Shape Restricted Estimation With a View Towards Astronomy*

Michael Woodroofe

March 20, 2006

1 Some Statistical Problems

It's best to begin with some examples.

Example 1 *Consider an experimental pain reliever whose dosage x may be controlled by a physician. If y is hours of relief, then we might suppose that $y = \mu(x) + \epsilon$ where μ is a non-decreasing function and ϵ is a random error.*

Example 2 *Consider a galaxy (or globular cluster), and let $V = (V_1, V_2, V_3)$ denote the velocity of a randomly selected star. Suppose that the galaxy is isotropic, so that the density of V is of the form*

$$f(v_1, v_2, v_3) = h(v_1^2 + v_2^2 + v_3^2).$$

Often astronomers can measure radial velocities. With a proper choice of coordinate axes, this becomes V_3. The density of V_3 is

$$f(v_3) = \int_{\boldsymbol{R}^2} h(v_1^2 + v_2^2 + v_3^2) dv_1 dv_2 = \pi \int_0^\infty h(t + v_3^2) dt = \pi \int_{v_3^2}^\infty h(t) dt$$

This is a symmetric function of v_3 and decreasing in $v_3 > 0$.

The examples illustrate two basic statistical problems, the monotone regression problem and the decreasing density estimation problem. In the monotone regression problem,

$$y_k = \mu(x_k) + \epsilon_k, \tag{1}$$

*Last revised on March 20, 2006

where μ is a monotone function, $\epsilon_1, \cdots, \epsilon_n$ are independent errors with mean 0 and variances $\sigma^2/w_1, \cdots, \sigma^2/w_n$, $w_1, \cdots, w_n > 0$ are known, and σ^2 may be known or unknown. In this case it is natural to minimize

$$SS = \sum_{i=1}^{n} w_i(y_i - \theta_i)^2 \tag{2}$$

over $-\infty < \theta_1 \leq \theta_2 \cdots \leq \theta_n < \infty$.

In the density estimation problem, one observes a sample of positive random variables $X_1 \cdots, X_n \sim f$, were f is non-increasing on $(0, \infty)$. The likelihood function is then

$$L(f) = \prod_{i=1}^{n} f(x_i).$$

Let $x_1 < \cdots < x_n$ denote the order statistics; then $L(f)$ is maximized when f is a left continuous step function with knots at $x_1, \cdots, x_n$. Thus, letting $\theta_j = \log[f(x_j)]$, it is required to maximize

$$\log L = \sum_{i=1}^{n} \theta_i, \tag{3}$$

subject to $-\infty < \theta_n \leq \cdots \leq \theta_1 < \infty$ and

$$\sum_{i=1}^{n} (x_i - x_{i-1}) e^{\theta_i} = 1.$$

2 Convex Optimization

Let Ω denote a closed, convex subset of $\mathbb{R}^n$; let $\psi : \Omega \to \mathbb{R}$ be a convex function; and consider the problem of minimizing ψ over Ω. About ψ suppose that ψ is continuously differentiable and that $\lim_{\|\theta\| \to \infty} \psi(\theta) = \infty$. Then ψ attains its minimum value on Ω; and a necessary and sufficient condition for θ^o to minimize ψ is that

$$(\theta - \theta^o)' \nabla \psi(\theta^o) \geq 0 \tag{4}$$

for all $\theta \in \Omega$, where $'$ denotes transpose and $\nabla \psi$ is the gradient of ψ. For the necessity, suppose that θ^o minimizes ψ. Let $\theta^1 \in \Omega$; and let

$$g(t) = \psi[t\theta^1 + (1-t)\theta^o]. \tag{5}$$

Then

$$0 \leq \frac{g(t) - g(0)}{t} \to (\theta - \theta^o)' \nabla \psi(\theta^o)$$

as $t \to 0$. For sufficiency, suppose that $\theta^o \in \Omega$ satisfies (4); let $\theta^1 \in \Omega$; and define g by (5). Then g is a convex function (of a real variable) for which $g(0) = \psi(\theta^o)$ and $g(1) = \psi(\theta^1)$. Thus,

$$\psi(\theta^1) - \psi(\theta^o) = \int_0^1 g'(t) dt,$$

where now $'$ denotes derivative. Now g is convex, so that g' is a non-decreasing function, and $g'(0) = (\theta^1 - \theta^o)'\nabla\psi(\theta^o) \geq 0$ by assumption. So, $g'(t) \geq 0$ for all $0 \leq t \leq 1$ and, therefore $\psi(\theta^1) - \psi(\theta^o) \geq 0$.

The set Ω is called a *cone* if $c\theta \in \Omega$ for all $c > 0$ for each $\theta \in \Omega$. In this case

$$\theta^{o\prime}\nabla\psi(\theta^o) = 0. \tag{6}$$

If Ω is a cone and contains a linear subspace L, then there is equality in (4) for all $\theta \in L$. Ω is called a *polyhedral cone* is Ω is of the form

$$\Omega = \{\theta \in I\!R^n : \gamma_i'\theta \geq 0,\ i = 1, \cdots, m\},$$

where $\gamma_1, \cdots, \gamma_m \in I\!R^n$ are linearly independent. The monotone regression problem Example 1 is of this form with $m = n-1$ and $\gamma_i = (0, \cdots, -1, 1, 0, \cdots, 0)'$. If Ω is a polyhedral cone and $\psi(x) = x'Ax + b'x$ for some positive definite matrix A and a $b \in I\!R^n$, then the maximization problem is called a *quadratic programming problem*. There are several existing numerical techniques for finding the minimum, notably *interior point* methods and *primal dual basis* methods.

3 Isotonic Estimation

Monotone Regression. Consider the model described in (1). Then minimizing SS is a quadratic programming problem; but in this special case there is a closed form solution. Let $W_k = w_1 + \cdots + w_k$; let G be a piecewise linear function for which $G(0) = 0$ and $G(W_k) = y_1 + \cdots + y_k$ for $k = 1, \cdots, n$; and let $\tilde{G}$ be the greatest convex minorant of G. Here the graph of G is called the cumulative sum diagram. Next let $\tilde{g}$ be the left hand derivative of $\tilde{G}$. Then the least squares estimators are $\hat{\theta}_k = \tilde{g}(W_k)$. In symbols,

$$\hat{\theta}_k = \tilde{g}(W_k) = \frac{d}{dw}[\text{GCM of CSD}]_{|w=W_k}.$$

An alternative expression is

$$\hat{\theta}_k = \max_{i \leq k} \min_{j \geq k} \frac{w_i y_1 + \cdots + w_j y_j}{w_i + \cdots + w_j}.$$

The process is illustrated with $n = 10,\ x_k = k/10,\ \mu(x) = x$, and $\epsilon_k \sim \text{Normal}(0, 1)$.

To see that $\hat{\theta}$ minimizes SS, it suffices to show that the sufficient condition (4) is satisfied. This will be show that in the special case that $w_i = 1$. In this case (6) is satisfied, and $\nabla\psi(\theta) = -2[y_1 - \theta_1, \cdots, y_n - \theta_n]'$. So, it suffices to show that

$$\sum_{j=1}^{n} \theta_j(y_j - \hat{\theta}_j) \leq 0 \tag{7}$$

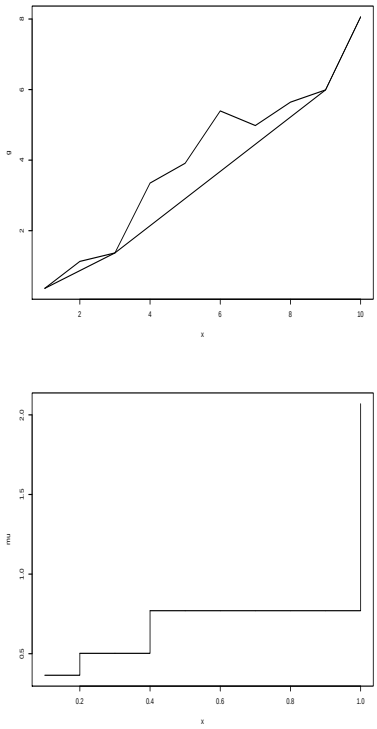

Figure 1: The GCM and MLE

for all non-decreasing sequences θ. It is clear from the graph that $y_1 + \cdots + y_n = \hat{\theta}_1 + \cdots + \hat{\theta}_n$. Let $\Delta\theta_j = \theta_j - \theta_{j-1}$. Then, summing by parts,

$$\sum_{j=1}^{n} \theta_j (y_j - \hat{\theta}_j) = \sum_{j=2}^{n} \Delta\theta_j [\tilde{G}_k - G_k] \leq 0,$$

so that (7) is satisfied.

Density Estimation. Now let f be a non-increasing density on $(0, \infty)$; let $X_1, \cdots, X_n \sim^{\text{ind}} f$ be a sample from f; and let $0 = x_1 < \cdots < x_n < \infty$ be the order statistics. Then the likelihood function can be maximized, using the techniques of convex optimization. To state the result, let

$$F^{\#}(t) = \frac{\#\{i \leq n : X_i \leq t\}}{n}$$

denote the emprical distribution function and let $\tilde{F}$ be its least concave majorant. Then the (non-parametric) maximum likelihood estimator of f is a step function with knots at $x_1, \cdots, x_n$. The value of $f(x_k)$ is the left hand slope of $\tilde{F}$ at x_k; in symbols

$$\tilde{f}(x_k) = \frac{d}{dx} \tilde{F}_{|x=x_k-}.$$

An alternative expression is

$$\tilde{f}(x_k) = \min_{i \le k} \max_{j \ge k} \frac{(j-i+1)}{n(x_j - x_{i-1})},$$

where $x_0 = 0$.

x	.585	1.262	3.138	3.142	3.979	4.560	7.151	10.261
$\tilde{f}(x)$	.214	.185	.152	.152	.152	.152	0.048	.040

Time in weeks since last reboot of eight work stations; $\hat{\nu} = 4.67$

Example. Let X denote the time since the last breakdown and repair of machinery. Then, supposing that breakdowns form a renewal process, the density of X is

$$f(x) = \frac{1}{\nu}[1 - G(x)],$$

where G is the distribution function of times between repairs. Observe that $\nu = 1/f(0)$, and

$$G(x) = 1 - \frac{f(x)}{f(0)}.$$

Clearly f is non-increasing. The data in the table below are the time since last reboot of computers in the University of Michigan Statistics Department. Machines were only rebooted when there was a problem.

For more detail on isotonic estimation, see Chapters 1 and 7 of [2]

4 Large Sample Properties

I will explain this for the density estimation problem. There are analogous results for the regression problem.

Consistency. Let h denote the *Hellinger metric* for densities

$$h^2(f,g) = \int_{-\infty}^{\infty} [\sqrt{f(x)} - \sqrt{g(x)}]^2 dx = 2[1 - \int_{-\infty}^{\infty} [\sqrt{f(x)g(x)} dx].$$

The first result is remarkable only for its generality: *If f is non-increasing then*

$$\lim_{n \to \infty} h(f, \tilde{f}_n) = 0 \ w.p.1.$$

Unfortunately, this result does not imply that $\tilde{f}_n(0)$ is consistent, and it isn't. In fact, *if $f(0) < \infty$, then*

$$\frac{\tilde{f}_n(0)}{f(0)} \Rightarrow \sup_{k \ge 1} \frac{k}{\Gamma_k},$$

$\Gamma_k = E_1 + \cdots + E_k$ and $E_1, E_2, \cdots$ are i.i.d. exponential random variable.

Limiting Distributions. Now suppose the f is positive and continuously differentiable near $t_0 > 0$, and let

$$\kappa = |4f(t_0)f'(t_0)|.$$

Then

$$n^{\frac{1}{3}}[\tilde{f}(t_0) - f(t_0)] \Rightarrow \kappa Z,$$

where

$$Z = \text{argmax}_{s \in \boldsymbol{R}} W(s) - s^2,$$

and W denotes a two-sided Brownian motion.

The Kiefer Wolfowitz Theorem. If f has bounded support and f' is continuous and negative throught the support, then

$$\lim n \to \infty \frac{n^{2/3}}{\log(n)} \sup_t |\tilde{F}_n(t) - F_n^{\#}(t)| = 0.$$

The knowledge that f is decreasing does no good for estimating the distribution function.

5 Wicksell's Problem

Suppose that X_1, X_2, X_3 have spherically symmetric density, but only X_1 and X_2 are observed. Let

$$Z = X_1^2 + X_2^2 + X_3^2 \sim f$$
$$Y = X_1^2 + X_2^2 \sim g$$

Then

$$g(y) = \frac{1}{2} \int_y^\infty \frac{f(z)dz}{\sqrt{z(z-y)}}.$$

Let

$$V(t) = \int_t^\infty \frac{g(y)dy}{\sqrt{y-t}}.$$

Then

$$V(t) = \cdots = \frac{1}{2}\pi \int_t^\infty \frac{f(z)dz}{\sqrt{z}}.$$

So, U is a decreasing function, and

$$f(t) = -\frac{2}{\pi} V'(t)$$

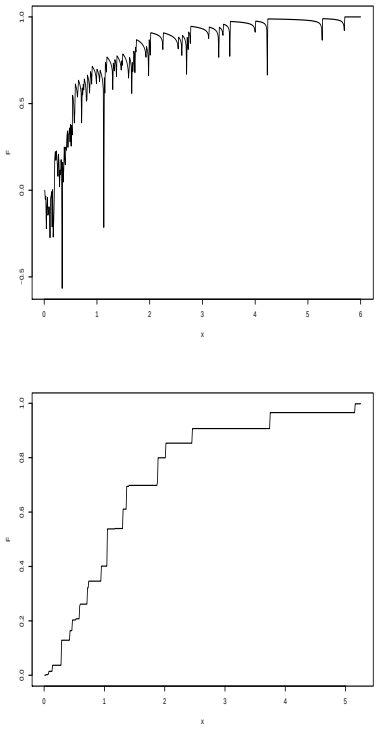

Figure 2: The ubiased and isotonic estimates of F for a sample of 100 from an exponential distribution

and

$$F(t) = 1 + \frac{2}{\pi}\int_t^\infty \sqrt{z} dV(z).$$

Now suppose

$$Y_1, \cdots, Y_n \sim g.$$

Let

$$V_n^{\#}(t) = \sum_{i:Y_i>t} \frac{1}{\sqrt{Y_i - t}}.$$

Then $V_n^{\#}(t)$ is an unbiased estimation or U for each t, but $U_n^{\#}$ is not monotone as a function of t. In fact, it has an infinite discontinuity at each Y_i. Let

$$U_n(t) = \int_0^\infty V n^{\#}(s) ds$$
$$\tilde{U}_n = \text{LCM} U_n$$
$$\hat{V}_n(t) = \tilde{U}_n'(t),$$

and

$$\hat{F}_n(t) = 1 + \frac{2}{\pi}\int_t^\infty \sqrt{z} d\hat{V}_n(z).$$

The example is adapted from [1]; [3] extend these techniques to obtain estimates of the distribution of dark matter in nearby dwarf spheroidal galaxies.

References

[1] Groeneboom, P. and K. Jongbloed (1995). (1995, *Annals of Statistics*)

[2] Robertson, Tim, Farrell Wright, and Richard Dykstra (1989). *Order Restricted Inference*, Wiley.

[3] Wang, et. al. (2006, *Astrophysical Journal*).