Quasi-Concave Density Estimation

Roger Koenker

University of Illinois, Urbana-Champaign

"The Shape of Things to Come"
Joel's Blancmange Conference: 6 November 2010

Joint work with Ivan Mizera, University of Alberta

Regularization for Density Estimation

Maximum likelihood estimation of densities

$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(X_i)$$

over any (reasonably) large class $\mathcal{F}$ yields . . .

Regularization for Density Estimation

Maximum likelihood estimation of densities

$$\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(X_i)$$

over any (reasonably) large class $\mathcal{F}$ yields ...

Dirac Catastrophe: Cai Guo-Qiang's "Transient Rainbow" New York, 2002

Regularization – Remedies for Ill-Posedness

Two general classes of treatments:

- Norm Constraints: $\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(X_i) - \lambda \|D^k h(f)\|$
  - Good (1971) $\|D\sqrt{f}\|_2^2$
  - Silverman (1982) $\|D^3 \log(f)\|_2^2$
  - Wahba/Gu (2002) $\|D^2 \log(f)\|_2^2$
  - Davies/Kovac (2004) $TV(f) = \|Df\|_1$
  - Koenker/Mizera (2005) $TV(D \log f) = \|D^2 \log f\|_1$

Regularization – Remedies for Ill-Posedness

Two general classes of treatments:

- Norm Constraints: $\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log f(X_i) - \lambda \|D^k h(f)\|$
  - Good (1971) $\|D\sqrt{f}\|_2^2$
  - Silverman (1982) $\|D^3 \log(f)\|_2^2$
  - Wahba/Gu (2002) $\|D^2 \log(f)\|_2^2$
  - Davies/Kovac (2004) $TV(f) = \|Df\|_1$
  - Koenker/Mizera (2005) $TV(D \log f) = \|D^2 \log f\|_1$
- Shape Constraints: $\max_{f \in \mathcal{F}} \{\sum_{i=1}^{n} \log f(X_i) \mid D^k h(f) \in \mathcal{K}\}$
  - Grenander (1956) f monotone
  - Rufibach/Dümbgen (2006) $\log f$ concave

On Tautology: The New, Improved Histogram

The simplest example of a total variation penalized density estimator is the tautstring estimator of Hartigan and Hartigan (1985) elaborated by Davies and Kovac (2001, 2004) and van de Geer and Mammen (1997).

- Make a $\pm\epsilon$ Kolmogorov tube around the empirical df.

On Tautology: The New, Improved Histogram

The simplest example of a total variation penalized density estimator is the tautstring estimator of Hartigan and Hartigan (1985) elaborated by Davies and Kovac (2001, 2004) and van de Geer and Mammen (1997).

- Make a $\pm\epsilon$ Kolmogorov tube around the empirical df.
- Attach a loose string to the points $(X_{(1)}, 0)$ and $(X_{(n)}, 1)$.

On Tautology: The New, Improved Histogram

The simplest example of a total variation penalized density estimator is the tautstring estimator of Hartigan and Hartigan (1985) elaborated by Davies and Kovac (2001, 2004) and van de Geer and Mammen (1997).

- Make a $\pm\epsilon$ Kolmogorov tube around the empirical df.
- Attach a loose string to the points $(X_{(1)}, 0)$ and $(X_{(n)}, 1)$.
- Pull the string taut.

On Tautology: The New, Improved Histogram

The simplest example of a total variation penalized density estimator is the tautstring estimator of Hartigan and Hartigan (1985) elaborated by Davies and Kovac (2001, 2004) and van de Geer and Mammen (1997).

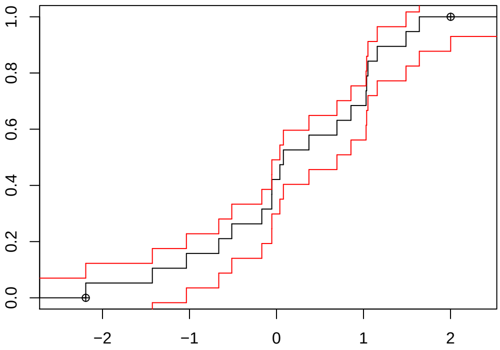
- Make a $\pm\epsilon$ Kolmogorov tube around the empirical df.
- Attach a loose string to the points $(X_{(1)}, 0)$ and $(X_{(n)}, 1)$.
- Pull the string taut.

This can be formalized as:

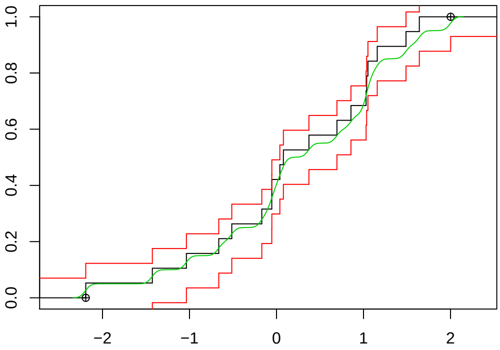
$$\hat{f} \equiv \hat{F}' = \text{argmin}_{F \in \mathcal{F}} \int (F_n(x) - F(x))^2 dF_n(x) + \lambda \text{TV}(F').$$

for some λ depending on ϵ.

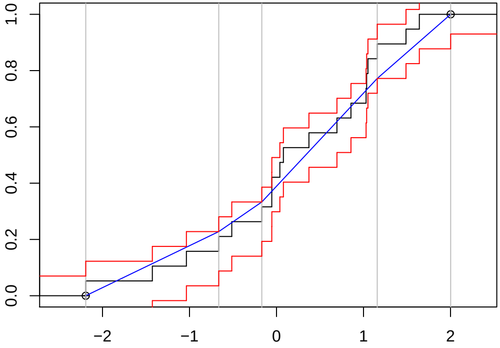
The Kolmogorov Tube



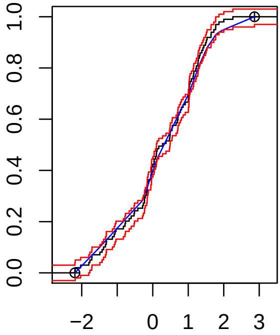
The Slack String

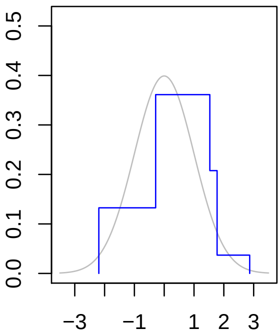


The Taut String

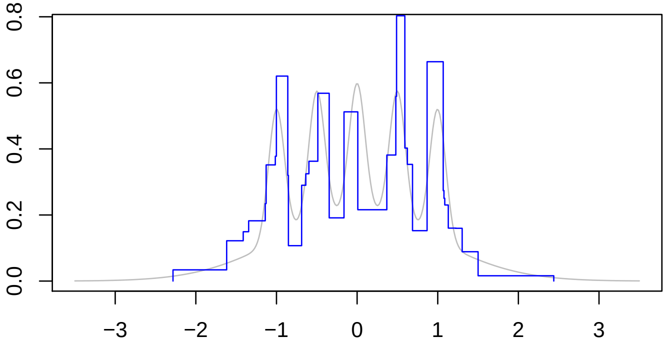


Taut String Densities are Piecewise Constant

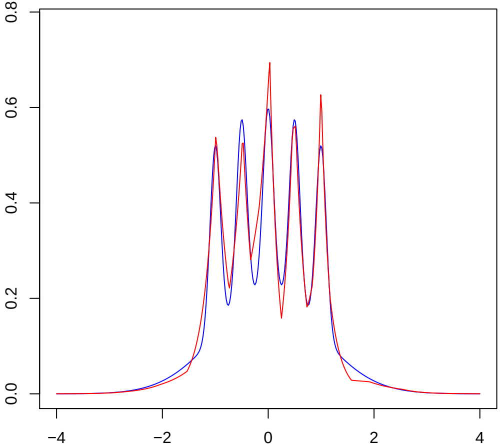




And Very Good at Estimating Modality



MLE's using TV Penalties on $(\log f)'$ Are Also Good



Shape Constrained Density Estimation: Early History

Grenander (1956) considered the maximum likelihood estimation of a monotone density:

$$\max\{\sum \log f(X_i) \mid f \searrow, \int f dx = 1\}$$

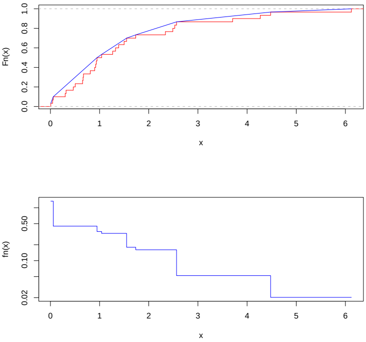
Solutions are piecewise constant functions with jumps at the observed $\{X_i\}$; they are derivatives of the least concave majorant, of the empirical distribution function, F_n.

Shape Constrained Density Estimation: Early History

Grenander (1956) considered the maximum likelihood estimation of a monotone density:

$$\max\{\sum \log f(X_i) \mid f \searrow, \int f dx = 1\}$$

Solutions are piecewise constant functions with jumps at the observed $\{X_i\}$; they are derivatives of the least concave majorant, of the empirical distribution function, F_n.



From Monotone to Unimodal Densities

If f is unimodal with a known mode then we can employ Grenander on each side of the mode to the same effect. Estimation of the mode can also be done so that the same rate is achievable with an estimated mode. Birgé (1997).

From Monotone to Unimodal Densities

If f is unimodal with a known mode then we can employ Grenander on each side of the mode to the same effect. Estimation of the mode can also be done so that the same rate is achievable with an estimated mode. Birgé (1997).

But unimodal densities aren't quite as appealing as they might at first appear. A more attractive class consists of strongly unimodal, or log-concave densities.

Definition A density $f : \mathbb{R}^d \to \mathbb{R}$ is log-concave if $g = -\log f$ is convex.

From Monotone to Unimodal Densities

If f is unimodal with a known mode then we can employ Grenander on each side of the mode to the same effect. Estimation of the mode can also be done so that the same rate is achievable with an estimated mode. Birgé (1997).

But unimodal densities aren't quite as appealing as they might at first appear. A more attractive class consists of strongly unimodal, or log-concave densities.

Definition A density $f : \mathbb{R}^d \to \mathbb{R}$ is log-concave if $g = -\log f$ is convex.

What's so great about log-concave densities?

Virtues of Log Concavity

- (Strong Unimodality) Convolutions of log-concave random variables are log concave. (Ibragimov (1956))

Virtues of Log Concavity

- (Strong Unimodality) Convolutions of log-concave random variables are log concave. (Ibragimov (1956))
- (Increasing Failure Rate) Hazard functions for log-concave random variables are increasing (Proschan (1965), Flinn and Heckman (1983))

Virtues of Log Concavity

- (Strong Unimodality) Convolutions of log-concave random variables are log concave. (Ibragimov (1956))
- (Increasing Failure Rate) Hazard functions for log-concave random variables are increasing (Proschan (1965), Flinn and Heckman (1983))
- (Monotone Likelihood Ratio) Log-concave densities have the MLR property for their location parameter:

$$f'(x-\theta)/f(x-\theta_0) \text{ is } \nearrow \text{ in } \theta.$$

and consequently the MLE (of location) is unique, and UMP tests exist ...

Virtues of Log Concavity

- (Strong Unimodality) Convolutions of log-concave random variables are log concave. (Ibragimov (1956))
- (Increasing Failure Rate) Hazard functions for log-concave random variables are increasing (Proschan (1965), Flinn and Heckman (1983))
- (Monotone Likelihood Ratio) Log-concave densities have the MLR property for their location parameter:

  $$f'(x - \theta)/f(x - \theta_0) \text{ is } \nearrow \text{ in } \theta.$$

  and consequently the MLE (of location) is unique, and UMP tests exist ...
- (Variation Diminishing Kernels) Kernel smoothing with log concave kernels insures that the number of modes of estimated density is decreasing in the bandwidth Silverman (1981) based on Karlin (1968).

Virtues of Log Concavity

- (Strong Unimodality) Convolutions of log-concave random variables are log concave. (Ibragimov (1956))
- (Increasing Failure Rate) Hazard functions for log-concave random variables are increasing (Proschan (1965), Flinn and Heckman (1983))
- (Monotone Likelihood Ratio) Log-concave densities have the MLR property for their location parameter:

$$f'(x-\theta)/f(x-\theta_0) \text{ is } \nearrow \text{ in } \theta.$$

  and consequently the MLE (of location) is unique, and UMP tests exist ...
- (Variation Diminishing Kernels) Kernel smoothing with log concave kernels insures that the number of modes of estimated density is decreasing in the bandwidth Silverman (1981) based on Karlin (1968).
- Many common densities are log concave: uniform, Gaussian, Laplacian, some Gammas, some Weibulls, ...

Virtues of Log Concavity

- (Strong Unimodality) Convolutions of log-concave random variables are log concave. (Ibragimov (1956))
- (Increasing Failure Rate) Hazard functions for log-concave random variables are increasing (Proschan (1965), Flinn and Heckman (1983))
- (Monotone Likelihood Ratio) Log-concave densities have the MLR property for their location parameter:

$$f'(x - \theta)/f(x - \theta_0) \text{ is } \nearrow \text{ in } \theta.$$

  and consequently the MLE (of location) is unique, and UMP tests exist ...
- (Variation Diminishing Kernels) Kernel smoothing with log concave kernels insures that the number of modes of estimated density is decreasing in the bandwidth Silverman (1981) based on Karlin (1968).
- Many common densities are log concave: uniform, Gaussian, Laplacian, some Gammas, some Weibulls, ...
- Numerous applications in virtually every corner of economic theory: search, signaling, reliability, auction design, pricing in differentiated product markets, and social choice all rely on log concavity conditions.

Beyond the Log Concave Horizon

Following Hardy, Littlewood and Polya (1934), recall that means of order ρ are defined as

$$M_\rho(a;p) = M_\rho(a_1, ..., a_n; p) = (\sum p_i a_i^\rho)^{1/\rho}$$

for p in the unit simplex, $\mathcal{S} = \{p \in \mathbb{R}^n_+ | \sum p_i = 1\}$.

Beyond the Log Concave Horizon

Following Hardy, Littlewood and Polya (1934), recall that means of order ρ are defined as

$$M_\rho(a; p) = M_\rho(a_1, ..., a_n; p) = (\sum p_i a_i^\rho)^{1/\rho}$$

for p in the unit simplex, $\mathcal{S} = \{p \in \mathbb{R}^n_+ | \sum p_i = 1\}$.

Examples: The classical means:

- $\rho = 1$ Arithmetic,
- $\rho = 0$ Geometric,
- $\rho = -1$ Harmonic.

Beyond the Log Concave Horizon

Definition (Avriel (1972)) A non-negative real function g defined on a convex set $C \subset \mathbb{R}^d$, is ρ-concave if for any $x_0, x \in C$ and $p \in \mathcal{S}$,

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

Note that

- concave functions are 1-concave,
- log-concave functions are 0-concave, ...
- σ-concaves are ρ-concave for all $\sigma > \rho$.
- $-\infty$-concaves are quasi-concave.

Moral: Some concaves are more concave than other concaves, but all are quasi-concave, that is they have convex level sets.

An Application to Voting and Social Choice

Caplin and Nalebuff (1992) consider a spatial model of voting in which agents have preferred positions in "issue space" according a ρ-concave density $f : \mathbb{R}^d \to \mathbb{R}$.

It is then demonstrated that the mean voter's preferred position is preferred by at least a proportion $1 - \delta$ of voters to any other proposed position, where

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

An Application to Voting and Social Choice

Caplin and Nalebuff (1992) consider a spatial model of voting in which agents have preferred positions in "issue space" according a ρ-concave density $f : \mathbb{R}^d \to \mathbb{R}$.

It is then demonstrated that the mean voter's preferred position is preferred by at least a proportion $1 - \delta$ of voters to any other proposed position, where

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

In the log-concave case, a simple computation then yields, for any d,

$$\delta(d, 0) = \lim_{\rho \to 0} \left(1 - \left[\frac{d + 1/\rho}{d + 1 + 1/\rho} \right]^{d+1/\rho} \right) = 1 - 1/e \approx .64.$$

An Application to Voting and Social Choice

Caplin and Nalebuff (1992) consider a spatial model of voting in which agents have preferred positions in "issue space" according a ρ-concave density $f : \mathbb{R}^d \to \mathbb{R}$.

It is then demonstrated that the mean voter's preferred position is preferred by at least a proportion $1 - \delta$ of voters to any other proposed position, where

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

In the log-concave case, a simple computation then yields, for any d,

$$\delta(d, 0) = \lim_{\rho \to 0} \left(1 - \left[\frac{d + 1/\rho}{d + 1 + 1/\rho} \right]^{d+1/\rho} \right) = 1 - 1/e \approx .64.$$

This generalizes the celebrated Black (1948) median voter result for (weakly) unimodal densities.

Nonparametric Maximum Likelihood

We can easily pose the problem:

$$\max_f \{\prod_{i=1}^{n} f(X_i) \mid f \text{ is a log-concave density}\}$$

$$(P) \qquad \min_g \{\sum_{i=1}^{n} g(X_i) \mid \int e^{-g(x)} dx = 1, \text{and } g \text{ is convex}\}$$

Nonparametric Maximum Likelihood

We can easily pose the problem:

$$\max_f \{\prod_{i=1}^n f(X_i) \mid f \text{ is a log-concave density}\}$$

$$(P) \qquad \min_g \{\sum_{i=1}^n g(X_i) \mid \int e^{-g(x)} dx = 1, \text{and } g \text{ is convex}\}$$

This is quite like the classical Grenander (1956) MLE for monotone densities. For $d = 1$ Rufibach (2007), and Pal, Woodroofe, and Meyer (2007) provide active set algorithms.

Nonparametric Maximum Likelihood

We can easily pose the problem:

$$\max_f \{ \prod_{i=1}^n f(X_i) \mid f \text{ is a log-concave density} \}$$

$$\text{(P)} \qquad \min_g \{ \sum_{i=1}^n g(X_i) \mid \int e^{-g(x)} dx = 1, \text{and } g \text{ is convex} \}$$

This is quite like the classical Grenander (1956) MLE for monotone densities. For $d = 1$ Rufibach (2007), and Pal, Woodroofe, and Meyer (2007) provide active set algorithms.

What about dimension $d > 1$? Koenker and Mizera (2010) suggest interior point methods, while Cule, Samworth and Stewart (2010) propose gradient methods.

A Characterization Lemma

Solutions to (P) are polyhedral convex functions of the form

$$\hat{g}(x) = \inf\left\{ \sum_{i=1}^{n} \lambda_i Y_i \mid x = \sum_{i=1}^{n} \lambda_i X_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geqslant 0 \right\},$$

where $\{X_i\}$ are the sample observations and the Y_i are freely varying, representing ordinates of the estimated density at the X_i's.

A Characterization Lemma

Solutions to (P) are polyhedral convex functions of the form

$$\hat{g}(x) = \inf\left\{\sum_{i=1}^{n}\lambda_i Y_i \mid x = \sum_{i=1}^{n}\lambda_i X_i, \sum_{i=1}^{n}\lambda_i = 1, \lambda_i \geqslant 0\right\},$$

where $\{X_i\}$ are the sample observations and the Y_i are freely varying, representing ordinates of the estimated density at the X_i's.

Implications:

- Reduces the problem to a finite, albeit n-dimensional, one.
- Solution log-densities are piecewise linear, i.e. polyhedral..
- Solution densities are piecewise exponential.
- Estimated densities vanish off the convex hull of the observations.

A Family of Convex Variational Problems

A functional version of our MLE problem (P) can be written as

$$\min_g \{ \int g dP_n + \int e^{-g} dx \mid g \in \mathcal{K} \}$$

where $\mathcal{K}$ denotes the cone of convex functions on $\mathcal{C}(X)$, the linear space of all bounded continuous functions on $\mathcal{H}(X)$, the convex hull of the $\{X_i\}$.

A Family of Convex Variational Problems

A functional version of our MLE problem (P) can be written as

$$\min_g \{ \int g dP_n + \int e^{-g} dx \mid g \in \mathcal{K} \}$$

where $\mathcal{K}$ denotes the cone of convex functions on $\mathcal{C}(X)$, the linear space of all bounded continuous functions on $\mathcal{H}(X)$, the convex hull of the $\{X_i\}$. It is useful to expand somewhat the class of these problems beyond the MLE log concave case, so we will rewrite this as,

$$\min_g \{ \int g dP_n + \int \psi(g) dx \mid g \in \mathcal{K} \}$$

Through the Looking Glass, Dually

Theorem Suppose that ψ is a decreasing convex function of a real variable with conjugate (Legendre transform) $\psi^*(y) = \sup_x\{yx - \psi(x)\}$, then the strong dual of the primal problem

$$(P) \qquad \min_g \{\int g dP_n + \int \psi(g) dx \mid g \in \mathcal{K}\}$$

is given by:

$$(D) \qquad \max_G \{-\int \psi^*(-f) dx \mid f = \frac{d(P_n - G)}{dx},\ G \in \mathcal{K}^*\}$$

where $\mathcal{K}^* = \{G \in \mathcal{C}^*(X) \mid \int g dG \geqslant 0 \text{ for all } g \in \mathcal{K}\}$, and $\mathcal{C}^*(X)$ is the space of signed Radon measures on $\mathcal{H}(K)$. Note that G must anihilate the atoms of P_n so that f is a density.

Dual Exhausts

Thus, for the original MLE log-concave example: $\psi(x) = e^{-x}$ we have $\psi^*(y) = -y \log(-y) + y$ giving the dual problem,

$$\max_f \{-\int f \log(f) dx \mid f = \frac{d(P_n - G)}{dx}, G \in \mathcal{K}^*\}$$

So the MLE problem becomes a maximum Shannon entropy problem.

Dual Exhausts

Thus, for the original MLE log-concave example: $\psi(x) = e^{-x}$ we have $\psi^*(y) = -y\log(-y) + y$ giving the dual problem,

$$\max_f \{-\int f \log(f) dx \mid f = \frac{d(P_n - G)}{dx}, G \in \mathcal{K}^*\}$$

So the MLE problem becomes a maximum Shannon entropy problem. Why Shannon? Why not some other (e.g. Renyi) entropy?

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

Dual Exhausts

Thus, for the original MLE log-concave example: $\psi(x) = e^{-x}$ we have $\psi^*(y) = -y\log(-y) + y$ giving the dual problem,

$$\max_f \{-\int f \log(f) dx \mid f = \frac{d(P_n - G)}{dx}, G \in \mathcal{K}^*\}$$

So the MLE problem becomes a maximum Shannon entropy problem. Why Shannon? Why not some other (e.g. Renyi) entropy?

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

The usual suspects (shades of Cressie-Read and Csiszár divergences):

- $\alpha = 1$ is Shannon (taking limits)
- $\alpha = 2$ is Pearson χ^2
- $\alpha = 1/2$ is Hellinger
- $\alpha = 0$ is (some form of) Empirical Likelihood

Don Juan in Hellinger

Our favorite alternative to Shannon is Renyi's $\alpha = 1/2$,

$$(D) \qquad \max_f \{-\int \sqrt{f}dx \mid f = \frac{d(P_n - G)}{dx}, G \in \mathcal{K}^*\}$$

Don Juan in Hellinger

Our favorite alternative to Shannon is Renyi's $\alpha = 1/2$,

$$(D) \qquad \max_f \{ -\int \sqrt{f} dx \mid f = \frac{d(P_n - G)}{dx}, G \in \mathcal{K}^* \}$$

$$(P) \qquad \min_g \{ \int g dP_n + \int g^{-1} dx \mid g \in \mathcal{K} \}$$

Here, $f = \psi'(g) = (g^{-1})' = -g^{-2}$, so $g = f^{-1/2}$ so the convexity constraint in (P) requires that $f^{-1/2}$ be concave.

Don Juan in Hellinger

Our favorite alternative to Shannon is Renyi's $\alpha = 1/2$,

$$(D) \qquad \max_f \{-\int \sqrt{f}dx \mid f = \frac{d(P_n - G)}{dx}, G \in \mathcal{K}^*\}$$

$$(P) \qquad \min_g \{\int g dP_n + \int g^{-1} dx \mid g \in \mathcal{K}\}$$

Here, $f = \psi'(g) = (g^{-1})' = -g^{-2}$, so $g = f^{-1/2}$ so the convexity constraint in (P) requires that $f^{-1/2}$ be concave.

- All Student's are admitted up to and including Cauchy.
- These are Avriel's ρ-concaves, with $\rho = \alpha - 1 = -1/2$.
- Recall that this class nests the log-concaves.

Algorithms and Actuality

Discrete implementations require two basic ingredients:

- Data: $\{X_1, \cdots, X_n\}$
- Undata: $\{v_1, \cdots, v_n\}$

Algorithms and Actuality

Discrete implementations require two basic ingredients:

- Data: $\{X_1, \cdots, X_n\}$
- Undata: $\{v_1, \cdots, v_n\}$

We parameterize $g = (g(v_i))_{i=1}^m \equiv (\gamma_i)_{i=1}^m$, thus:

- $\int \psi(g)dx \approx \sum s_i \psi(g(v_i)) \equiv s^\top \Psi(\gamma)$ Riemann Sum
- $\int g dP_n = \sum g(X_i) = w^\top L\gamma$ Linear Interpolation
- $g \in \mathcal{K} \Leftrightarrow D\gamma \geqslant 0 \quad D = \nabla^2$ Convex Cone Constraint

Algorithms and Actuality

Discrete implementations require two basic ingredients:

- Data: $\{X_1, \cdots, X_n\}$
- Undata: $\{v_1, \cdots, v_n\}$

We parameterize $g = (g(v_i))_{i=1}^m \equiv (\gamma_i)_{i=1}^m$, thus:

- $\int \psi(g)dx \approx \sum s_i \psi(g(v_i)) \equiv s^\top \Psi(\gamma)$ Riemann Sum
- $\int g dP_n = \sum g(X_i) = w^\top L\gamma$ Linear Interpolation
- $g \in \mathcal{K} \Leftrightarrow D\gamma \geqslant 0 \quad D = \nabla^2$ Convex Cone Constraint

Yielding the primal and dual problems:

$$(P) \qquad \{w^\top L\gamma + s^\top \Psi(\gamma) \mid D\gamma \geqslant 0\} = \min!$$

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

The Discrete Charm of the Duality

$$(P) \qquad \{w^\top L\gamma + s^\top \Psi(\gamma) \mid D\gamma \geqslant 0\} = \min!$$

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

Theorem: (Sanity Check) In (P) suppose that for a vector of ones, ι, $w^\top L\iota = 1$ and $D\iota = 0$, then solutions f and g are strongly dual and satisfy:

$$f(v_i) = \psi'(g(v_i)) \quad i = 1, \cdots, m,$$

and $\int f(x)dx = \sum s_i f(v_i) = 1$, and $f(v_i) \geqslant 0$.

The Discrete Charm of the Duality

$$(\mathrm{P}) \qquad \{w^\top L\gamma + s^\top \Psi(\gamma) \mid D\gamma \geqslant 0\} = \min!$$

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

Theorem: (Sanity Check) In (P) suppose that for a vector of ones, ι, $w^\top L\iota = 1$ and $D\iota = 0$, then solutions f and g are strongly dual and satisfy:

$$f(v_i) = \psi'(g(v_i)) \quad i = 1, \cdots, m,$$

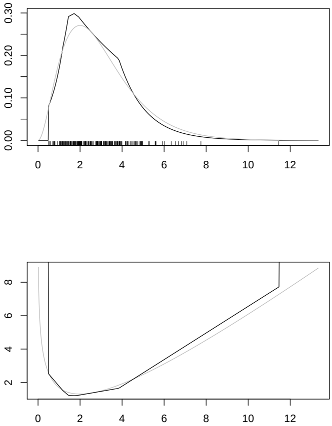
and $\int f(x)dx = \sum s_i f(v_i) = 1$, and $f(v_i) \geqslant 0$.

The argument for the integrability constraint is especially simple and revealing:

$$s^\top f \equiv \iota^\top Sf = \iota^\top Lw + \iota^\top D^\top h = 1$$

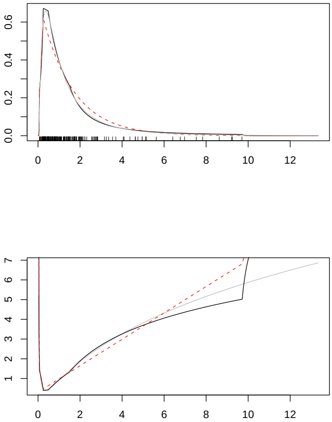
Since $D = \nabla^2$ the same argument implies that $\int xf(x)dx = \int xdP_n$.

A Gamma Example



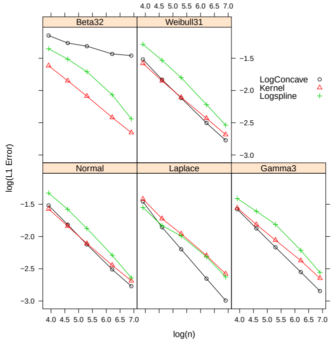
Log-concave Maximum Likelihood Estimator of a Gamma(3) Density

A Log-Normal Example

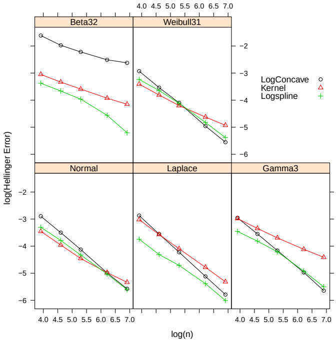


Log-concave and -1/2-concave Estimates of a Log-Normal Density

Simulation Evidence for Log-Concave Estimator



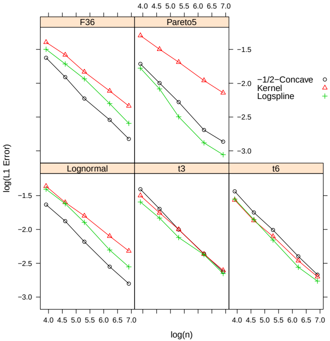
(a) L1 Error



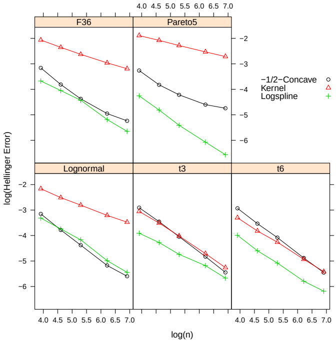
(b) Hellinger Error

Comparison of 3 Estimators: {Log-Concave, Kernel, Logspline}, for 5 Target Densities: {Beta(3,2), Weibull(3,1), Normal, Laplace, Gamma(3)}, with 5 sample sizes $\{50, 100, 200, 500, 1000\}$ and 500 replications.

Simulation Evidence for Hellinger Estimator



(a) L1 Error



(b) Hellinger Error

Comparison of 3 Estimators: $\{-1/2$-Concave, Kernel, Logspline$\}$, for 5 Target Densities: $\{$F(3,6), Pareto(5), Lognormal, t_3, $t_6\}$, with 5 sample sizes $\{50, 100, 200, 500, 1000\}$ and 500 replications.

Empirical Rates of Convergence

A naïve way to summarize the foregoing figures is to estimate a simple model for the implied rate of convergence for each of estimators:

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

where y_{ij} denotes a cell average of one of our two error criteria for one of our three estimators, for target density i and sample size n_j.

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

Estimated Convergence Rates for Log Concave Target Densities

Criterion	−1/2-Concave	Kernel	Logspline
L1 Error	−0.405 (0.004)	−0.324 (0.008)	−0.386 (0.01)
Hellinger	−0.751 (0.034)	−0.355 (0.023)	−0.672 (0.019)

Estimated Convergence Rates for −1/2-Concave Target Densities

An Historical Bivariate Example

"Student" (W.S. Gosset) in his celebrated 1908 paper writes:

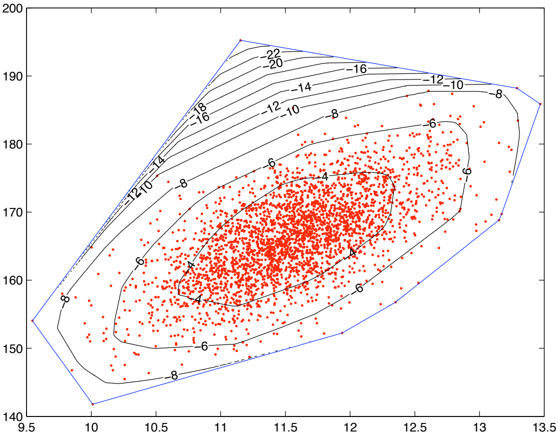
> *Before I had succeeded in solving my problem analytically, I had endeavoured to do so empirically. The material used was a correlation table containing the height and left middle finger measurements of 3000 criminals, from a paper by W. R. Macdonell. The measurements were written out on 3000 pieces of cardboard, which were then very thoroughly shuffled and drawn at random. Finally each consecutive set of 4 was taken as a sample . . .*

TABLE III. 3000 *Criminals*. Height (feet and inches).

Left Middle Finger (millimetres)	4' 7" 9/16—8" 9/16	8" 9/16—9" 9/16	9" 9/16—10" 9/16	10" 9/16—11" 9/16	5' 11" 9/16—0" 9/16	0" 9/16—1" 9/16	1" 9/16—2" 9/16	2" 9/16—3" 9/16	3" 9/16—4" 9/16	4" 9/16—5" 9/16	5" 9/16—6" 9/16	6" 9/16—7" 9/16	7" 9/16—8" 9/16	8" 9/16—9" 9/16	9" 9/16—10" 9/16	10" 9/16—11" 9/16	6' 11" 9/16—0" 9/16	0" 9/16—1" 9/16	1" 9/16—2" 9/16	2" 9/16—3" 9/16	3" 9/16—4" 9/16	4" 9/16—5" 9/16	Totals
9·4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
5	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
6	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
7	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
8	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
9	—	—	1	—	1	—	1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	3
10·0	1	—	—	1	2	—	2	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	7
1	—	—	—	1	3	1	—	1	1	—	—	—	—	—	—	—	—	—	—	—	—	—	7
2	—	—	2	2	2	1	—	2	—	1	—	—	—	—	—	—	—	—	—	—	—	—	10
3	—	1	1	3	2	2	3	5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	17
4	—	—	1	1	2	3	3	4	3	3	—	—	—	—	—	—	—	—	—	—	—	—	20
5	—	—	—	1	3	7	6	4	3	1	3	1	—	1	—	—	—	—	—	—	—	—	30
6	—	—	—	1	4	5	9	14	6	3	1	—	—	1	—	—	—	—	—	—	—	—	44
7	—	—	1	2	4	9	14	16	15	7	3	1	2	—	—	—	—	—	—	—	—	—	74
8	—	—	—	2	5	6	14	27	10	7	1	2	1	—	—	—	—	—	—	—	—	—	75
9	—	—	—	—	2	6	14	24	27	14	10	4	1	—	—	—	—	—	—	—	—	—	102
11·0	—	—	—	2	6	12	15	31	37	27	17	10	6	—	—	—	—	—	—	—	—	—	163
1	—	—	—	3	3	12	22	26	24	26	24	7	4	1	—	—	—	—	—	—	—	—	152
2	—	—	—	3	2	7	21	30	38	29	27	20	4	1	—	—	—	—	—	—	—	1	183
3	—	—	—	1	—	5	10	24	26	39	26	24	7	2	—	—	—	—	—	—	—	—	164
4	—	—	—	—	3	4	9	29	56	58	26	22	10	11	—	—	—	—	—	—	—	—	228
5	—	—	—	—	—	5	11	17	33	57	38	34	25	11	2	—	—	—	—	—	—	—	233
6	—	—	—	—	2	1	4	13	37	39	48	38	27	12	2	2	—	1	—	—	—	—	226
7	—	—	—	—	—	2	9	17	30	37	48	45	24	9	9	2	—	—	—	—	—	—	232
8	—	—	—	—	1	—	2	11	15	35	41	34	29	10	5	1	—	—	—	—	—	—	184
9	—	—	—	—	1	1	2	12	10	27	32	35	19	10	9	3	1	—	—	—	—	—	162
12·0	—	—	—	—	—	—	1	4	8	19	42	39	22	16	8	2	2	—	—	—	—	—	163
1	—	—	—	—	—	—	—	2	4	13	22	28	15	27	10	4	1	—	—	—	—	—	126
2	—	—	—	—	—	—	1	2	5	6	23	17	16	11	8	1	1	—	—	—	—	—	91
3	—	—	—	—	—	—	—	—	4	8	10	13	20	23	6	5	—	—	—	—	—	—	89
4	—	—	—	—	—	—	1	1	1	2	7	12	4	7	7	1	—	—	1	—	—	—	44
5	—	—	—	—	—	—	—	1	—	1	3	12	11	8	6	8	—	2	—	—	—	—	52
6	—	—	—	—	—	—	—	—	—	1	—	3	5	7	8	6	3	1	1	—	—	—	35
7	—	—	—	—	—	—	—	—	—	1	1	7	5	5	8	2	2	—	—	—	—	—	31
8	—	—	—	—	—	—	—	—	—	—	1	2	3	1	8	5	3	1	1	—	—	—	25
9	—	—	—	—	—	—	—	—	—	—	—	1	2	2	—	1	1	—	—	—	—	—	7
13·0	—	—	—	—	—	—	—	—	—	—	3	—	1	—	1	—	2	1	—	—	—	—	8
1	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—	—	—	—	—	—	2
2	—	—	—	—	—	—	—	—	—	—	1	1	—	1	—	3	—	—	—	—	—	—	6
3	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	1	—	—	—	2
4	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0
5	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—	—	—	—	1
Totals	1	1	6	23	48	90	175	317	393	462	458	413	264	177	97	46	17	7	4	0	0	1	3000
Means	100	103	102·6	107·0	107·8	109·4	110·6	111·8	113·3	114·8	116·5	117·7	118·6	120·1	122·2	123·9	125·9	126·4	127·7	—	—	112	

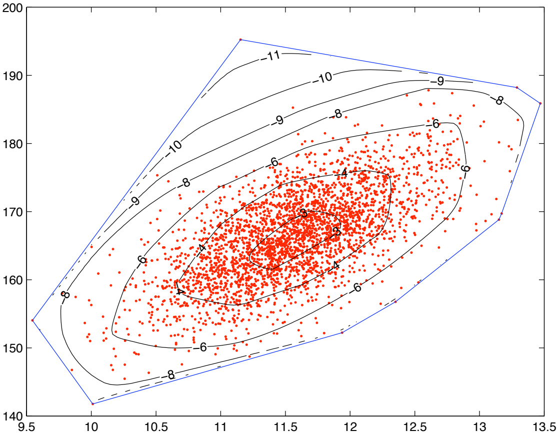
216 *On Criminal Anthropometry*

Student's Middle Fingers



Bivariate Log-Concave Estimate

Student's Middle Fingers, Again



Bivariate −1/2-Concave Hellinger Estimate

Regularization for Density Estimation

- An old idea (Good, Vapnik, . . .) whose time has come?

Regularization for Density Estimation

- An old idea (Good, Vapnik, . . .) whose time has come?
- Bayes (in mufti) procedures that shrink toward *a priori* plausible models for norm constraints.

Regularization for Density Estimation

- An old idea (Good, Vapnik, . . .) whose time has come?
- Bayes (in mufti) procedures that shrink toward *a priori* plausible models for norm constraints.
- Shape constraints also regularize thereby offering a middle ground between parametric and nonparametric modeling.

Regularization for Density Estimation

- An old idea (Good, Vapnik, ...) whose time has come?
- Bayes (in mufti) procedures that shrink toward *a priori* plausible models for norm constraints.
- Shape constraints also regularize thereby offering a middle ground between parametric and nonparametric modeling.
- ML estimation of log-concave densities is especially appealing on economic theory grounds.

Regularization for Density Estimation

- An old idea (Good, Vapnik, . . .) whose time has come?
- Bayes (in mufti) procedures that shrink toward *a priori* plausible models for norm constraints.
- Shape constraints also regularize thereby offering a middle ground between parametric and nonparametric modeling.
- ML estimation of log-concave densities is especially appealing on economic theory grounds.
- But other maximum entropy estimators of ρ-concave densities are also attractive and permit a broader (algebraic) class of tail behavior.

Regularization for Density Estimation

- An old idea (Good, Vapnik, . . .) whose time has come?
- Bayes (in mufti) procedures that shrink toward *a priori* plausible models for norm constraints.
- Shape constraints also regularize thereby offering a middle ground between parametric and nonparametric modeling.
- ML estimation of log-concave densities is especially appealing on economic theory grounds.
- But other maximum entropy estimators of ρ-concave densities are also attractive and permit a broader (algebraic) class of tail behavior.
- Why density estimation? Because it is a stepping stone toward the hegemony of semi-parametrics.