QUASI-CONCAVE DENSITY ESTIMATION

ROGER KOENKER AND IVAN MIZERA

Abstract. 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

Log-concave probability densities play a crucial role in a wide variety of economic models. In reliability theory, search and contract theory, social choice, and a broad range of game theoretic contexts it has proven convenient to posit densities whose logarithm is concave. Surveys by An (1998) and Bagnoli and Bergstrom (2005) catalogue applications of log-concavity in virtually every corner of the economic literature. Much of the credit for recognizing the importance of log-concavity may be traced back to work by Schoenberg and Karlin on total positivity beginning in the late 1940's and culminating in Karlin (1968).

Karlin's work 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. Barlow and Proschan (1975) pioneered the application of log-concavity in reliability theory. Flinn and Heckman (1983) considered search models in which log-concavity of the wage-offer density implied that the exit rate from unemployment is increasing in the rate of arrival of job offers. Goldberger (1983) employed related methods to study selection bias in regression models under non-Gaussian conditions. Heckman and Honoré (1990) use it to generalize the implications of the Roy model. In auction models with bidder's values determined as $v_i = \nu + u_i$ with u_i iid with density f across bidders, the v_i's are affiliated if and only if f is log-concave.

Log-concavity plays an important role in many multivariate settings as well. Caplin and Nalebuff (1991b) describe a mean-voter model in which voter preferences have a linear structure on a d-dimensional Euclidean space with preference parameters

Version: May 23, 2008. This research was partially supported by NSF grant SES-05-44673, and by the Natural Sciences and Engineering Research Council of Canada.

$\alpha \in \mathbb{R}^d$ distributed according to a log concave density. It is then shown that the mean voter's preferred outcome is preferred to any other proposed alternative in pairwise comparisons by at least a proportion, $(1 - 1/e) \approx .64$, of other voters. In a companion paper, Caplin and Nalebuff (1991a), study models of imperfect (spatial) competition employing similar preference assumptions. A notable feature of both of these papers is that they treat an extended class of ρ-concave densities that include not only the log-concave ones, but ultimately the class of all quasi-concave densities.

Most of the familiar parametric densities employed in economics are log-concave: the uniform, normal, exponential, logistic, Weibull, gamma, all belong; others like the Student t densities fail to be log-concave, but are ρ-concave. Distributions like the Student t family with algebraic tails play an increasingly important role in economics and finance, and this motivates our interest in estimating the broader class of ρ-concave densities. It is relatively easy to determine whether a given parametric density is log-concave or satisfies weaker conditions for ρ-concavity, but in some empirical settings we may wish to impose some form of concavity without simultaneously imposing a particular parametric form. But beyond these parametric families lies *terra incognita*.

Our objective in this paper is to introduce a general framework for non-parametric estimation of a class of densities that satisfy a concavity requirement. This framework will enable us to treat not only univariate densities, but multivariate densities as well. We will consider not only maximum likelihood as a fidelity criterion, but also other measures like Hellinger discrepancy. Log-concave densities will play an important role in the sequel, but we will also consider weaker concavity constraints that permit algebraic tail behavior. The common feature of all of our estimated densities is that they will be required to be quasi-concave, in the sense of Arrow and Enthoven (1961).

2. Classes of Quasi-Concave Probability Densities

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 taking 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 use the convention $-\log 0 = +\infty$.

Unimodality of concave functions implies that log-concave densities are unimodal. However, unimodal densities need not be log-concave; none of the Student t_ν densities are log-concave for $\nu < +\infty$. Laplace densities, with their exponential tail behavior, are log-concave; but heavier—algebraic, say—tails are prohibited. This prohibition motivates a relaxation of the log-concavity requirement.

A natural hierarchy of concave functions can be built on the foundation of the weighted means of order ρ studied in Hardy, Littlewood, and Pólya (1934):

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

for p in the unit simplex, $\mathcal{S} = \{p \in \mathbb{R}^n | p \geq 0, \sum p_i = 1\}$. The familiar arithmetic, geometric and harmonic means correspond to $\rho = 1$, $\rho = 0$ and $\rho = -1$ respectively. Following Avriel (1972), a non-negative, real function f, defined on a convex set $C \subset \mathbb{R}^d$ is said to be ρ-*concave* if for any $x_0 \in C$, $x_1 \in C$, and $p \in \mathcal{S}$,

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

In this terminology log-concave functions are 0-concave, and concave functions are 1-concave.

The hierarchy of ρ-concave density functions was introduced to the economics literature by Caplin and Nalebuff, who demonstrated its relevance in spatial models of voting and imperfect competition. In their mean voter model they showed that if the distribution of preference parameters has a ρ-concave density, then the mean voter's preferred outcome is preferred by at least a proportion $1 - \delta$ of voters to any other proposed outcome in pairwise comparisons, where

$$\delta(d,\rho) = 1 - \left[\frac{d + 1/\rho}{d + 1 + 1/\rho}\right]^{d+1/\rho}.$$

Evaluating this expression as $\rho \to 0$ yields the log-concave limit $\delta_0 = 1 - 1/e \approx .64$ irrespective of the dimension d. This result can be interpreted as a generalization of Black's (1948) celebrated median voter result for univariate spatial voting with unimodal preferences. Caplin and Nalebuff's result has intriguing connections to Tukey's half-space depth in multivariate statistics, see Mizera (2002).

Since $M_\rho(a,p)$ 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,

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

and therefore are *quasi-concave*, a class that includes all ρ-concave functions.

In view of the importance of log-concave densities throughout economics and statistics it seems surprising that estimation of and inference about such densities has been so neglected. Concavity constraints have played a significant role in non-parametric regression since the seminal work of Hildreth (1954), but shape restricted estimation of densities has has received little attention until very recently. Our effort to help remedy this neglect begins by considering estimation of log-concave densities, then broadens in scope to include estimation of other ρ-concave densities.

3. Maximum Likelihood Estimation of Log-Concave Densities

Recently, Rufibach and Dümbgen (2006) and Pal, Woodroofe, and Meyer (2006), investigated an earlier proposal of Eggermont and LaRiccia (2001) to estimate log-concave probability densities via maximum likelihood. Only the univariate case was considered; nevertheless, the general multivariate formulation is straightforward. Let $X = \{X_1, \cdots, X_n\}$ be a collection of data points in $\mathbb{R}^d$. We assume 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 the X_i behave like a sample from a (probability) density f_0 and $n \geq d$. Thinking of the X_i as a random sample from an unknown, log-concave density f_0 on $\mathbb{R}^d$, we can write the minimization problem for the *maximum likelihood estimate* of that density as

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

It is convenient to recast (1) in terms of $g = -\log f$, the estimate becoming $f = \mathrm{e}^{-g}$,

$$\sum_{i=1}^{n} g(X_i) = \min_g ! \qquad \text{subject to } g \in \mathcal{K}, \text{ and } \int \mathrm{e}^{-g}\, dx = 1, \tag{2}$$

where $\mathcal{K}$ stands for the set of all convex functions on $\mathbb{R}^d$. In the spirit of the convention adopted above, the objective function of (2) is equal to $+\infty$, unless all X_i are in the domain of g.

It is well-known that naïve, unrestricted maximum likelihood estimation is bound to fail when applied in the general density estimation context: once "log-concave" is dropped from the formulation of (1), any sequence of putative maximizers is attracted to the "Dirac catastrophe", the linear combination of point masses located at the data points. A possible way out is regularization—a well-posed problem can be obtained by adding a roughness/complexity penalty to the likelihood function. Various proposals in this vein can be found in Good (1971), Silverman (1982), Gu (2002), and Koenker and Mizera (2008).

Regularization can also be achieved by imposing shape constraints. This line of development dates back to the celebrated Grenander (1956) nonparametric maximum likelihood estimator for monotone densities. While monotonicity serves to regularize the maximum likelihood estimator, unimodality *per se*—somewhat surprisingly—does not. Regularization is achieved only through somewhat more stringent shape constraints. Strong unimodality, that is to say log-concavity, provides a leading example, which, as we will demonstrate in the next section, can be formulated as,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \mathrm{e}^{-g}\, dx = \min_{g \in \mathcal{C}(X)} \qquad \text{subject to} \quad g \in \mathcal{K}, \tag{3}$$

where $\mathcal{C}(X)$ is the collection of all continuous functions on $\mathcal{H}(X)$. The crucial difference between (2) and (3) is that the latter is a *convex* problem, while the former not.

An advantage of shape-constrained regularization is that it is not encumbered by the need to select additional tuning parameters, as with more familiar norm-constraints, but on the other hand, it is limited in scope: applicable only to situations when the shape-constraint is plausible for the unknown density.

4. Quasi-Concave Density Estimation

Generalizing (3) to accommodate further possibilities we replace e^{-g} in the integral by a generic convex function $\psi(g)$. We obtain a new objective function Ψ in the *primal formulation*

$$\Psi(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g)\,dx = \min_{g\in\mathcal{C}(X)}! \qquad \text{subject to } g \in \mathcal{K}. \tag{4}$$

Again, we consider the objective function $\Psi(g)$ to be equal to $+\infty$, unless all X_i are in the domain of g. In what follows, we will assume that the function ψ is convex, nonincreasing, and finite on $(0, +\infty)$; its domain is thus $[\omega, +\infty)$ or $(\omega, +\infty)$, where in the latter case ω may be $-\infty$. If the domain is $(\omega, +\infty)$, then we assume that for any $c \geq 0$, the limit, for $x \to \omega$, of $\psi(x) - cx$ is $+\infty$; in particular, $\psi(x) \to +\infty$ when $x \to \omega$. Note that $+\infty$ is the limit also for $x \to +\infty$ and $c > 0$, due to the nondecreasing character of ψ. Given these conditions, we distinguish two types of ψ functions:

A. those bounded from below, for which we assume that the bound is zero and $\psi(x) \to 0$ when $x \to +\infty$; and that $\psi(x)$ is decreasing on the set of all x such that $\psi(x) > 0$;
B. those unbounded from below, for which we assume that they are decreasing everywhere on their domain.

For Type A functions, like $\psi(g) = \mathrm{e}^{-g}$ appearing in (3), the integral $\int \psi(g)\,dx$ is equal to that restricted to the domain of g, and is thus either finite or $+\infty$. For Type B functions more caution is called for, because a straightforward treatment may bring formal complications regarding the existence of integrals; we need to handle the integrals in a more delicate manner, in the spirit of singular integrals of calculus of variations, but because we use Type B functions in this paper only marginally, we do not pursue a full development here.

By definition, the objective function $\Psi(g)$ is infinite whenever $g(X_i) = +\infty$ for some X_i. On the other hand, any function g equal to some positive constant on $\mathcal{H}(X)$ and $+\infty$ elsewhere yields $\Psi(g) < \infty$ (this is true also for Type B functions, because for those, g can be set equal everywhere to a constant such that $\psi(g) = 0$). These considerations motivate the choice of $\mathcal{C}(X)$ as the optimization domain: it is sensible to assume that the domain of any g contains the closed set $\mathcal{H}(X)$, and then

any convex function is continuous on the interior of its domain. It is not hard to verify that Ψ is convex, and because the intersection of $\mathcal{C}(X)$ and $\mathcal{K}$ is convex, we conclude that (4) is indeed a feasible convex problem. Unlike in (3), $\psi(g)$ is not necessarily the estimated density f; the relationship of g to f is revealed in Theorem 3.

The existence of the solution of (4) follows from the characterization of the shape of the estimates obtained through (4). For every collection (X, Y) of points $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$, we define a function

$$g_{(X,Y)}(x) = \inf\left\{\sum_{i=1}^{n} \lambda_i Y_i \mid x = \sum_{i=1}^{n} \lambda_i X_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0\right\}. \tag{5}$$

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. For fixed X, we will denote the collection of all functions $g_{(X,Y)}$ by $\mathcal{G}(X)$. The convention $\inf \emptyset = +\infty$ used in (5) means that the domain of $g_{(X,Y)}$ is equal to $\mathcal{H}(X)$; hence $\mathcal{G}(X) \subset \mathcal{K} \subset \mathcal{C}(X)$. 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; consequently, $g_{(X,Y)}$ is the maximum of convex functions with this property, the lower convex hull of points (X_i, Y_i).

The collection (X, Y) determines $g_{(X,Y)}$ uniquely, by virtue of its definition (5). Given X, we call a vector Y with components $Y_i \in \mathbb{R}$ *discretely convex relatively 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 that are discretely convex relatively to X. Note also that for any discretely convex Y, $g_{(X,Y)}(x) \geq \min_i Y_i$ for all x.

Theorem 1. *For every convex function h on $\mathbb{R}^d$, there is $g \in \mathcal{G}(X)$ such that $\Psi(g) \leq \Psi(h)$, with the strict inequality holding whenever $h \notin \mathcal{G}(X)$ and $\mathcal{H}(X)$ has nonempty interior.*

As promised, the theorem demonstrates that all solutions of (4) can be found among functions from $\mathcal{C}(X)$, by showing that they in fact belong to its subset $\mathcal{G}(X)$. Note that the optimization domain of the original maximum likelihood formulation (2) is a subset of that of (3), the instance of (4) for $\psi(g) = \mathrm{e}^{-g}$; therefore the conclusion of Theorem 1 is valid also for (2).

Moreover, Theorem 1 shows that due to the one-one correspondence between $\mathcal{G}(X)$ and $\mathcal{D}(X)$, the problem (4) is essentially finite-dimensional, parameterized by the values $Y_i = g(X_i)$. This facilitates the existence proof. The collection X of data points $X_1, \ldots, X_n$ from $\mathbb{R}^d$ will be considered in *general position* if every subset with $d+1$ elements forms a non-degenerate d-dimensional simplex.

Theorem 2. *Suppose that the collection X of $n > d+1$ data points $X_1, \ldots, X_n$ is in general position. Let the domain of ψ be the interval with endpoints ω and $+\infty$. If*

(i) $\omega = -\infty$ and ψ is of Type A, or if (ii) $\omega > -\infty$, then the primal formulation (4) *has a solution $g \in \mathcal{C}(X)$; in the case (ii), $g(x) \geq \omega$ for all x. If ψ is strictly convex, then this solution is unique.*

The assumption of general position serves mainly to contain the combinatorial character of the proof within reasonable limits; we do not believe that minor violations of the general position requirement are really capable of endangering the existence of the solution. Of course, a problem arises if, say, all data points in the plane lie on one line. Recall that in a probabilistic model with X_i independently sampled from a distribution absolutely continuous with respect to the Lebesgue measure, the collection X is in general position with probability 1.

The statement and proof of Theorem 2 does not exclude the possibility that the solution g is equal to ω for less than $d+1$ points X_i. Since this means that $\psi(x) = +\infty$, it is clear that it cannot happen for $d + 1$ or more points, because then $\psi(g)$ will be infinite over a nondegenerate simplex and then its integral, and consequently the objective function will be infinite. Nevertheless, Theorem 2 does not exclude the possibility that $\psi(g)$ is infinite at some points, as long as the integral is finite. However, for the instances considered below this possibility is usually ruled out by further bounds on the maximum of the density estimate.

5. Duality, Entropy and Divergences

The conjugate of the function ψ, defined as

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

is itself convex, being a sup of affine functions. Since ψ is decreasing, there are no affine functions with the positive slope minorizing the graph of ψ, hence $\psi^*(y) = +\infty$ for all $y > 0$. Finally, if ψ is differentiable on the (nonempty) interior of its domain, then ψ^* is the Legendre transformation of ψ,

$$\psi^*(y) = y\psi'^{-1}(y) - \psi(\psi'^{-1}(y)).$$

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 cone dual to $\mathcal{K}$ is

$$\mathcal{K}^* = \{G \in \mathcal{C}^*(X) \mid \int g \, dG \geq 0 \text{ for all } g \in \mathcal{K}\}.$$

Theorem 3. *The strong (Fenchel) dual of the primal formulation* (4) *is*

$$-\int \psi^*(-f)\, dy = \max_f ! \qquad \textit{subject to } f = \frac{d(P_n - G)}{dy}, \quad G \in \mathcal{K}^*, \tag{6}$$

in the sense that all $\Psi(g)$, the values of the primal objective for g satisfying the constraints of (4)*, dominate those of* (6)*, and both problems have optimal solutions, g*

and f respectively, for which the respective values of the primal and dual objective functions coincide. Any function f satisfying the constraints of (6) *is a probability density with respect to the Lebesgue measure: $f \geq 0$ and $\int f\,dx = 1$. If ψ is differentiable on the interior of its domain, then the dual and primal optimal solutions satisfy the relationship $f = -\psi'(g)$.*

It should be emphasized that the expression of absolute continuity in the (6) is a requirement of $F = P_n - G$; the dual objective function is defined as the conjugate to the primal objective function of (4), 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. The nonnegativity of f is a consequence of $\psi^*(-y)$ being infinite for $y < 0$. In practical algorithmic settings it may be prudent to enforce $f \geq 0$ in the dual explicitly, as a feasibility constraint.

An immediate consequence of Theorem 3 is that we can reformulate the maximum likelihood problem posed in (3) as an equivalent maximum (Shannon) entropy problem.

Corollary 1. *Maximum likelihood estimation of a log-concave density as posed in* (3) *has an equivalent dual formulation,*

$$-\int f \log f\,dy = \max_f !\qquad \textit{subject to } f = \frac{d(P_n - G)}{dy}, \quad G \in \mathcal{K}^*, \tag{7}$$

whose solution satisfies the relationship $f = \mathrm{e}^{-g}$, where g is the solution of (3)*. In particular, the solution of* (3) *satisfies $\int \mathrm{e}^{-g} = 1$, therefore problems* (2) *and* (3) *are equivalent.*

The emergence of the Shannon entropy here 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 picking f closest in the Kullback-Leibler divergence to the uniform distribution on $\mathcal{H}(X)$, from all f satisfying the dual constraints.

While the equivalence of (2) and (3) could be much less expediently shown by elementary means, it is important to emphasize that the real value of the dual connection is in the vista of new possibilities it opens. To explore the link to potential alternatives, we considered the family of entropies originally introduced for $\alpha > 0$ by Rényi (1961, 1965),

$$(1-\alpha)^{-1} \log\Big(\int f^\alpha(x)dx\Big), \qquad \alpha \neq 1, \tag{8}$$

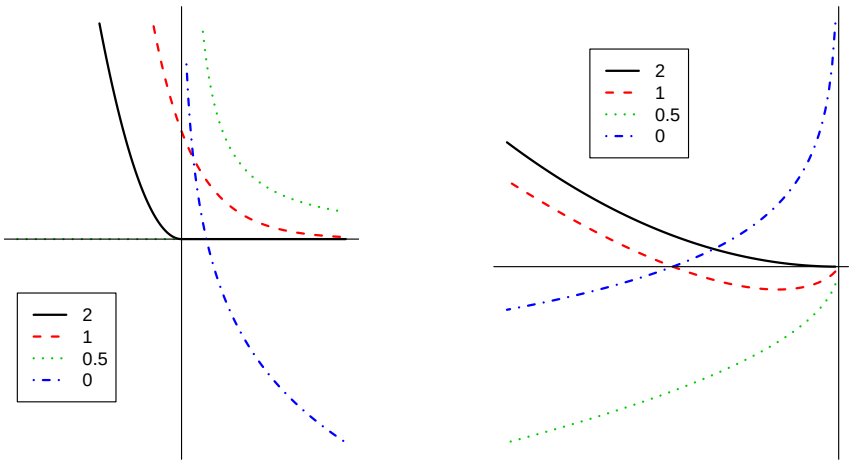

FIGURE 1. Primal ψ (left) and dual ψ^* (right) for selected $\alpha \geq 0$ from the Rényi family of entropies.

as an extension of the limiting case for $\alpha = 1$, the Shannon entropy. For $\alpha \neq 1$, maximizing (8) is equivalent to the maximization of

$$\frac{\operatorname{sgn}(1-\alpha)}{\alpha} \int f^\alpha = -\operatorname{sgn}(\alpha - 1) \int \frac{f^\alpha}{\alpha}. \tag{9}$$

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

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

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

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

where β and α are conjugates in the usual sense that $1/\beta + 1/\alpha = 1$; see Figure 1. All these ψ function satisfy the assumptions made in Section 4: they are convex and decreasing, their domain contains $(0, +\infty)$, and the boundary limit behavior requirement is also satisfied. Moreover, they are all strictly convex and differentiable on their domains. Finally, for $\alpha > 0$ they are all bounded from below by 0, so there

is no problem with the Lebesgue existence of the integral in the primal formulation, whose general form for $\alpha \neq 1$ corresponding to (8) can be written in a unified way as

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{1}{|\beta|}\int |g|^{\beta}\,dx = \min_{g\in\mathcal{C}(X)}\,! \tag{10}$$

as well as the relation between the dual and primal solutions, $f = |g|^{\beta-1}$.

Several particular instances merit special attention.

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

$$-\frac{1}{n}\sum_{i=1}^{n} h(X_i) + \frac{1}{\beta}\int h^{\beta}\,dx = \min_{g\in\mathcal{C}(X)}\,! \qquad \text{subject to } g \in \mathcal{K}. \tag{11}$$

After substituting $f^{1/(\beta-1)}$ for h into the objective function of (11), multiplying by β, and rewriting in terms of α we obtain

$$-\left(\frac{\alpha}{\alpha-1}\right)\frac{1}{n}\sum_{i=1}^{n} f^{\alpha-1}(X_i) + \int f^{\alpha}\,dx. \tag{12}$$

This form of the objective function recalls the "minimum density power divergence estimators", proposed, for $\alpha \geq 1$, by Basu, Harris, Hjort, and Jones (1998) in the context of estimation in parametric families.

5.2. **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

$$-\int f^2(y)dy = \max_{f}!\qquad \text{subject to } f = \frac{d(P_n - G)}{dy},\quad G \in \mathcal{K}^*. \tag{13}$$

The primal formulation can be written in a particularly simple form,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \frac{1}{2}\int g^2\,dx = \min_{g\in\mathcal{C}(X)}\,! \qquad \text{subject to } g \in \mathcal{K}, \tag{14}$$

which can be interpreted as a variant of the minimum Pearson χ^2 criterion; a similar motive can be traced also 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 (14) by $-f$ and appropriately modifying the cone constraint gives a variant of the "least-squares estimator", proposed by Groeneboom, Jongbloed, and Wellner (2001) to estimate a *convex* (and decreasing) density.

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 $\alpha \leq 1$. Apart from the celebrated log-concave case $\alpha = 1$, a promising alternative would seem to be Rényi entropy with $\alpha = 1/2$.

5.3. **Hellinger.** The choice of $\alpha = 1/2$ in the Rényi system leads to the dual

$$\int \sqrt{f} dy = \max_f ! \qquad \text{subject to } f = \frac{d(P_n - G)}{dy}, \quad G \in \mathcal{K}^*, \tag{15}$$

and primal

$$\frac{1}{n}\sum_{i=1}^n g(X_i) + \int \frac{1}{g}\, dx = \min_{g \in \mathcal{C}(X)} ! \qquad \text{subject to } g \in \mathcal{K}. \tag{16}$$

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 $\nu \geq 1$ satisfy this requirement. The dual problem (15) 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 $\mathcal{H}(X)$.

5.4. **New Frontiers.** 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 (see Figure 1)

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

The resulting dual, adapted directly from (6), is

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

and the primal becomes

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

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 (7), 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

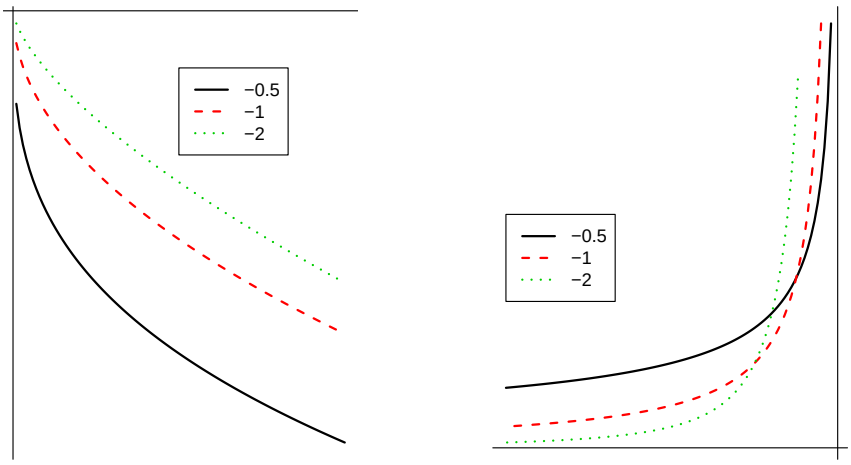

FIGURE 2. Primal ψ (left) and dual ψ^* (right) for selected $\alpha < 0$.

divergence. In parametric estimation the latter objective is sometimes associated with empirical likelihood, while the former is associated with exponential tilted empirical likelihood. See, for example, Shennach (2007).

One could continue in this fashion marching inexorably toward weaker and weaker concavity requirements. There appears to be no obstacle in considering also $\alpha < 0$; the general form (10) of the primal is still applicable. The shape constraints corresponding to negative α encompass a wider and wide class of quasi-concave densities, eventually arriving at the -∞-concave constraint, at which point we would have sanctioned all of the quasi-concave densities. However, a formal complication, demanding some finer treatment in the spirit of singular integrals of calculus of variations, arises for $\alpha \leq 0$ with the Lebesgue existence of the integrals involved in the primal formulations, because ψ is no longer bounded from below; see Figure 2. In our experience numerical solution for $\alpha \leq 0$ also becomes more fragile. As a practical matter, it seems therefore prudent to focus our attention to $\alpha = 1$ and $1/2$, the log-concavity and the (Hellinger) -1/2-concave cases. These estimators cover (respectively) the exponential and algebraic tail behaviors—those that are generally considered to be plausible in applications—and are tractable from both theoretical and numerical viewpoints.

6. FISHER CONSISTENCY OF THE HELLINGER ESTIMATOR

Consistency of these estimators in dimension one for selected α has been addressed by several authors: Pal, Woodroofe, and Meyer (2006) prove consistency in the

Hellinger metric for the log-concave ($\alpha = 1$) case, Groeneboom, Jongbloed, and Wellner (2001) establish consistency and rates of convergence for $\alpha \in \{1, 2\}$ in the uniform metric. The proofs of these results are highly technical and rely strongly on features of the univariate setting. However, it is plausible that analogous results hold for other α and dimensions greater than one. While such extensions of the theory remain a subject of on-going research, it can be noted that the crucial prerequisite, Fisher consistency, can be verified in a quite straightforward manner. Consider, for example, the (Hellinger) case of $\alpha = 1/2$ with primal objective function (16),

$$\int \frac{1}{\sqrt{f}}\, dP_n + \int \sqrt{f}\, dx. \tag{17}$$

Replacing dP_n by $f_0\, dx$, where f_0 is the unknown target density yields,

$$2 \int \sqrt{f_0}\, dx \leq \int \frac{f_0}{\sqrt{f}}\, dx + \int \sqrt{f}\, dx, \tag{18}$$

which follows from the inequality,

$$\sqrt{f_0 f} \leq \frac{f_0 + f}{2}.$$

This establishes that the population version of the (primal) objective function is uniquely minimized at f_0. Feasibility, of course, requires that f_0 be $-1/2$-concave.

7. Computational Methods and Experience

To illustrate the application of the foregoing methods we briefly consider some examples. Figure 3 illustrates a maximum likelihood estimate of a Gamma density on 3 degrees of freedom based on a sample of 200 observations. The sample is represented in the upper panel by the vertical bars along the horizontal axis—a so-called rug plot. The estimated density has the characteristic piecewise exponential appearance implied by the piecewise linear form of the estimated log-density shown in the lower panel of the Figure. In each panel the target Gamma function is shown in grey.

Details regarding the numerical implementation of our methods appear in Appendix B. We have 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. 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. Rufibach (2007) has implemented an active set algorithm for the univariate case, and Cule, Samworth, and Stewart (2007) have recently implemented a promising steepest descent

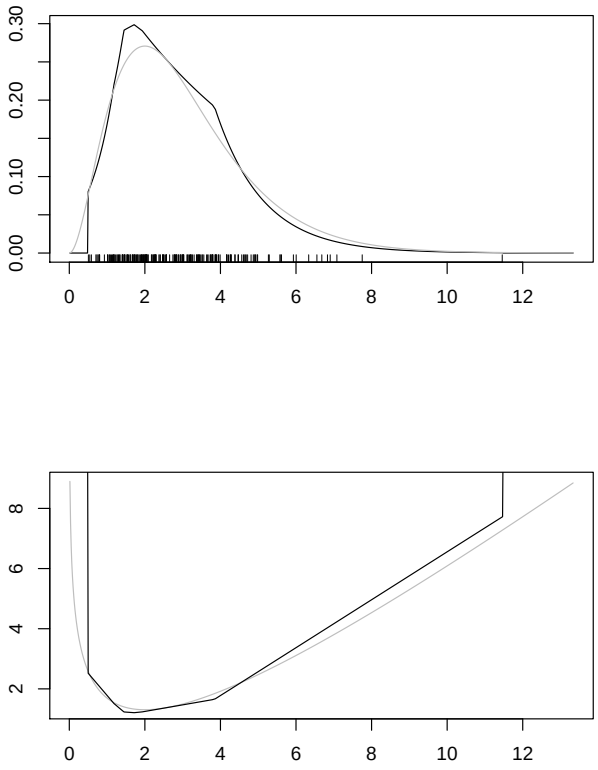

FIGURE 3. Log-concave Maximum Likelihood Estimator of a Gamma Density

algorithm for multivariate log-concave estimation that may be adaptable to other quasi-concave density estimation problems.

In Figure 4 we illustrate two estimates of a lognormal density, again log densities are plotted in the lower panel. As is apparent from the gray curve representing the true log density, the lognormal is *not* log-concave so the piecewise exponential log-concave estimate has some difficulties. However, the lognormal is $1/\sqrt{f}$ concave so the Hellinger estimate depicted as the dashed line is able to cope somewhat better with this case.

In the foregoing examples we have used 800 equally spaced "undata" values, v_i, in the interval $[0, 12]$; this is relatively modest, but sufficient to make the Riemann approximation for the integrability constraint very accurate. Note that the estimated density has discontinuities at the extreme order statistics; this is most apparent in the lower panel. These discontinuities may appear to be somewhat unfortunate at

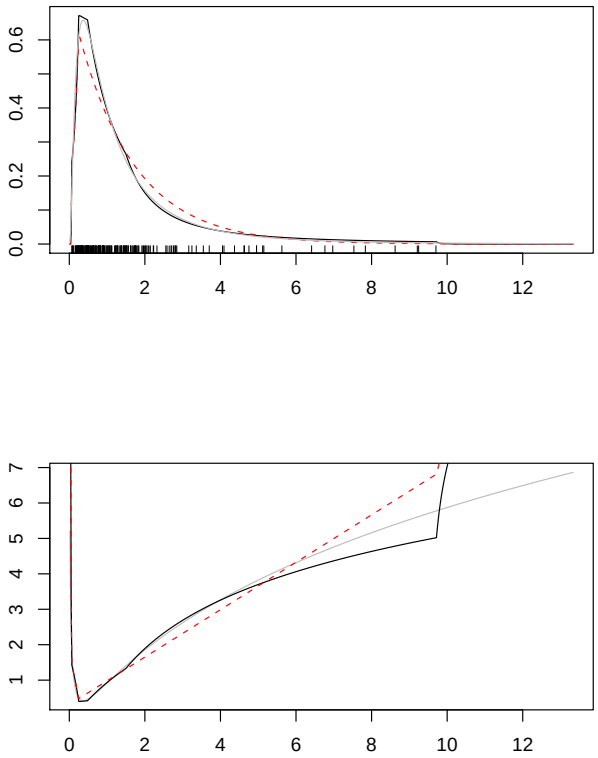

FIGURE 4. Two Estimates of a Lognormal Density

first glance, but they must be considered a "feature" not a "bug" in view of Lemma 1. Conventional kernel density estimation, it may be worth remarking, has the opposite problem—spreading mass into the negative half-line where it doesn't belong.

To illustrate our approach in a simple bivariate setting we reconsider the well-known MacDonell (1902) data on the heights and left middle finger lengths of 3000 British criminals. This data was employed by W.S. Gosset in preliminary simulation work described in "Student" (1908).

Figure 5 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 anomalous, at least for any density with exponential tail behavior. The maximum likelihood estimate of the log-concave model in Figure 6 has very similar central contours, but the outer contours fall off much more

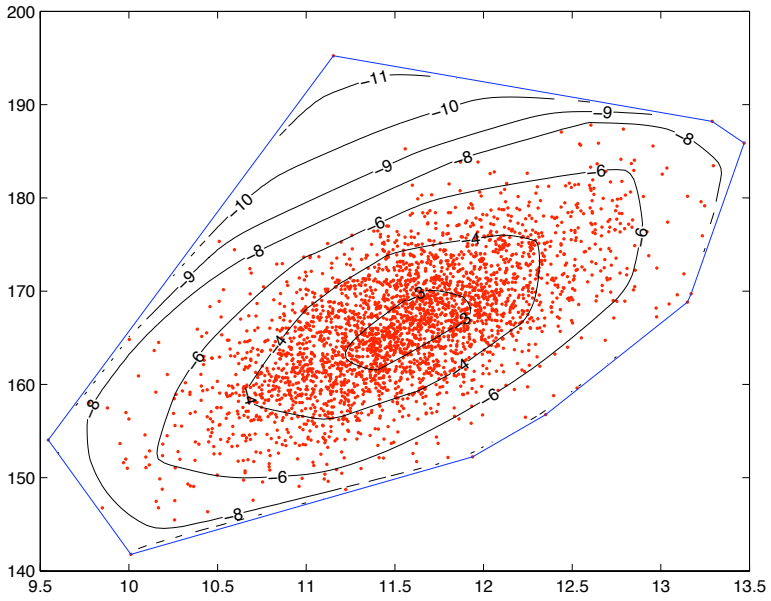

FIGURE 5. Hellinger (-1/2-concave) Estimate of Student's Criminals Density: Contours are labeled in units of log-density.

rapidly implying that the log-concave estimate assigns much smaller probability to the region near the unusual point.

8. 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 Renyi entropy family.

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 (2007) 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 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.

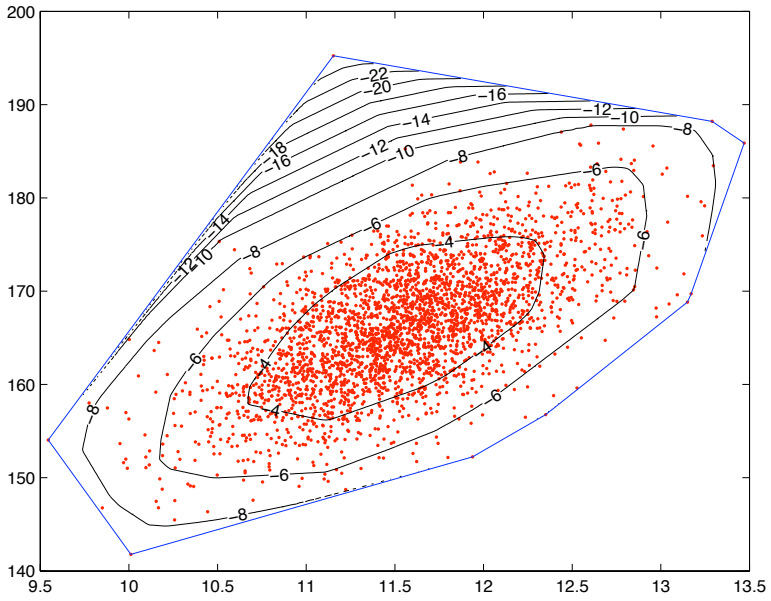

FIGURE 6. Log-concave Estimate of Student's Criminals Density: Contours are labeled in units of log-density.

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 would 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 1. Given h convex, put $Y_i = h(X_i)$ and take $g = g_{(X,Y)}$, the function defined by (5). The convexity of h implies that $h(x) \leq g(x)$ for every x; since ψ is nonincreasing, we have

$$c = \int \psi(h)\,dx - \int \psi(g)\,dx \geq 0. \tag{19}$$

The definition of $g_{(X,Y)}$ implies that $h(X_i) = g_{(X,Y)}(X_i) = Y_i$; therefore the rest of Ψ remains unchanged, and (19) implies that $\Psi(g) \leq \Psi(h)$.

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 polyhedrality of $\mathcal{H}(X)$ implies through Theorem 10.2 of Rockafellar (1970) that h is upper semicontinuous relative to $\mathcal{H}(X)$ at any $x \in \operatorname{conv} X$; 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 $\Psi(g) < \Psi(h)$ follows. ∎

Lemma 1. *Let $U_1, \ldots, U_{d+1}$ be $d+1$ points in $\mathbb{R}^d$ forming the vertices of a nondegenerate simplex S. Suppose that ψ is a nonincreasing, convex function. If g is a convex function such that $g(U_i) = Z_i$, then*

$$\int \psi(g)\,dx \geq \int_S \psi(g)\,dx \geq V\psi(\bar{Z}), \tag{20}$$

where V is the volume of S and $\bar{Z} = (Z_1 + \cdots + Z_{d+1})/(d+1)$.

Proof. By the convexity of g, the affine function h such that $h(U_i) = Z_i$ majorizes g on S. Using the formula for the integral of an affine function over a simplex and the fact that ψ is nonincreasing, we obtain

$$V\psi(\bar{Z}) = V\psi\left(\frac{1}{V}\int_S h\,dx\right) \leq V\psi\left(\frac{1}{V}\int_S g\,dx\right) \leq \int_S \psi(g)\,dx,$$

by the Jensen inequality applied with respect to the uniform distribution on S. ∎

Proof of Theorem 2. As shown by Theorem 1, problem (4) can be parametrized by Y_i, the putative values of $g(X_i)$, because there is one-one correspondence between $\mathcal{D}(X)$, the set of discretely convex vectors relatively to X, and the set of functions $\mathcal{G}(X)$. However, this correspondence is not a linear mapping (except when $d = 1$), so we have to show first that (4) remains a convex problem when parametrized by the elements of $\mathcal{D}(X)$, or, in a redundant way, by the elements of $\mathbb{R}^d$. For any $\alpha, \beta \geq 0$

such that $\alpha + \beta = 1$,

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

Since ψ is nonincreasing and convex, we obtain

$$\begin{aligned} \int \psi\big(g_{(X,\alpha Y+\beta Z)}(x)\big)\,dx &\leq \int \psi\big(\alpha g_{(X,Y)}(x) + \beta g_{(X,Z)}(x)\big)\,dx \\ &\leq \int \alpha\psi(g_{(X,Y)}(x)) + \beta\psi(g_{(X,Z)}(x))\,dx \\ &= \alpha \int \psi(g_{(X,Y)}(x))\,dx + \beta \int \psi(g_{(X,Z)}(x))\,dx, \end{aligned}$$

which was required to establish that the function $Y \mapsto \int \psi(g_{(X,Y)}(x))\,dx$ is convex; it is also finite for every Y with 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. Subsequently,

$$\begin{aligned} \Phi(Y) = \Psi(g_{(X,Y)}) &= \frac{1}{n}\sum_{i=1}^n g_{(X,Y)}(X_i) + \int \psi(g_{(X,Y)})\,dx \\ &= \frac{1}{n}\sum_{i=1}^n Y_i + \int \psi(g_{(X,Y)})\,dx \end{aligned}$$

is a (proper) convex function. The convexity constraint in (4) translates as the condition requiring Y to lie in $\mathcal{D}(X)$, a convex subset of $\mathbb{R}^d$; indeed, if there are convex functions g, h such that $Y_i = g(X_i)$ and $Z_i = h(X_i)$, then $\alpha Y_i + \beta Z_i = \alpha g(X_i) + \beta h(X_i)$ and $\alpha g + \beta h$ is a convex function.

In fact, the convexity of the problem is needed only to establish the continuity of the objective function; this in turn implies the existence of the solution, through the existence of a bounded sublevel set. For the existence proof, we use the modified function $\Phi_{\mathcal{D}}$ equal to Φ for every $Y \in \mathcal{D}(X)$, and otherwise equal to $+\infty$. To show the existence of Y minimizing $\Phi_{\mathcal{D}}$, it is sufficient to show the existence of a bounded sublevel set $\{Y \mid \Phi_{\mathcal{D}}(Y) \leq c\}$. This is implied by coercivity, the property ensuring that $\Phi_{\mathcal{D}}(Y) \to +\infty$ whenever Y is approaching the infinity—there is a component Y_i^k of Y with a limit point equal to $-\infty$ or $+\infty$.

We will do the proof first for the case (i), the case with $\omega = -\infty$, and then describe the adaptation for finite ω. Thus, suppose that Y^k is a sequence approaching infinity. If only a finite number of Y^k are in $\mathcal{D}(X)$, then the limit of $\Phi_{\mathcal{D}}(Y)$ is $+\infty$ and there is nothing to prove. Otherwise, we can assume, passing to a subsequence if necessary, that $Y^k \in \mathcal{D}(X)$ for every k, and each componentwise sequence Y_i^k either converges to ω, or to $+\infty$, or to a limit within $(\omega, +\infty)$. Let I be the set of all i such that $Y_i^k \to \omega$; there is $C > \omega$ such that $Y_i^k \geq C$ for all k and $i \notin I$. There are three possibilities.

(I) The set I is empty. Since ψ is nonincreasing,

$$\int \psi(g_{(X,Y^k)}(x))\,dx \leq \int \psi(\min_i Y_i^k)\,dx \leq V\psi(C),$$

where V is the volume of $\mathcal{H}(X)$. There must be at least one t such that $Y_i^k \to \infty$; the summation term in $\Phi_{\mathcal{D}}$ then makes $\Phi_{\mathcal{D}}(Y^k) \to +\infty$.

(II) The set I is nonempty, with cardinality $m < d+1$. We form a simplex S whose set of vertices, J, consists of all m points of I, together with $d+1-m$ other points. There are two possibilities now:

(IIa) If $\bar{Y}_J^k = \sum_{i\in J} Y_i^k/(d+1) \to \omega$, then we apply Lemma 1 to show that

$$\begin{aligned}
\Phi_{\mathcal{D}}(Y^k) &= \frac{1}{n}\sum_{i=1}^n Y_i^k + \int \psi(g_{(X,Y^k)})\,dx \\
&= \frac{1}{n}\sum_{i\in J} Y_i^k + \frac{1}{n}\sum_{i\notin J} Y_i^k + \int \psi(g_{(X,Y^k)})\,dx \\
&\geq \frac{1}{n}\sum_{i\in J} Y_i^k + \frac{C(n-d-1)}{n} + \int_S \psi(g_{(X,Y^k)})\,dx \\
&\geq \frac{d+1}{n}\bar{Y}_J^k + \frac{C(n-d-1)}{n} + V\psi(\bar{Y}_J^k) \to +\infty
\end{aligned}$$

where V is now the volume of S. The convergence follows from the assumption on ψ guaranteeing the convergence of $\psi(x) - cx$ to $+\infty$ when $x \to \omega$, for any $c \geq 0$.

(IIb) If $\bar{Y}_J^k$ for our selection of the additional vertices of S stays bounded from below from ω, then this can happen only when $Y_{i_0}^k \to +\infty$ for some $i_0 \in J \setminus I$. (Here we are using the assumption that $\omega = -\infty$.) We drop the index i_0 from J then, and add some other vertex—some other index to J. If then $\bar{Y}_J^k \to \omega$, we can proceed like in (IIa). Otherwise, $\bar{Y}_J^k$, for new J, stays bounded from below by $B > \omega$; we take the index i_0 which we omitted, and observe that

$$\begin{aligned}
\Phi_{\mathcal{D}}(Y^k) &= \frac{1}{n}\sum_{i\in J} Y_i^k + \frac{1}{n}Y_{i_0}^k + \frac{1}{n}\sum_{i\notin J\cup\{i_0\}} Y_i^k + \int \psi(g_{(X,Y)})\,dx \\
&\geq B + \frac{1}{n}Y_{i_0}^k + \int \psi(g_{(X,Y^k)})\,dx \geq B + \frac{1}{n}Y_{i_0}^k \to +\infty.
\end{aligned}$$

using the fact that ψ is of Type A and hence bounded from below by zero. Note that it is this alternative that uses the assumption (i).

(III) The cardinality of I is $m \geq d+1$. The discrete convexity of Y^k implies the following. Whenever S is a simplex whose vertices X_i satisfy $Y_i^k \to \omega$, that is, $i \in I$, then any other X_i from this simplex satisfies the same: $Y_i^k \to \omega$ and $i \in I$. Therefore, the convex hull of the set X_I consisting of all X_i with $i \in I$ does not contain any point X_i with $i \notin I$. We can form a triangulation of X_I now: a collection of d-dimensional simplexes with vertices in X_I, whose intersections are lower-dimensional simplices with vertices in X_I, and whose union is the convex hull of X_I. We denote simplices of this triangulation by S_j, their volumes by V_j, and the averages of Y_i corresponding to their vertices X_i by $\bar{Y}_J$. Let K_i be, for any $X_i \in X_I$, the number of simplices S_j into which X_i belongs; apparently, $K_i \geq 1$, and for $X_i \notin X_I$ we set $K_i = 0$); let $K = \max_i K_i$. Since $Y_i^k \to \omega$ for all $i \in I$, we may assume that $Y_i^k \leq A$ for some $A > \omega$, passing to a subsequence if necessary. We obtain

$$\begin{aligned}
\Phi_{\mathcal{D}}(Y^k) &= \frac{1}{n}\sum_{i=1}^{n} Y_i^k + \int \psi(g_{(X,Y^k)})\,dx \\
&= \frac{1}{n}\sum_{i\in I} Y_i^k + \frac{1}{n}\sum_{i\notin I} Y_i^k + \int \psi(g_{(X,Y^k)})\,dx \\
&\geq \frac{1}{n}\sum_{i\in I} K_i Y_i^k - \frac{1}{n}\sum_{i\in I}(K_i-1)A + \frac{C(n-m)}{n} + \sum_{j\in J}\int_{S_j} \psi(g_{(X,Y^k)})\,dx \\
&\geq \sum_{j\in J}\frac{d+1}{n}\bar{Y}_j^k - \frac{mKA}{n} + \frac{C(n-m)}{n} + \sum_{j\in J} V_j \psi(\bar{Y}_j^k) \to +\infty,
\end{aligned}$$

using the same argument for the convergence as in the case (IIa).

All the arguments above go through also for $\omega > -\infty$, except for the case (IIb). However, in this case we only need (I) to prove the existence of a bounded sublevel set; as there is no problem with (I), we can prove it under the assumption (ii). The only anomaly is that some of the components of the solution Y may be equal to ω, which in turn may render $\psi(g(x))$ infinite for some x—if the domain of ψ is $(\omega, +\infty)$ and not $[\omega, +\infty)$. Nevertheless, the validity of (III) under (ii), as well as the fact that the uniform distribution on $\mathcal{H}(X)$ yields finite objective function of (4), imply that—even in such case—the integral of $\psi(g)$ must be finite. ∎

Proof of Theorem 3. We rewrite (4) in a form suitable for the application of Fenchel duality theorems,

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \int \psi(g)\,dx + \iota_{\mathcal{K}}(g) = \Phi(g) + \Upsilon(g) = \inf_g ! \tag{21}$$

where Φ is the objective function of (4) and $\Upsilon = \iota_{\mathcal{K}}$ is the indicator function of $\mathcal{K}$ (equal to 0 for any $g \in \mathcal{K}$ and to $+\infty$ otherwise). The Fenchel dual of (21) is then

$$-\Phi^*(G) - \Upsilon^*(-G) = \max_G ! \tag{22}$$

The conjugate of the indicator of a convex cone $\mathcal{K}$ is the indicator of $-\mathcal{K}^*$; see, for instance, Attouch, Buttazzo, and Michaille (2006), Example 9.1.1.(2). The term $-\Upsilon^*(-G)$ in the objective can be therefore interpreted as a constraint $-G \in -\mathcal{K}^*$, that is, $G \in \mathcal{K}^*$. From the definition of the conjugate of Φ, we obtain

$$\begin{aligned} \Phi^*(G) &= \sup_g \left(\langle G, g \rangle - \frac{1}{n} \sum_{i=1}^n g(X_i) - \int \psi(g)\, dx \right) \\ &= \sup_g \left(\langle G - P_n, g \rangle - \int \psi(g)\, dx \right) = \Psi^*(G - P_n) \end{aligned} \tag{23}$$

where the sup is taken over all g from

$$\operatorname{dom} \Psi = \operatorname{dom} \Phi = \{ g \in \mathcal{C}(X) \mid \int \psi(g)\, dx < +\infty \},$$

and Ψ^* is the conjugate of the functional

$$\Psi(g) = \int \psi(g)\, dx, \tag{24}$$

whose form is given by Corollary 4A of Rockafellar (1971): if G is absolutely continuous with respect to the Lebesgue measure, then

$$\Psi^*(G) = \int \psi^* \left(\frac{dG}{dx} \right) dx, \tag{25}$$

otherwise $\Psi^*(G) = +\infty$. All these facts and expressions (23) and (25) result in (6).

The standard constraint qualification in this context is that of Rockafellar (1966), Theorem 1: we need to find some g where both Φ and Υ are finite and one of them is continuous. A function constant on $\operatorname{conv} X$, say $g(x) = 1$ for all $x \in \operatorname{conv} X$, satisfies this requirement. 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.

Having verified the constraint qualification, 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 the dual objective is finite, then f is nonnegative. If $G \in \mathcal{K}^*$, then $\langle G, f \rangle \geq 0$ for every $f \in \mathcal{K}$ and

$$0 \leq \langle G, 1 \rangle = -\langle G, -1 \rangle \leq 0,$$

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

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

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

The existence of the optimal solution for the primal follows from Theorem 2. Once a solution exists, the primal and dual solutions g and f must satisfy the extremality (Karush-Kuhn-Tucker) conditions.

■

Appendix B. Computational Methods

Our generic primal problem (4) involves minimizing an objective function consisting of a linear component representing likelihood, or some generalized notion of fidelity, plus a non-linear component representing the integrability constraint subject to a cone constraint imposing concavity. In special cases, the integrability term $\int \psi(g(x))dx$ can be written explicitly, however more generally one needs a more pragmatic approach. We adopt a quite straightforward Riemann approximation that relies on the introduction of artificial data to control the accuracy of the approximation. We will briefly describe this technique, and then turn to the cone constraint.

B.1. **The Integrability Constraint.** Our objective is to approximate the integral,

$$\int_{\mathcal{H}(X)} \psi(g(x))dx \approx \sum_{i=1}^{m} \psi(g(x_i))v_i.$$

For $\psi(g) = e^g$, this is unnecessary since the integral can be evaluated explicitly, but this appears to be generally impractical. Of course, when the number of points m is modest the approximation may be poor, so we augment the observed data $X_1, \cdots, X_n$ with artificial *undata* points $u_1, \cdots, u_{m-n}$, which taken together form the points of evaluation $x_1, \cdots, x_m$. In dimension one the u_i just fill in the gaps between the order statistics, and the v_i's are simply the averages of the adjacent spacings between the ordered x_i's. In dimension two the u_i's add new vertices to the triangulation formed by the observed data, and the v_i depend upon the areas of the adjacent triangles. In principle, this can be extended to higher dimensions with areas replaced by volumes of adjacent tetrahedrons in dimension three, and so forth.

Denoting the vector of function values, $\gamma_i = g(x_i)$ for $i = 1, \cdots, m$, we can write the finite dimensional version of the primal problem as,

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

where $\Psi(\gamma)$ denotes the m-vector with typical element $\psi(g(x_i)) = \psi(\gamma_i)$, L is an "evaluation operator" that simply selects the data elements from γ, zeroing out the undata elements, and w is an n-vector of observation weights, typically $w_i \equiv 1/n$. In

this formulation, the additive form of the convex component of the objective function fits nicely into the existing frameworks for convex optimization; in particular the interior point algorithms of the PDCO algorithm of Saunders (2003), and the Mosek implementation of Andersen (2006) are designed for just such formulations.

B.2. **The Convexity Constraint.** Convexity of the transformed density is imposed by the convex cone constraint $D\gamma \geq 0$. In dimension one D is just a tridiagonal matrix that does second differencing, adapted to the possibly unequal spacing of the x_i's. In dimension two, when the fitted function, $\hat{g}$ is piecewise linear on a triangulation, $\mathcal{T}$, the convexity constraint requires that the gradient jump across every edge of the triangulation yields a convex contribution. For each edge this contribution is determined by the function values corresponding to the vertices of the quadralateral associated with the edge. Gradients on adjacent triangles are linear in their respective function values at the vertices, so again this yields linear inequality restrictions, one for each edge. Further details are spelled out in Koenker and Mizera (2004), Theorems 2.1 and 3.1. Sparsity of the resulting D matrix is crucial; only four non-zero elements in each of the rows of D need to be stored and manipulated by the optimization procedure. This dramatically reduces the computational burden and the memory requirements of the problem. Again, higher dimensional problems have a similar structure.

The foregoing description assumes that we have a fixed triangulation of the x_i's. Ideally, we would like to adapt the triangulation, $\mathcal{T}$, and perhaps even the position of the undata points to the problem at hand. Unfortunately, optimization with respect to these choices appears quite impractical at present. In most of our experiments we have either generated u_i's randomly, that is as a homogeneous Poisson process on the set $\mathcal{H}(X)$, or alternatively on a regular lattice, and in either case employed the Delone triangulation. This choice for $\mathcal{T}$ has some advantages from a numerical stability viewpoint, since it is known to maximize the minimum angle over all the triangles of $\mathcal{T}$, but further work on adaptive schemes for triangulation certainly merit future investigation.

An alternative that circumvents some of the ambiguity of the triangulation approach is to revert to a regular rectangular grid for the undata points, u_i. Again, the approximation error of the integral constraint is easily controlled, but the concavity constraint requires a new approach. Finite differencing in the bivariate case yields the Hessian estimates:

$$\begin{aligned}
H_{xx}(x,y) &= g(x+h,y) - 2g(x,y) + g(x-h,y) \\
H_{yy}(x,y) &= g(x,y+h) - 2g(x,y) + g(x,y-h) \\
H_{xy}(x,y) &= [g(x+h,y+h) - g(x+h,y-h) \\
&\qquad -g(x-h,y+h) + g(x-h,y-h)]/4 \\
H_{yx}(x,y) &= H_{xy}(x,y)
\end{aligned}$$

Concavity at each of the grid points can then be formulated as a semi-definite programming constraint, ensuring that these Hessians are negative definite at each grid point (x, y). However, in the bivariate setting such constraints can be reformulated as rotated quadratic cone constraints. The latter approach facilitates implementation in Mosek for the Hellinger case, requiring only that we constrain the signs of the diagonal elements of H and its determinant.

B.3. **Discrete Duality.** Associated with the primal problem (P) is the dual problem,

$$\{-s^\top \Psi^*(f) \mid Sf = -w^\top L + D^\top h, f \geq 0, D^\top h \geq 0\} = \max! \tag{D}$$

Here h is an m-vector of dual variables and f is an m-vector of function values representing the density evaluated at the v_i's, and $S = \mathrm{diag}(s)$. 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 1.

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

$$f(v_i) = \psi'(g(v_i)) \quad \textit{for } i = 1, \cdots, m, \tag{E}$$

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.

Proposition 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 f of (D) satisfy $s^\top f = 1$ and $f \geq 0$.*

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.

REFERENCES

AN, M. Y. (1998): "Log-concavity versus log-convexity: a complete characterization," *Journal of Economic Theory*, 80, 350–369.

ANDERSEN, E. D. (2006): "The MOSEK Optimization Tools Manual, Version 4.0," Available from `http://www.mosek.com`.

ARROW, K. J., AND A. C. ENTHOVEN (1961): "Quasi-Concave Programming," *Econometrica*, 29, 779–800.

ATTOUCH, H., G. BUTTAZZO, AND G. MICHAILLE (2006): *Variational analysis in Sobolev and BV spaces: applications to PDEs and optimization*, MPS-SIAM Series on Optimization 6. SIAM, Philadelphia.

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

BAGNOLI, M., AND T. BERGSTROM (2005): "Log-concave probability and its applications," *Economic Theory*, 26, 445–469.

BARLOW, R. E., AND F. PROSCHAN (1975): *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, New York.

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.

BLACK, D. (1948): "On the Rationale of Group Decision Making," *Journal of Political Economy*, 56, 23–34.

CAPLIN, A., AND B. NALEBUFF (1991a): "Aggregation and imperfect competition: on the existence of equilibrium," *Econometrica*, 59, 25–59.

——— (1991b): "Aggregation and social choice: a mean voter theorem," *Econometrica*, 59, 1–23.

CULE, M., R. SAMWORTH, AND M. STEWART (2007): "Computing the Maximum Likelihood Estimator of a Multidimensional Log-Concave Density," preprint.

EGGERMONT, P., AND V. LARICCIA (2001): *Maximum Penalized Likelihood Estimation*. Springer-Verlag.

FLINN, C. J., AND J. J. HECKMAN (1983): "Are unemployment and out of the labor force behaviorally distinct labor force states?," *Journal of Labor Economics*, 1, 28–42.

GOLDBERGER, A. S. (1983): "Abnormal Selection Bias," in *Studies in Econometrics, Time Series, and Multivariate Statistics*, ed. by S. Karlin, T. Amemiya, and L. A. Goodman, pp. 67–84. Academic Press.

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

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

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

GU, C. (2002): *Smoothing spline ANOVA models*. Springer-Verlag, New York.

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

HECKMAN, J. J., AND B. E. HONORÉ (1990): "The empirical content of the Roy model," *Econometrica*, 58, 1121–1149.

HILDRETH, C. (1954): "Point Estimates of Ordinates of Concave Functions," *J. of Am. Stat. Soc.*, 49, 598–619.

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

KOENKER, R., AND I. MIZERA (2004): "Penalized triograms: total variation regularization for bivariate smoothing," *J. R. Stat. Soc. (B)*, 66, 145–163.

——— (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. Witkowský.

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

Mizera, I. (2002): "On Depth and Deep Points: A Calculus," *The Annals of Statistics*, 30, 1681–1736.

Müller, S., and K. Rufibach (2007): "Smooth Tail Index Estimation," preprint.

Pal, J. K., M. Woodroofe, and M. Meyer (2006): "Estimating a Polya frequency function," in *A Festschrift in Memory of Yehuda Vardi*, IMS Lecture Notes-Monograph Series. Institute of Mathematical Statistics, to appear.

Rényi, A. (1961): "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Vol I*, pp. 547–561. University of California Press, Berkeley.

——— (1965): "On the foundations of information theory," *Revue d'Institut International de Statistique*, 33, 1–14.

Rockafellar, R. T. (1966): "Extensions of Fenchel's duality theorem for convex functions," *Duke Mathematical Journal*, 33, 81–89.

——— (1970): *Convex analysis*. Princeton University Press, Princeton.

——— (1971): "Integrals which are convex functionals, II," *Pacific Journal of Mathematics*, 39, 439–469.

Rufibach, K. (2007): "Computing Maximum Likelihood Estimators of a Log-Concave Density: Function," *J. Statistical Computation and Simulation*, 77, 561–574.

Rufibach, K., and L. Dümbgen (2006): "Maximum likelihood estimation of a log-concave density: basic properties and uniform consistency," Preprint.

Saunders, M. A. (2003): "PDCO: A Primal-Dual interior solver for convex optimization," `http://www.stanford.edu/group/SOL/software/pdco.html`.

Shennach, S. M. (2007): "Point estimation with exponentially tilted empirical likelihood," *Ann. Statist.*, 35, 634–672.

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.

Zeidler, E. (1985): *Nonlinear functional analysis and its applications III: Variational methods and optimization*. Springer-Verlag, New York.