Shape constrained inference

introduction & brief overview

Geurt Jongbloed

Delft Institute of Applied Mathematics

TU Delft

EYSM2015, September 2, 2015

Praha

EYSM 1993, Palanga, Lithuania

Dmitri Kramkov
Christophe Croux
Hein Putter
Ian Dryden
Tarmo Koll
Martin Baxter
Mindaugas Bloznelis
……

Outline

- Fitting a statistical model
- Isotonic regression
- Estimating a monotone density
- The current durations model
- Application to French fertility data
- More examples and directions

Fitting a statistical model

Observations: $X_1, \ldots, X_n$, i.i.d. $\sim F_0$
$F_0 \in \mathcal{F}$, statistical model

- $\mathcal{F}$: normal distribution functions
- $\mathcal{F}$: df's with $|F''| \leq 10$
- $\mathcal{F}$: concave df's on $[0, \infty)$
- $\mathcal{F}$: convex-concave df's on $[0, \infty)$

Observations: $(x_1, Y_1), \ldots, (x_n, Y_n)$, where

$$Y_i = r_0(x_i) + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n \sim N(0, \sigma^2)$, i.i.d. and $x_1 < \cdots < x_n$

Statistical model: $r_0 \in \mathcal{M}$

$\mathcal{M}$: $r_0(x) = \alpha + \beta x$

$\mathcal{M}$: $|r_0''|$ is continuous

$\mathcal{M}$: r_0 is increasing

Isotonic Regression

Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$x_1 < x_2 < \cdots < x_n$$

Least Squares estimate minimizes:

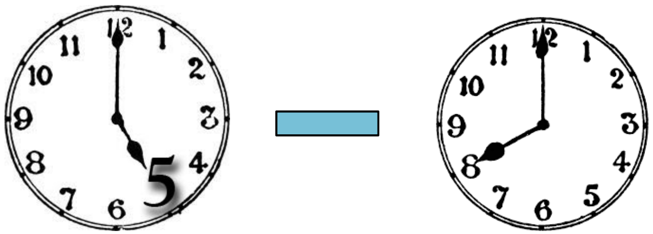
$$Q(r) = \frac{1}{2}\sum_{i=1}^{n}(r(x_i) - y_i)^2 \quad \text{such that } r \uparrow$$

Note: Q only depends on r *at* the x_i's

⇒ write r_i for $r(x_i)$

⇒ minimization over subset of $\mathbb{R}^n$

Examples



Solution to the minimization problem

$$Q(r) = \frac{1}{2}\sum_{i=1}^{n}(r_i - y_i)^2 \qquad r_1 \le r_2 \le \cdots \le r_n$$

$$\left.\begin{array}{l} \mathbf{1}.\ P_0 = (0,0) \\ \mathbf{2}.\ P_i = \left(i, \sum_{j=1}^{i} y_j\right),\ 1 \le i \le n \end{array}\right\} \quad \text{CSD}$$

3. Construct greatest convex minorant of

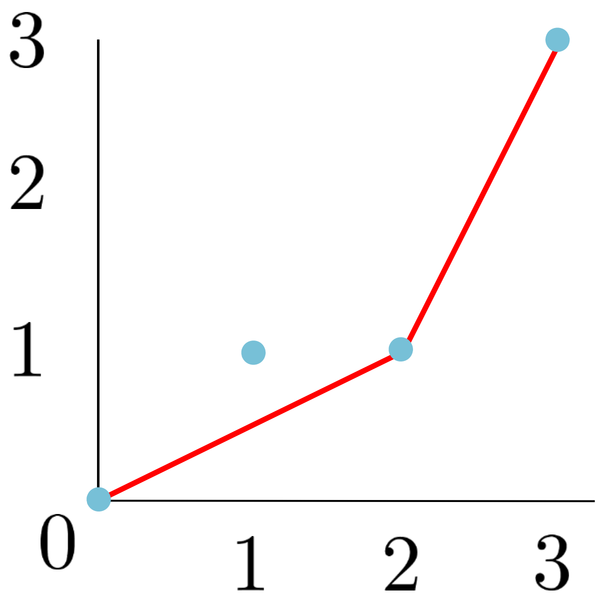
$$\{P_i \,:\, 0 \le i \le n\}$$

$\Rightarrow$ $\hat{r}_i$: left derivative of GCM at P_i

Example: $(2,1),(4,0),(7,2)$

1. $P_0 = (0,0)$ **2.** $P_i = \left(i, \sum_{j=1}^{i} y_j\right), \; 1 \le i \le n$

3. GCM of $\{P_i : 0 \le i \le n\}$



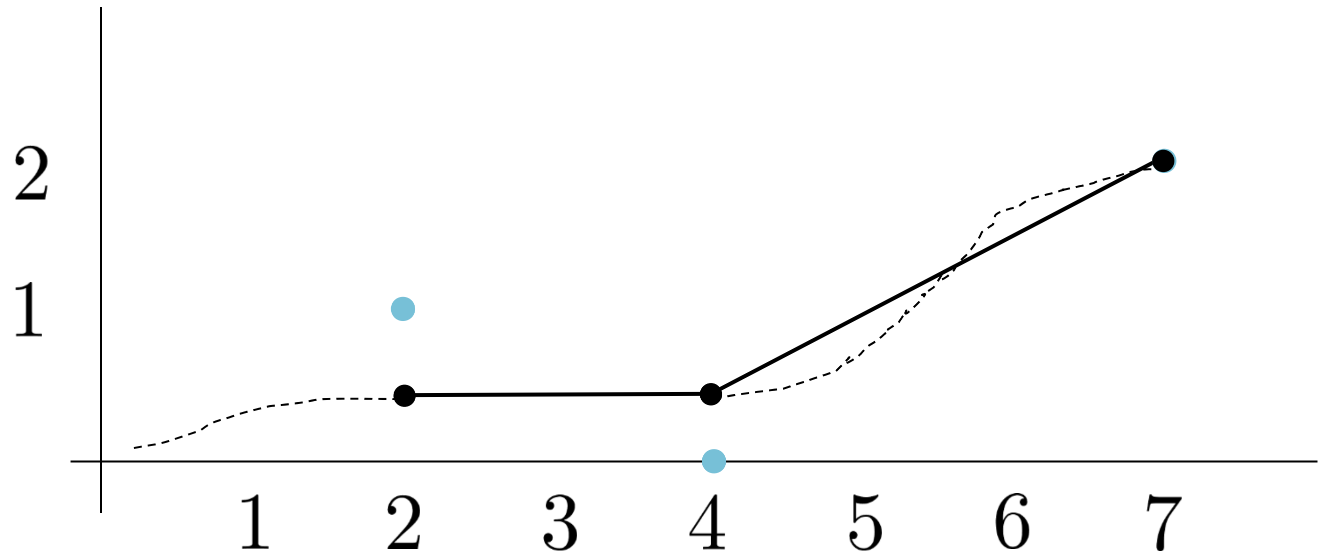
$$\hat{r}_1 = \frac{1}{2} = \hat{r}_2$$

$$\hat{r}_3 = 2$$

$\hat{r}_i$: left derivative of GCM at P_i

Example: $(2,1),(4,0),(7,2)$

$$\hat{r} = \left(\frac{1}{2}, \frac{1}{2}, 2\right)$$



Estimating a decreasing density

$X_1, X_2, \ldots, X_n \sim g$, i.i.d.

g: decreasing on $[0, \infty)$

Data: $x_1, x_2, \ldots, x_n$

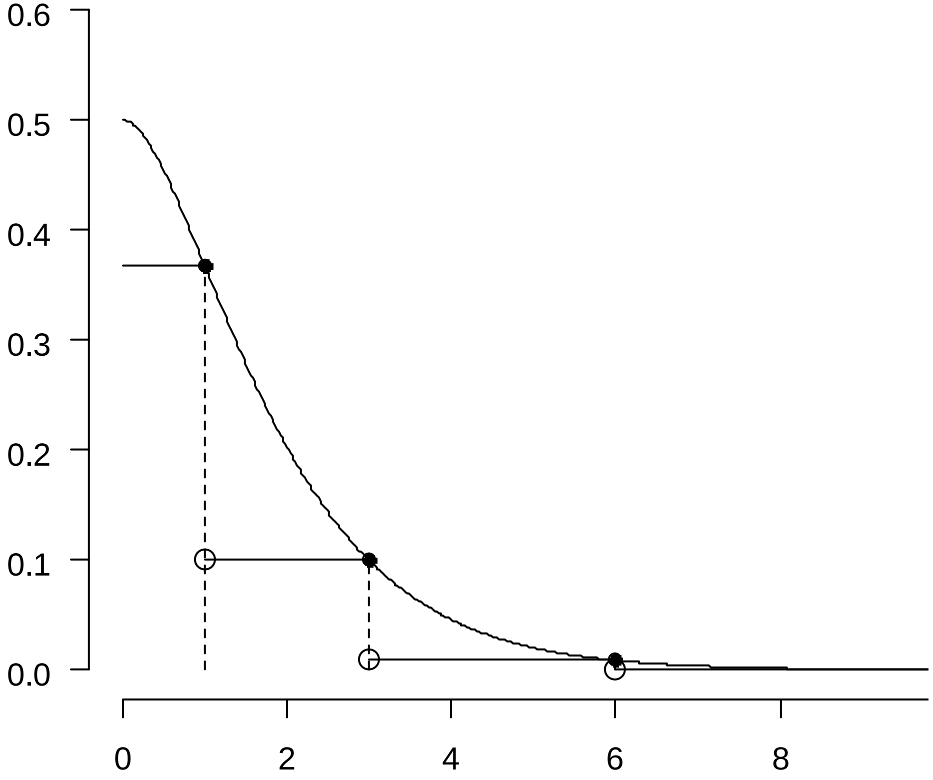
Estimate g based on $x_1, \ldots, x_n$

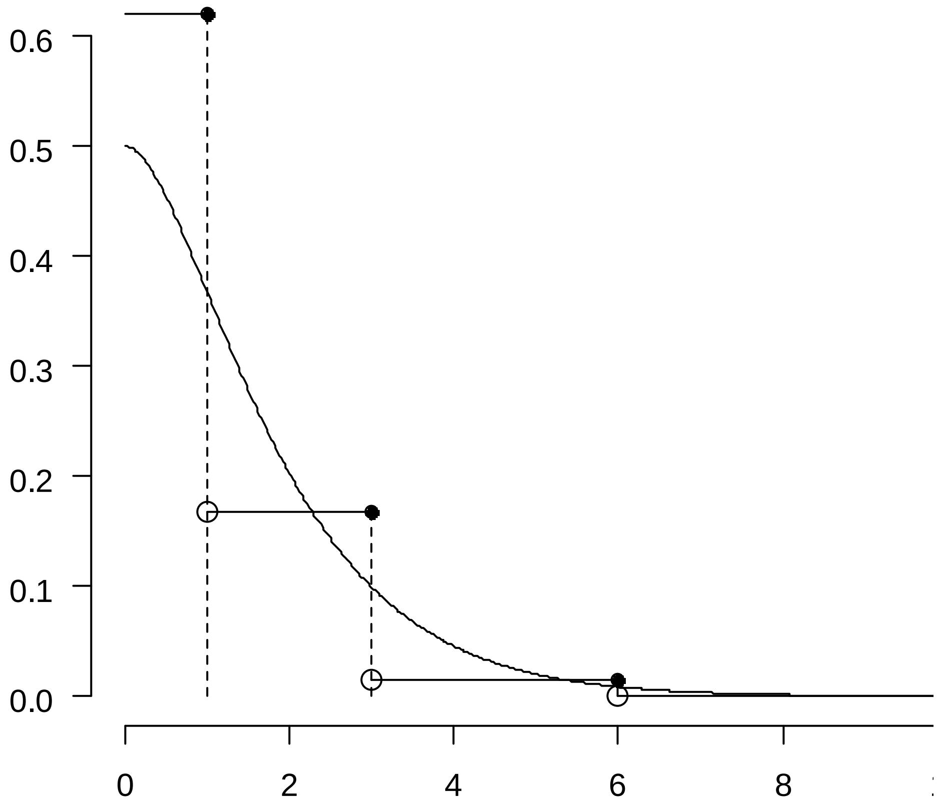
Log likelihood: $\ell(g) = \sum_{i=1}^{n} \log g(x_i)$

MLE for decreasing density

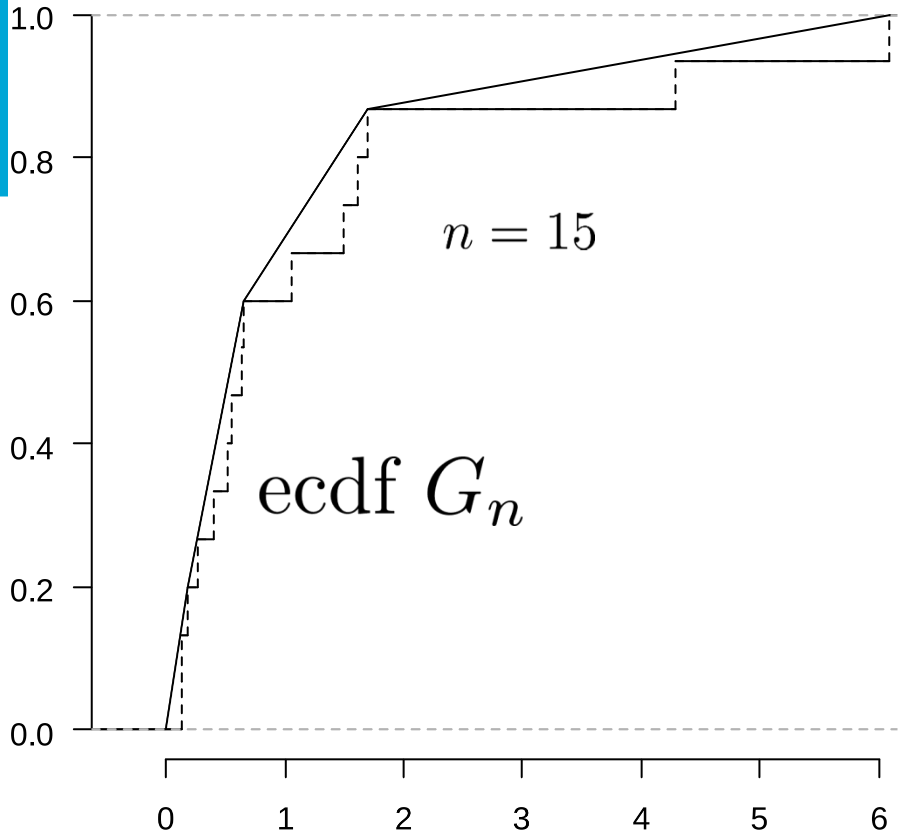
$$\ell(g) = \sum_{i=1}^{n} \log g(x_i)$$

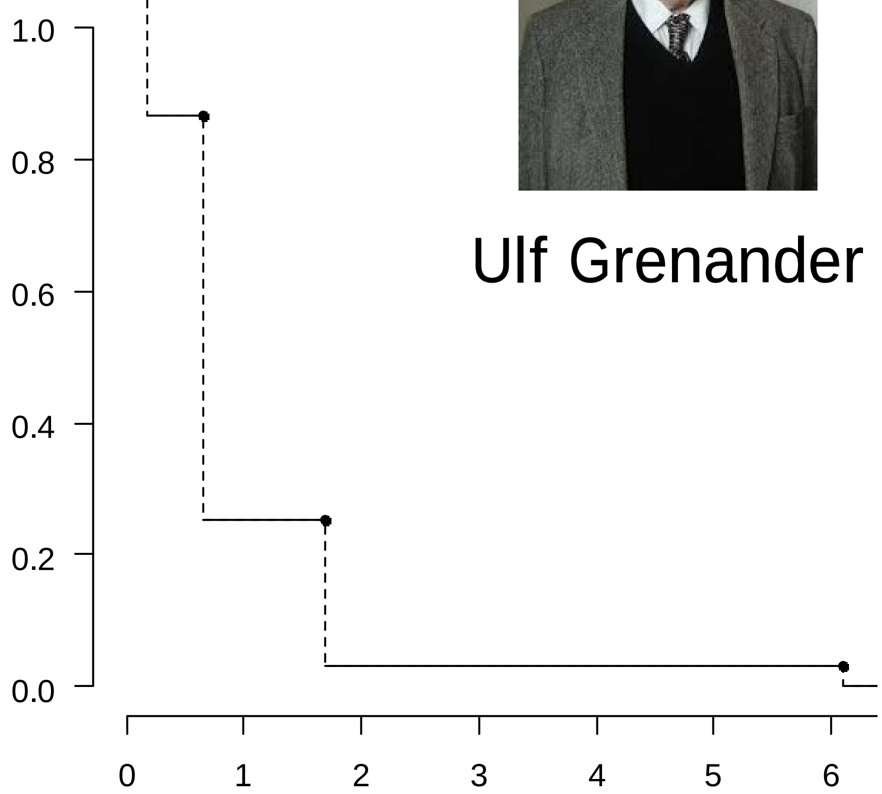
Maximizer: piecewise constant
left continuous





Construction of ML estimator





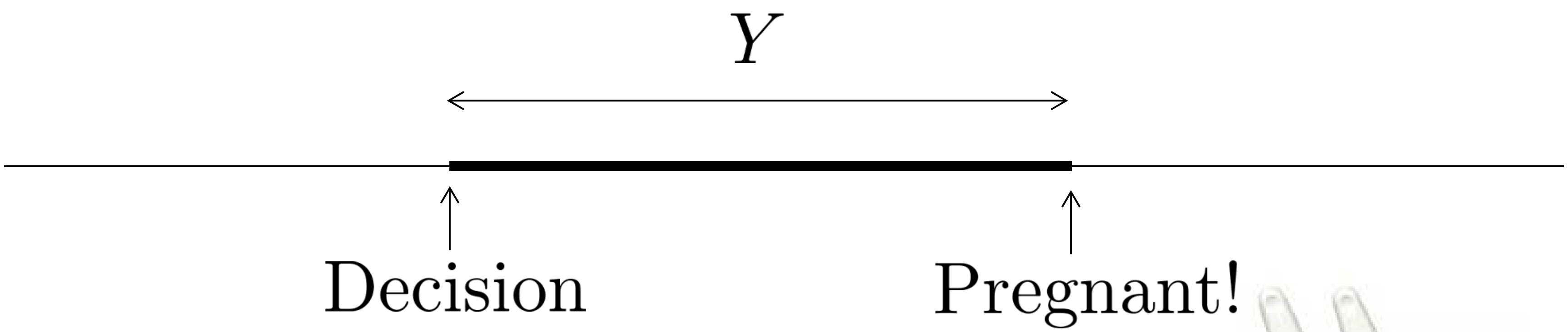
Ulf Grenander

Why would one be interested in

estimating a decreasing density?

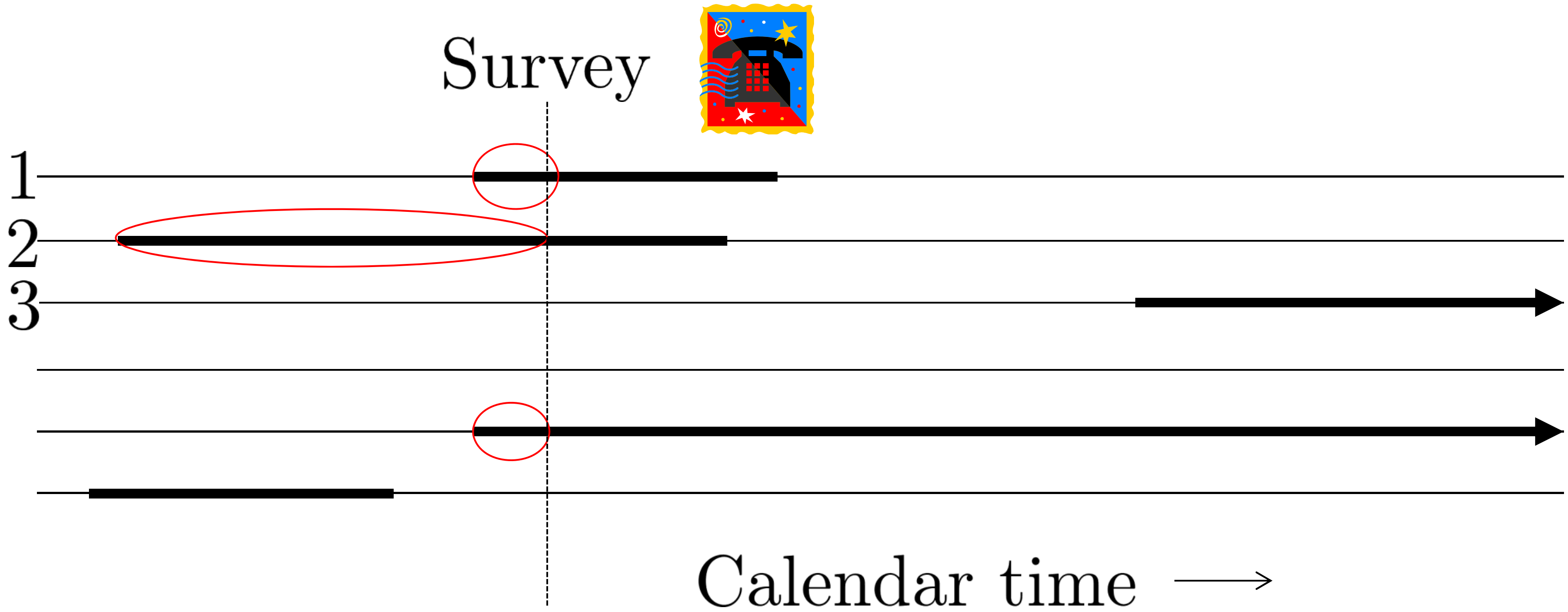
The current durations model

How long does it take to become pregnant?



Slama, R. *et al* (2012)
Estimation of the frequency of involuntary infertility on a nation-wide basis.
HUMAN REPRODUCTION

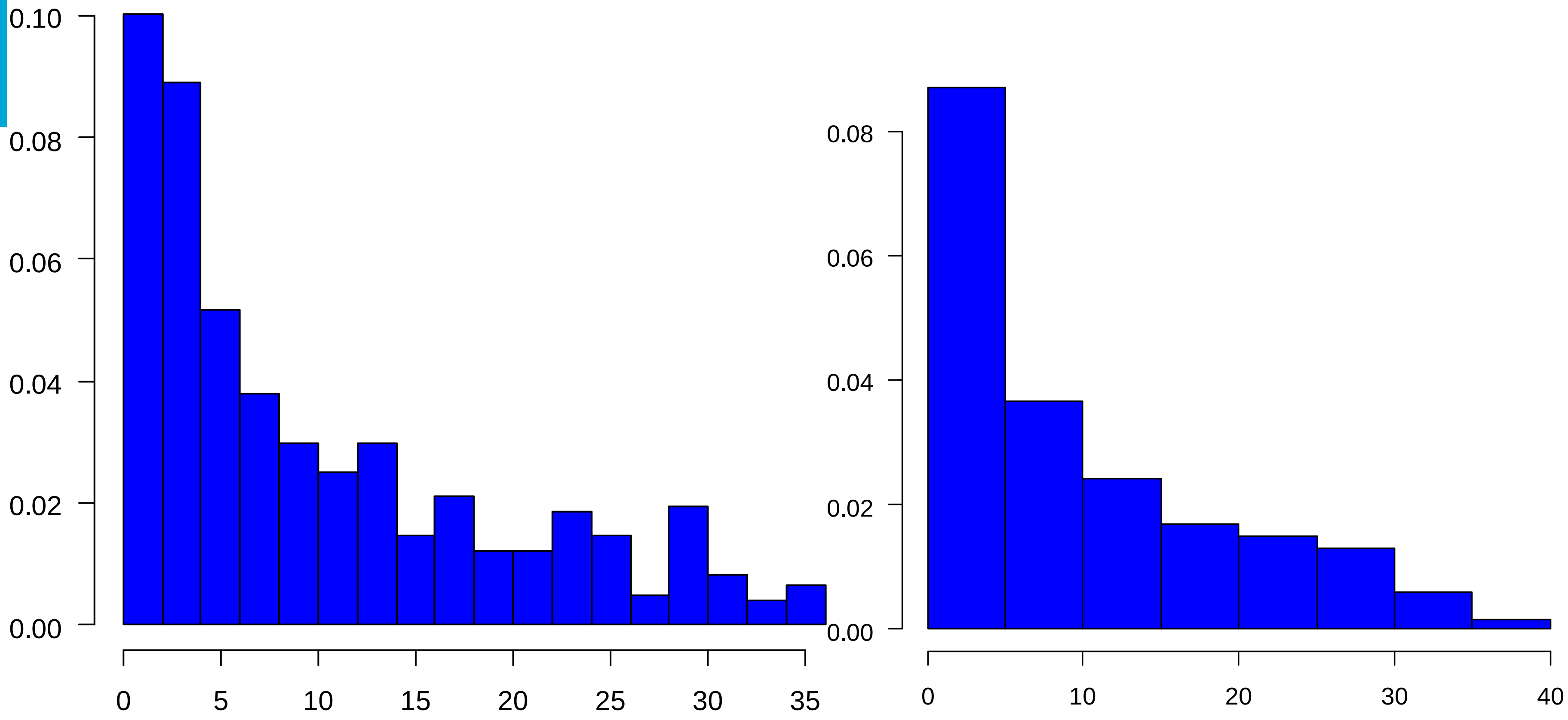
Cross-sectional study in France:



Question: when did 'your trial' start?

⟹ $X_1, X_2, \ldots, X_n$

Histogram of 618 values ≤ 36



Note: women with *larger* duration over-represented

Given the woman is in the sample, her current duration is draw from the *length biased* distribution associated with F_Y

$$\tilde{F}_Y(y) = \frac{\int_0^y x\, dF_Y(x)}{\int (1 - F_Y(x))\, dx}$$

X *smaller* than actual duration

It is a uniform random fraction of $\tilde{Y}$ taken from $\tilde{F}_Y$

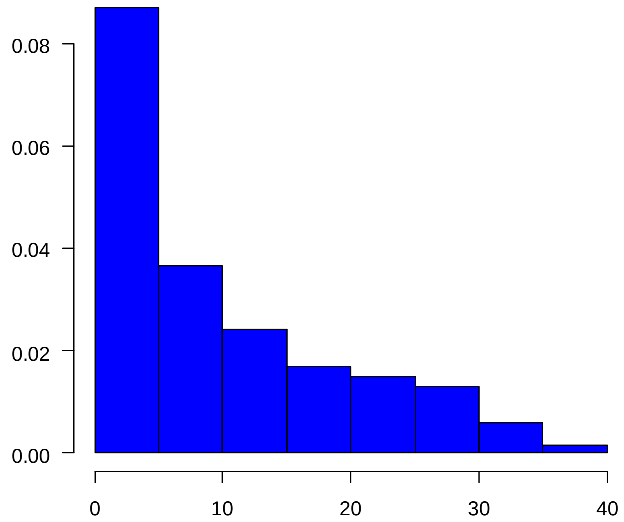
$$X = U\tilde{Y}$$

Relation distribution X and Y:

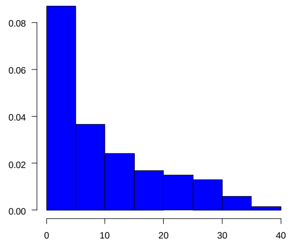
$$g_X(x) = \frac{1 - F_Y(x)}{\int y\, dF_Y(y)}, x \geq 0$$

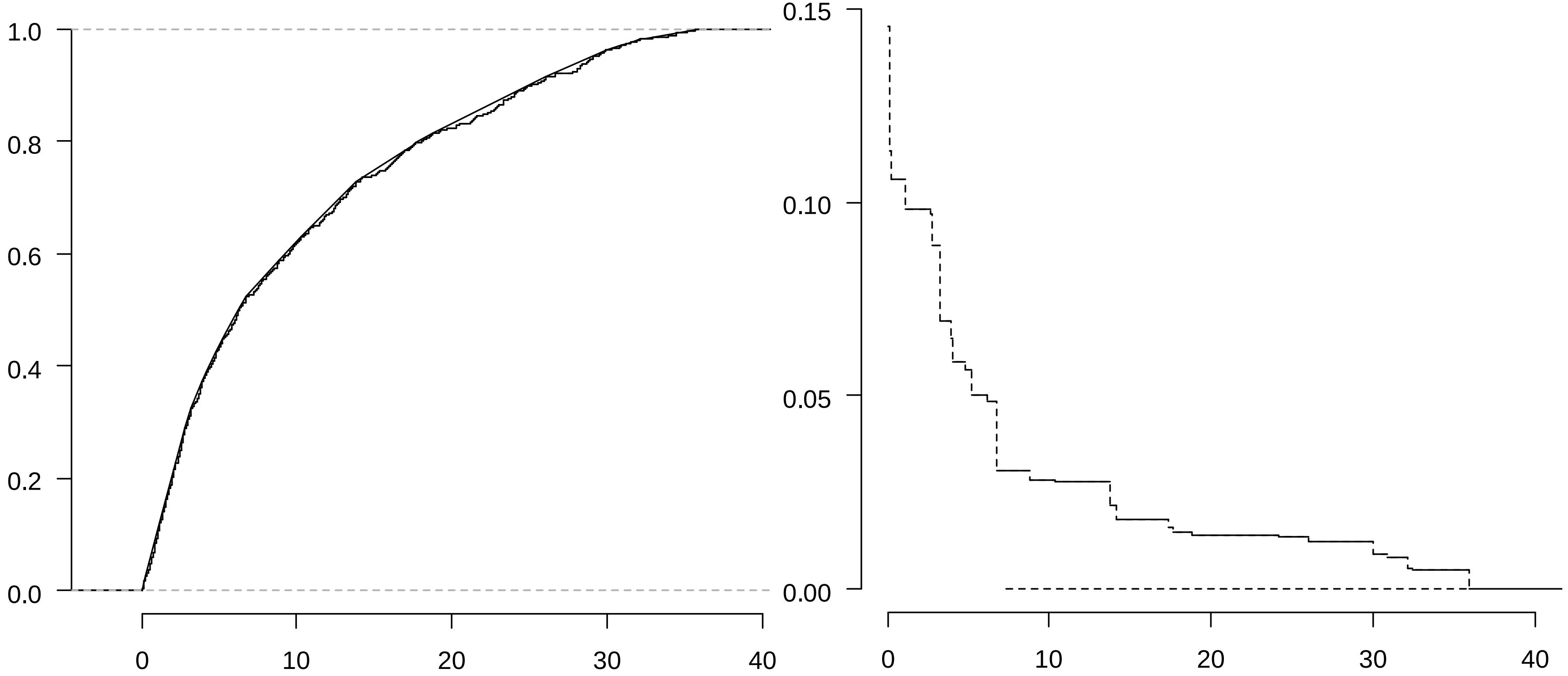
sampling density bounded, decreasing on $(0, \infty)$

$$F_Y(y) = 1 - \frac{g_X(y)}{g_X(0)}$$



MLE for the French fertility data

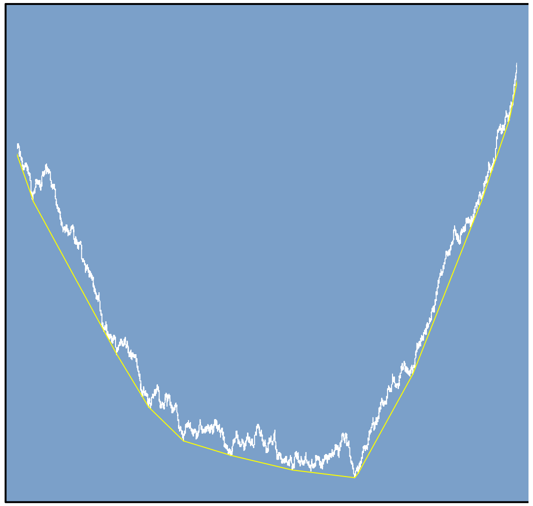


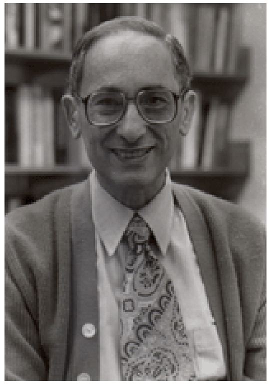


Asymptotic distribution of $\hat{g}_n(x)$ for $x > 0$.

$$n^{1/3} \frac{\hat{g}_n(x) - g_X(x)}{(g'_X(x) g_X(x)/2)^{1/3}}$$

$$\Rightarrow^D \text{ slocom } \left[t \mapsto W(t) + t^2\right]_{t=0}$$





Herman Chernoff

- Estimate $g_X(0)$
- Construct (pointwise) confidence band for g_X, F_Y

The MLE at zero

$$\hat{g}_n(0) = \max_{1 \le i \le n} \frac{\hat{G}_n(X_{(i:n)})}{X_{(i:n)}} = \max_{1 \le i \le n} \frac{i}{nX_{(i:n)}}$$

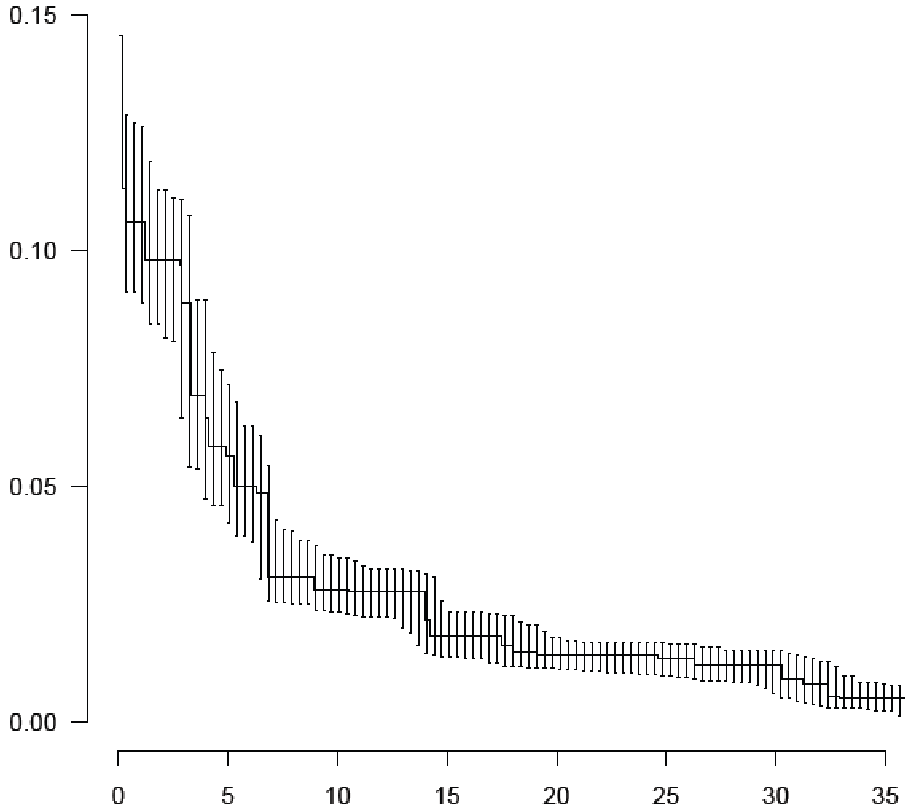
$$\ge \frac{1}{nX_{(1:n)}} = \frac{g_X(0)}{g_X(0)nX_{(1:n)}} \approx \frac{g_X(0)}{\mathrm{Exp}(1)}$$

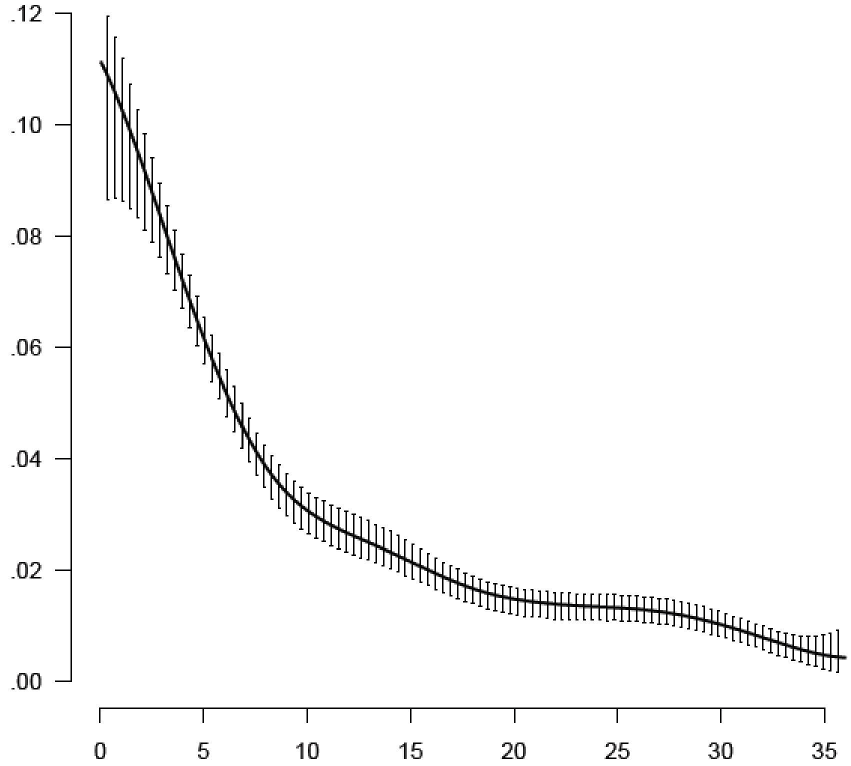
⇒ MLE is *inconsistent* at zero!

Confidence set for $g_X(x)$

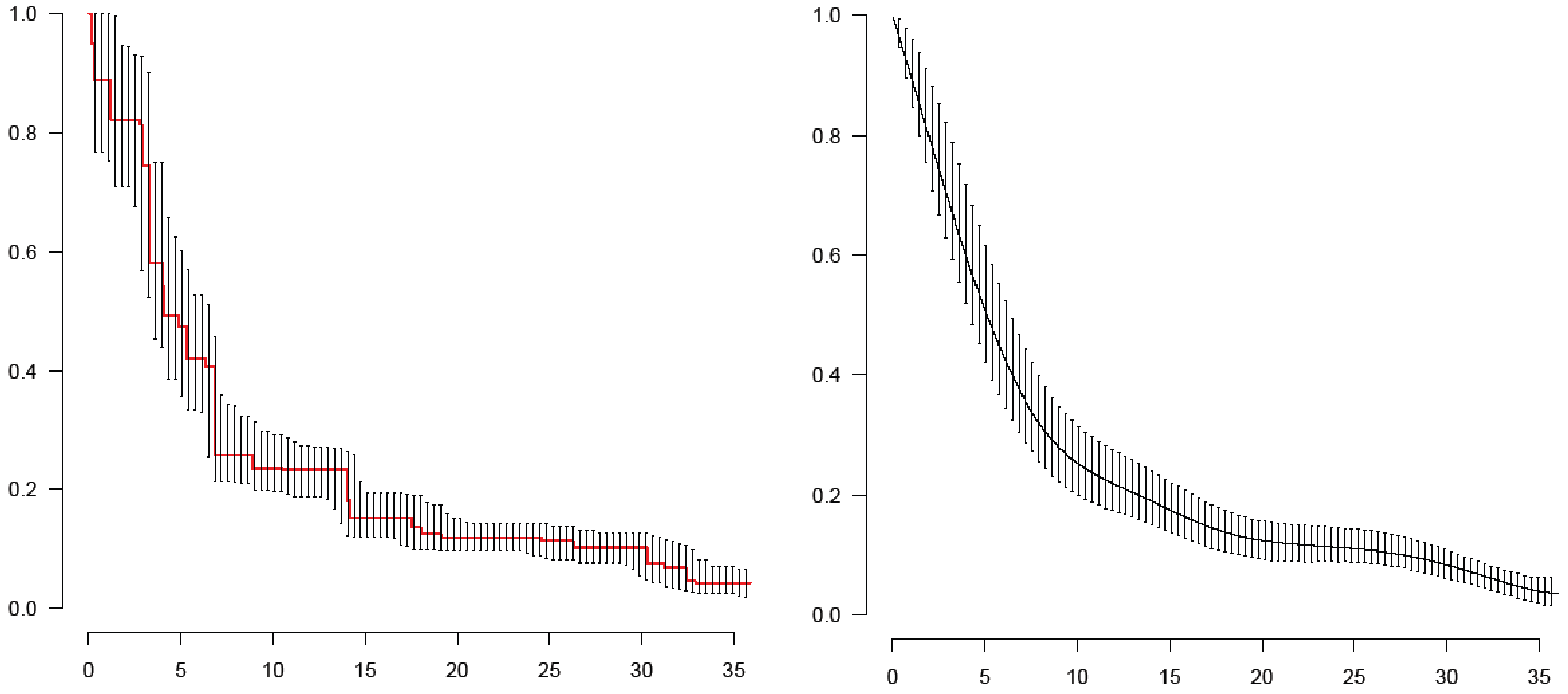
Likelihood Ratio approach

Bootstrap approach with smoothed estimate





Confidence sets for duration survival function $\frac{g_X(x)}{g_X(0)}$



More examples

The current status model

$$g(t,\delta) = \delta F(t)h(t) + (1-\delta)(1-F(t))h(t)$$

$$(t,\delta) \in [0,\infty) \times \{0,1\}$$

Min-uniform

$$G(z) = zF(z),\ z \in (0,1)\ \ F(x) = G(x)/x,\ x \in (0,1)$$

Competing risk with current status data

Bivariate interval censoring

Oriented cylinder model

many more

Summary

Small vs big statistical models

Isotonic regression: model and solution

Estimating a decreasing density: MLE

The current durations model

Some more examples

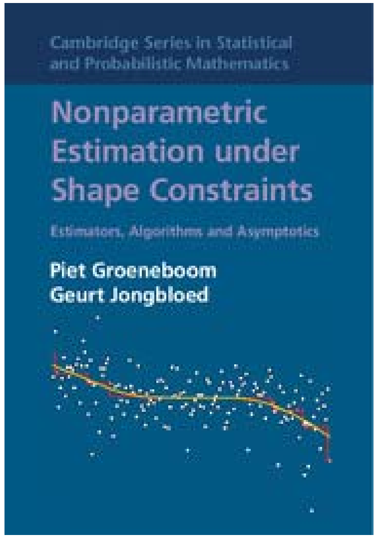
References

Banerjee, M. and Wellner, J.A. (2001)
Likelihood ratio tests for monotone functions
Annals of Statistics

Grenander, U. (1956)
On the theory of mortality measurement
Skand. Aktuarietidskr.

Groeneboom, P. and Jongbloed, G. (2014)
Nonparametric estimation under shape constraints
Cambridge University Press

Groeneboom, P. and Jongbloed, G. (2015)
Nonparametric confidence intervals for monotone functions
to appear in Annals of Statistics



Silvapulle, M.J. and Sen, P.K. (2005)
Constrained statistical inference: inequality, order and shape constraints Wiley

Robertson, T., Wright, F.T. and Dykstra, R.L. (1988)
Order restricted statistical inference. Wiley

Slama, R. *et al* (2012)
Estimation of the frequency of involuntary infertility on a nation-wide basis.
Human reproduction

Woodroofe, M. and Sun, J. (1993)
A penalized MLE of $f(0+)$ when f is nonincreasing.
Statistica Sinica

Thanks for your attention