem (2.5) involves minimizing an objective function consisting of a linear component, representing likelihood or some generalized notion of fidelity, plus a nonlinear component, representing the integrability constraint. Minimization is then subject to a cone constraint imposing convexity. We will first describe our procedure for enforcing convexity, and then turn to the integrability constraint.

B.1. The convexity constraint. In dimension one convexity of piecewise linear functions can be imposed easily by enforcing linear inequality constraints on a set of function values, $\gamma_i = g(\xi_i)$ at selected points $\xi_1, \xi_2, \dots, \xi_m$. For ordered ξ_i's, the convex cone constraint can be written as $D\gamma \ge 0$ for a tridiagonal matrix D that does second differencing, adapted to the possible unequal spacing of the ξ_i's.

In dimension two, enforcing convexity becomes more of a challenge. Ideally, we would utilize knowledge of the polyhedral character of the optimal g, established by Theorem 2.1, and implying that the optimal g is piecewise linear over *some* triangulation of the observations $X_i \in \mathbb{R}^2$. Once we knew the triangulation, it is again straightforward to impose convexity: each interior edge of the triangulation generates one linear inequality on the coefficients γ. Unfortunately, the complexity of traveling over a binary tree of possible triangulations of the observed points makes finding the optimal one difficult. The algorithm implemented in the `logConcDEAD` package of Cule, Gramacy and Samworth (2009), for computing log-concave estimates, exploits special features of the log-concave MLE problem and thus does not appear to be easily generalizable to our other settings. Finite-element methods involving fixed (Delaunay) triangulation of an expanded set of vertices were also ultimately deemed unsatisfactory.

A superior choice, one that circumvents the difficulties of the finite-element, fixed triangulation approach, relies on finite differences. Convexity is imposed directly at points on a regular rectangular grid using finite-differences to compute the discrete Hessian:

$$\begin{aligned}
H_{11}(\xi_1, \xi_2) &= g(\xi_1 + \delta, \xi_2) - 2g(\xi_1, \xi_2) + g(\xi_1 - \delta, \xi_2),\\
H_{22}(\xi_1, \xi_2) &= g(\xi_1, \xi_2 + \delta) - 2g(\xi_1, \xi_2) + g(\xi_1, \xi_2 - \delta),\\
H_{12}(\xi_1, \xi_2) &= [g(\xi_1 + \delta, \xi_2 + \delta) - g(\xi_1 + \delta, \xi_2 - \delta)\\
&\qquad - g(\xi_1 - \delta, \xi_2 + \delta) + g(\xi_1 - \delta, \xi_2 - \delta)]/4,\\
H_{21}(\xi_1, \xi_2) &= H_{12}(\xi_1, \xi_2).
\end{aligned}$$

Convexity is then enforced by imposing positive semidefiniteness. These constraints—convexity at each of the grid points (ξ_{1i}, ξ_{2i})—produce a semi-definite programming problem. In the bivariate setting the semi-definiteness of each H can be reformulated as a rotated quadratic cone constraint; we need only constrain the signs of the diagonal elements of H and its determinant. This simplifies the implementation of the Hellinger estimator in MOSEK. For the relatively fine grid used for Figure 3 solution requires about 25 seconds, considerably quicker than the log-concave estimate of Figure 4 computed with the implementation of Cule, Gramacy and Samworth (2009).

B.2. The integrability constraint. For certain special ψ, one can evaluate the integral term $\int \psi(g(x))\,dx$ in the objective function of (2.5) explicitly—as was done for $\psi(g) = e^{-g}$ by Cule, Samworth and Stewart (2010). While such a strategy may also be possible for certain other specific ψ, we adopt a more pragmatic approach based on a straightforward Riemannian approximation

$$\int_{\mathcal{H}(X)} \psi(g(x))\,dx \approx \sum_{i=1}^{m} \psi(g(\xi_i))s_i. \tag{B.1}$$

Here, s_i are weights derived from the configuration of ξ_i. Of course, with only a modest number of ξ_i's such an approximation may be poor; in dimension one we therefore augment the initial collection of the observed data $X_1, \dots, X_n$ by filling the gaps between their order statistics by further grid points, to ensure that the resulting grid (not necessarily uniformly spaced) and consisting of the observed data points as well as the new grid points, provides a sufficiently accurate approximation (B.1). The s_i's are then simply the averages of the adjacent spacings between the ordered ξ_i's. Given the size of problems modern optimization software can successfully handle, it is no problem to add an abundance of new points in dimension one.

In dimension two, the approximation (B.1) is based on the uniformly spaced grid of the points used in the finite-difference approach described in the previous subsection. As the original data points X_i may no longer lie among the grid points ξ_i, we have to modify the fidelity component of the objective function: instead of obtaining $g(X_i)$ directly, we obtain it via linear interpolation from the values of g at the vertices of the rectangles enclosing X_i. As long as the grid is sufficiently fine, the difference is minimal. We use this approach often also in dimension one, as it provides better numerical stability especially for fine grids and large data sets.

B.3. Discrete duality. Adopting the procedures described above, we can write the finite-dimensional version of the primal problem as

$$\{w^\top L\gamma + s^\top \Psi(\gamma) | D\gamma \geq 0\} = \min!, \tag{P}$$

where $\Psi(\gamma)$ denotes now the m-vector with typical element $\psi(g(\xi_i)) = \psi(\gamma_i)$, L is an "evaluation operator" which either selects the data elements from γ, or performs the appropriate linear interpolation from the neighboring ones, so that $L\gamma$ denotes the n-vector with typical element, $g(X_i)$, and w is an n-vector of observation weights, typically $w_i \equiv 1/n$.

Associated with the primal problem (P) is the dual problem

$$\{-s^\top \Psi^*(-\phi) | S\phi = -w^\top L + D^\top \eta, \phi \geq 0, D^\top \eta \geq 0\} = \max!. \tag{D}$$

Here, η is an m-vector of dual variables and ϕ is an m-vector of function values representing the density evaluated at the ξ_i's, and $S = \operatorname{diag}(s)$. The vector Ψ^* is the convex conjugate of Ψ defined coordinate-wise with typical element $\psi^*(y) = \sup_x\{yx - \psi(x)\}$. Problems (P) and (D) are strongly dual in the sense of the following result, which may viewed as the discrete counterpart of Theorem 3.1.

PROPOSITION B.1. *If ψ is convex and differentiable on the interior $\mathcal{I}$ of its domain, then the corresponding solutions of* (P) *and* (D) *satisfy*

$$f(\xi_i) = \psi'(g(\xi_i)) \qquad \text{for } i = 1, \dots, m, \tag{E}$$

whenever the elements of g are from $\mathcal{I}$ and the elements of f are from the image of $\mathcal{I}$ under ψ'.

For $\Psi(x)$ with typical element $\psi(x) = e^{-x}$ we have Ψ^* with elements $\psi^*(y) = -y \log y + y$, so the dual problem corresponding to maximum likelihood can be interpreted as maximizing the Shannon entropy of the estimated density subject to the constraints appearing in (D). Since g was interpreted in (P) as $\log f$, this result justifies our interpretation of solutions of (D) as densities provided that they satisfy our integrability condition. This is easily verified and thus justifies the implicit Lagrange multiplier of one on the integrability constraint in (P), giving a discrete counterpart of Theorem 3.1.

PROPOSITION B.2. *Let ι denote an m-vector of ones, and suppose in* (P) *that* $w^\top L\iota = 1$ *and* $D\iota = 0$. *Then solutions* ϕ *of* (D) *satisfy* $s^\top \phi = 1$ *and* $\phi \geq 0$.

The crucial element of the proof is that the differencing operator D annihilates the constant vector and therefore the result extends immediately to other norm-type penalties as well as to the other entropy objectives that we have discussed. Indeed, since the second difference operator representing our convexity constraint annihilates any affine function it follows by the same argument that the mean of the estimated density also coincides with the sample mean of the observed X_i's.

Acknowledgments. We are grateful to Lutz Dümbgen, Kaspar Rufibach, Guenther Walther and Jon Wellner for sending us preprints of their work, to the referees for their very constructive comments, and to Mu Lin for help with the bright star example and Fisher consistency proof.

REFERENCES

ANDERSEN, E. D. (2006). The MOSEK optimization tools manual, Version 4.0. Available at http://www.mosek.com.

AVRIEL, M. (1972). r-convex functions. *Math. Program.* **2** 309–323. MR0301151

BALABDAOUI, F., RUFIBACH, K. and WELLNER, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Ann. Statist.* **37** 1299–1331. MR2509075

BASU, A., HARRIS, I. R., HJORT, N. L. and JONES, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85** 549–559. MR1665873

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719

BORELL, C. (1975). Convex set functions in d-space. *Period. Math. Hungar.* **6** 111–136. MR0404559

BORWEIN, J. and LEWIS, A. S. (2006). *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, New York. MR2184742

CAPLIN, A. and NALEBUFF, B. (1991). Aggregation and social choice: A mean voter theorem. *Econometrica* **59** 1–23. MR1085582

CHANG, G. T. and WALTHER, G. (2007). Clustering with mixtures of log-concave distributions. *Comput. Statist. Data Anal.* **51** 6242–6251. MR2408591

CULE, M., GRAMACY, R. and SAMWORTH, R. (2009). LogConcDEAD: An R package for maximum likelihood estimation of a multivariate log-concave density. *J. Statist. Software* **29** 1–20.

CULE, M., SAMWORTH, R. and STEWART, M. (2010). Computing the maximum likelihood estimator of a multidimensional log-concave density. *J. Roy. Statist. Soc. Ser. B*. To appear.

CULE, M. and SAMWORTH, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electron. J. Statist.* **4** 254–270.

DE FINETTI, B. (1949). Sulle stratificazioni convesse. *Ann. Mat. Pura Appl.* **30** 173–183. MR0043491

DHARMADHIKARI, S. and JOAG-DEV, K. (1988). *Unimodality, Convexity and Applications*. Academic Press, Boston. MR0954608

DÜMBGEN, L., HÜSLER, A. and RUFIBACH, K. (2007). Active set and EM Algorithms for log-concave densities based on complete and censored data. Available at arXiv:0707.4643v2.

DÜMBGEN, L. and RUFIBACH, K. (2009). Maximum likelihood estimation of a log-concave density: Basic properties and uniform consistency. *Bernoulli* **15** 40–68. MR2546798

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. *Skand. Aktuarietidskr.* **39** 125–153. MR0093415

GROENEBOOM, P., JONGBLOED, G. and WELLNER, J. A. (2001). Estimation of a convex function: Characterizations and asymptotic theory. *Ann. Statist.* **29** 1653–1698. MR1891742

GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York. MR1876599

HALL, P. and PRESNELL, B. (1999). Density estimation under constraints. *J. Comput. Graph. Statist.* **8** 259–277. MR1706365

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1934). *Inequalities*. Cambridge Univ. Press, London.

HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (1993). *Convex Analysis and Minimization Algorithms I, II*. Springer, Berlin. MR1261420

HOFFLEIT, D. and WARREN, W. H. (1991). *The Bright Star Catalog*, 5th ed. Yale Univ. Observatory, New Haven, CT.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. Probability* **1** 221–233. Univ. California Press, Berkeley. MR0216620

KARLIN, S. (1968). *Total Positivity*. Stanford Univ. Press, Stanford, CA.

KOENKER, R. and MIZERA, I. (2008). Primal and dual formulations relevant for the numerical estimation of a probability density via regularization. *Tatra Mt. Math. Publ.* **39** 255–264. MR2452059

KOOPERBERG, C. and STONE, C. J. (1991). A study of logspline density estimation. *Comput. Statist. Data Anal.* **12** 327–347.

MACDONELL, W. R. (1902). On criminal anthropometry and the identification of criminals. *Biometrika* **1** 177–227.

MIZERA, I. (2002). On depth and deep points: A calculus. *Ann. Statist.* **30** 1681–1736. MR1969447

MÜLLER, S. and RUFIBACH, K. (2009). Smooth tail-index estimation. *J. Stat. Comput. Simul.* **79** 1155–1167. MR2572422

PAL, J. K., WOODROOFE, M. and MEYER, M. (2007). Estimating a Polya frequency function. In *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond* (R. Liu, W. Strawderman and C.-H. Zhang, eds.). *IMS Lecture Notes-Monograph Series* **54** 239–249. IMS, Beachwood, OH. MR2459196

PRÉKOPA, A. (1973). On logarithmic concave measures and functions. *Acta Sci. Math. (Szeged)* **34** 335–343. MR0404557

RÉNYI, A. (1961). On measures of entropy and information. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* 547–561. Univ. California Press, Berkeley. MR0132570

RÉNYI, A. (1965). On the foundations of information theory. *Rev. d'Inst. Int. Statist.* **33** 1–14. MR0181483

ROCKAFELLAR, R. T. (1966). Extensions of Fenchel's duality theorem for convex functions. *Duke Math. J.* **33** 81–89. MR0187062

ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton Univ. Press, Princeton. MR0274683

ROCKAFELLAR, R. T. (1971). Integrals which are convex functionals, II. *Pacific J. Math.* **39** 439–469. MR0310612

ROCKAFELLAR, R. T. (1974). *Conjugate Duality and Optimization*. SIAM, Philadelphia. MR0373611

RUFIBACH, K. (2007). Computing maximum likelihood estimators of a log-concave density function. *J. Stat. Comput. Simul.* **77** 561–574. MR2407642

RUFIBACH, K. and DÜMBGEN, L. (2009). Logcondens: Estimate a log-concave probability density from iid observations. R package Version 1.3.4. Available at http://CRAN.R-project.org/package=logcondens.

SAUNDERS, M. A. (2003). PDCO: A primal-dual interior solver for convex optimization. Available at http://www.stanford.edu/group/SOL/software/pdco.html.

SCHENNACH, S. (2007). Estimation with exponentially tilted empirical likelihood. *Ann. Statist.* **35** 634–672. MR2336862

SEREGIN, A. and WELLNER, J. A. (2009). Nonparametric estimation of multivariate convex-transformed densities. Available at arXiv:0911.4151v1.

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43** 97–99. MR0610384

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810. MR0663433

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

STUDENT (1908). The probable error of the mean. *Biometrika* **6** 1–23.

WALTHER, G. (2001). Multiscale maximum likelihood analysis of a semiparametric model with applications. *Ann. Statist.* **29** 1298–1319. MR1873332

WALTHER, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *J. Amer. Statist. Assoc.* **97** 508–513. MR1941467

WALTHER, G. (2009). Inference and modeling with log-concave distributions. *Statist. Sci.* **24** 319–327.

ZEIDLER, E. (1985). *Nonlinear Functional Analysis and Its Applications III: Variational Methods and Optimization*. Springer, New York. MR0768749

DEPARTMENT OF ECONOMICS
UNIVERSITY OF ILLINOIS
410 DAVID KINLEY HALL
1407 W. GREGORY, MC-707
URBANA, ILLINOIS 61801
USA
E-MAIL: rkoenker@uiuc.edu

DEPARTMENT OF MATHEMATICAL
AND STATISTICAL SCIENCES
UNIVERSITY OF ALBERTA
CAB 632, EDMONTON, AB
T6G 2G1 CANADA
E-MAIL: mizera@stat.ualberta.ca