Maximum likelihood estimation of a log-concave density based on censored data

Dominic Schuhmacher

Institute of Mathematical Statistics and Actuarial Science
University of Bern

Joint work with Lutz Dümbgen and Kaspar Rufibach

Swiss Statistics Seminar, 25 October 2011

Overview

1. A review of log-concave density estimation for exact data
2. Log-concave density estimation for censored data
   - Problem formulation
   - Theoretical results
   - An EM algorithm
   - Simulated and real data examples

Part 1: Log-concave density estimation for exact data

Problem

Problem: Nonparametric estimation of the density f of a distribution P on $\mathbb{R}$ based on independent observations $X_1, X_2, \ldots, X_n$ using maximum likelihood.

Log-likelihood:

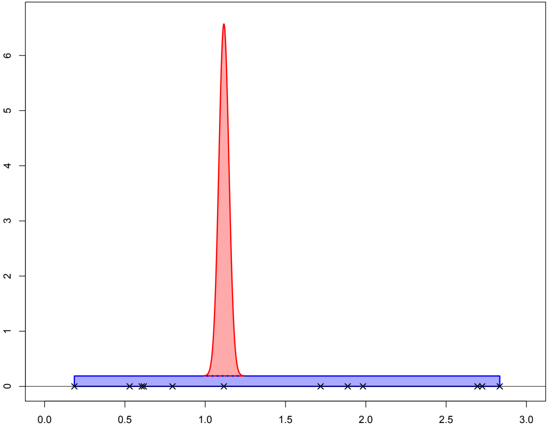
$$\ell(f) := \sum_{i=1}^{n} \log f(X_i),$$

not bounded from above on space of all densities f.

Our approach: impose shape constraint.
Unimodality is not enough.

Unimodality is not enough!

Distribute mass $1/2$ uniformly over $[x_{(1)}, x_{(n)}]$ (blue) and concentrate mass $1/2$ more and more closely around one individual point (red).



Log-concavity

A density *f* is called *log-concave* if $\varphi := \log \circ f\colon \mathbb{R} \to [-\infty, \infty)$ is a concave function.

Turns out that class $\mathcal{F}_{\mathrm{lc}}$ of log-concave densities is a large, rather flexible class with many nice properties.

Properties of $\mathcal{F}_{\mathrm{lc}}$

- Every log-concave density is unimodal.
- Ibragimov (1956): f log-concave if and only if $f * g$ is unimodal for every unimodal g
  (Log-concavity has been called “strong unimodality”).
- $\mathcal{F}_{\mathrm{lc}}$ is closed under convolution and weak limits.
- S, Hüsler and Dümbgen (2009/11), strong continuity property: weak convergence implies pointwise convergence of densities, and conv. in an exponentially weighted total variation distance: $\int \exp(\delta|x|)|f_n(x) - f(x)|\, dx \to 0$ for some $\delta > 0$.
- Every log-concave density has subexponential tails and a non-decreasing hazard rate.

Parametric families of log-concave distributions

The class of log-concave densities contains the following parametric families:

- Normal, logistic, exponential, Laplace, uniform, Gumbel
- Gamma and Weibull with shape parameter ≥ 1
- Beta with both shape parameters ≥ 1

Log-concave density estimation

Nonparametric maximum likelihood estimation under log-concavity constraint has been studied independently in

- Rufibach (2006) and Dümbgen and Rufibach (2009)
- Pal, Woodroofe, and Meyer (2007)

Task: for $n \geq 2$ find

$$\hat{f}_n \in \arg\max_{f \in \mathcal{F}_{\mathrm{lc}}} \sum_{i=1}^{n} \log f(X_i),$$

where $\mathcal{F}_{\mathrm{lc}} := \{f : \mathbb{R} \to \mathbb{R}_+ \text{log-concave density}\}$.

First results

Equivalent problem (by “Silverman’s trick”)**:** find

$$\widehat{\varphi}_n \in \arg\max_{\varphi \in \Phi} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) - \int_0^\infty \exp\varphi(x)\, dx}_{=:L(\varphi)}$$

where Φ is set of all concave (and upper semicontinuous, say) functions $\mathbb{R} \to [-\infty, \infty)$.

Shape: $\widehat{\varphi}_n$ piecewise linear, changes of slope only possible at data points X_i, $\operatorname{dom}(\widehat{\varphi}_n) = [X_{(1)}, X_{(n)}]$.

L depends on φ only via $\varphi(X_1), \ldots, \varphi(X_n)$. Is strictly concave and coercive; defined on a closed, convex cone $\subset \mathbb{R}^n$.

$\Longrightarrow$ **Existence and uniqueness** of NPMLE $\hat{f}_n$.

Further properties of $\hat{f}_n$ (simplified)

(Dümbgen & Rufibach, 2009, and Balabdaoui, Rufibach & Wellner, 2009)

- **Mean($\hat{f}_n$) $=$ empirical mean, Variance($\hat{f}_n$) $\leq$ empirical variance.**
- **Uniform consistency:** $\|\hat{f}_n - f\|_\infty \to 0$, and $\|\widehat{F}_n - F\|_\infty \to 0$.
- **Rate-optimality and adaptivity:** For any compact interval $T \subset \mathrm{int}\{f > 0\}$ and a true f that is Hölder continuous with exponent $\beta \in [1, 2]$, we have
$$\max_{x \in T} |\hat{f}_n(x) - f(x)| = O_p\left(\left(\frac{\log(n)}{n}\right)^{\beta/(2\beta+1)}\right).$$
- **Pointwise limit distributions:** If $f(x_0) > 0$, $\varphi \in C^2$ around x_0, and $\varphi''(x_0) \neq 0$:
$$n^{2/5}\left(\hat{f}_n(x_0) - f(x_0)\right) \xrightarrow{\mathscr{D}} c(x_0, f) H''(0),$$
where H is so-called “lower invelope” of $Y(t) = \int_{0 \wedge t}^{0 \vee t} W(s)\, ds - t^4$, W two-sided Brownian motion starting in 0.
- **results for derived quantities:** mode, distance between knots.

Computation

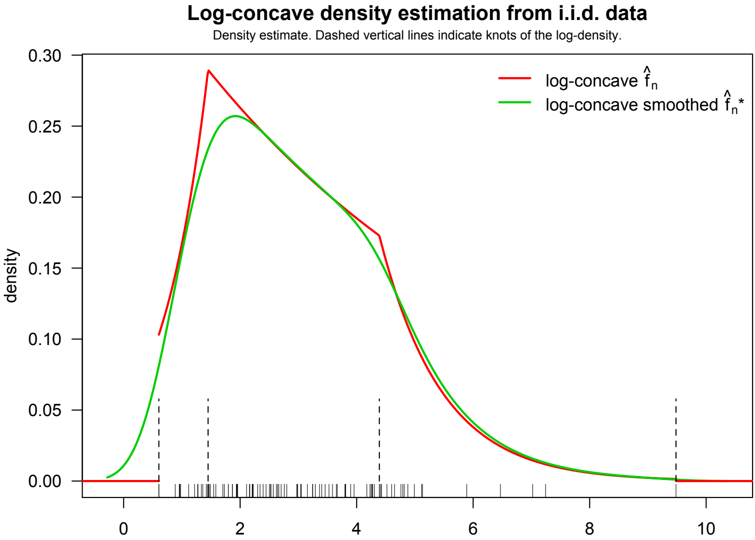
We have a concave maximization problem over a closed convex cone in $\mathbb{R}^n$.

n potentially large.

There is a fast active set algorithm by Dümbgen and Rufibach (2011), available in the R-package `logcondens`.

Simulation example

Sample 100 points from $\Gamma(3, 1)$.



Related work

- Multivariate case (Cule, Samworth and Stewart, 2010, JRSS B); R package `LogConcDEAD` (Cule, Gramacy and Samworth, 2009, J Stat. Soft.).
- General approximation theory with applications to additive regression models with log-concave error distribution (Dümbgen, Samworth and S, 2010, Ann. Stat.).
- Discrete case (Balabdaoui and Rufibach, preprint 2011).

Part 2: Log-concave density estimation for censored data

Censored data

For simplicity of notation: estimate densities on $\mathbb{R}_+$; think of distribution of time to a certain event.

Instead of exact values X_i we observe intervals $\widetilde{X}_i$ containing X_i. The following cases are possible:

- $\widetilde{X}_i = \{X_i\}$;
- $\widetilde{X}_i = (L_i, R_i]$ where $0 \leq L_i < R_i < \infty$;
- $\widetilde{X}_i = (L_i, \infty)$ where $L_i \geq 0$.

Write generally $L_i := \inf \widetilde{X}_i$, $R_i := \sup \widetilde{X}_i$.

We would like to estimate the "underlying distribution of the X_is". (It is not clear what this means without further specification.)

A concrete censoring model

Assume that the i-th individual is inspected at times T_{ij}, $1 \leq j \leq K_i$, where $0 =: T_{i0} < T_{i1} < T_{i2} < \ldots < T_{i,K_i}$.
If $K_i = \infty$, assume that $T_{i,\infty} := \sup_j(T_{ij}) = \infty$.

Let furthermore Y_{ij} be an indicator for exact observability in the j-th inter-inspection interval

$$I_{ij} := \begin{cases} (T_{ij}, T_{i,j+1}] & \text{if } 0 \leq j < K_i; \\ (T_{ij}, \infty) & \text{if } j = K_i. \end{cases}$$

Then the X_i translate into observations $\widetilde{X}_i$ as follows:

$$\widetilde{X}_i := \begin{cases} X_i & \text{if } X_i \in I_{ij} \text{ and } Y_{ij} = 1; \\ I_{ij} & \text{if } X_i \in I_{ij} \text{ and } Y_{ij} = 0. \end{cases}$$

Note: The K_i, T_{ij}, and Y_{ij} may be arbitrarily dependent random variables.

Special cases

Right-censoring: if event happens prior to a single inspection time, it is observed exactly, otherwise censored. $K_i \equiv 1$, $I_{i0} \equiv 1$, $I_{i1} \equiv 0$.

Current status model: it is only known whether event happened before or after a single inspection time. $K_i \equiv 1$, $I_{i0} \equiv I_{i1} \equiv 0$.

Mixed case interval-censoring: for each individual it is known in which of a number of contiguous intervals the event happened. K_i finite-valued, $I_{ij} \equiv 0$ for all j.

Rounding and binning: X_is are known up to rounding to closest integer (say). $K_i \equiv \infty$, $I_{ij} \equiv 0$, $T_{ij} \equiv j + 1/2$ for all i, j.

Conditional log-likelihood

Assume that conditionally on all K_i, T_{ij}, and Y_{ij} the random variables $X_1, \ldots, X_n$ are i.i.d. with density $f = \exp \circ \varphi$.

The normalized log-likelihood for φ is then

$$\begin{aligned}\ell(\varphi) &:= \ell(\varphi; \widetilde{X}_1, \ldots, \widetilde{X}_n) \\ &:= \frac{1}{n} \sum_{i=1}^{n} \left[1\{L_i = R_i\} \varphi(X_i) + 1\{L_i < R_i\} \log \left(\int_{L_i}^{R_i} \exp \varphi(x)\, dx \right) \right],\end{aligned}$$

where $L_i = \inf \widetilde{X}_i$, $R_i = \sup \widetilde{X}_i$.

Likelihood maximization

Maximize $\ell(\varphi)$ over all $\varphi \in \Phi$ that satisfy $\int \exp \varphi(x)\, dx = 1$.

Equivalently: Maximize

$$L(\varphi) = \frac{1}{n} \sum_{i=1}^{n} \Big[1\{L_i = R_i\} \varphi(X_i) + 1\{L_i < R_i\} \log \Big(\int_{L_i}^{R_i} \exp \varphi(x)\, dx \Big) \Big] - \int_0^{\infty} \exp \varphi(x)\, dx$$

over $\varphi \in \Phi$.

Shape requirements for maximizer $\widehat{\varphi}$

Let $0 \le \tau_1 < \tau_2 < \ldots < \tau_q \le \infty$ be the endpoints of the data intervals, i.e. $\{\tau_1, \ldots, \tau_q\} = \{L_i \colon 1 \le i \le n\} \cup \{R_i \colon 1 \le i \le n\}$.

Then any $\widehat{\varphi} \in \arg\max_{\varphi \in \Phi} L(\varphi)$ satisfies

- $[\min_i R_i, \max_i L_i] \ \subset \ \mathrm{dom}\big(\widehat{\varphi}\big) \ \subset \ [\tau_1, \tau_q]$
- $\widehat{\varphi}$ is linear on every interval $(\tau_j, \tau_{j+1}) \subset \mathrm{dom}(\widehat{\varphi}) \setminus \bigcup_{i=1}^n [L_i, R_i]$.

Shape simplifications for maximizer

Theorem (shape)

Suppose that $\arg\max_{\varphi\in\Phi} L(\varphi) \neq \emptyset$.

Then there exists a maximizer $\widehat{\varphi}$ *with* $\mathrm{dom}(\widehat{\varphi}) = [\tau_{j_1}, \tau_{j_2}] \cap \mathbb{R}_+$ *for some* j_1, j_2 *that is piecewise linear on* $\mathrm{dom}(\widehat{\varphi})$ *with at most one change of slope in any interval* (τ_j, τ_{j+1}).

There is no change of slope in

- *the two extreme intervals* $(\tau_{j_1}, \tau_{j_1+1})$ *and* $(\tau_{j_2-1}, \tau_{j_2})$;
- *the interval* $(\tau_{j_1+1}, \tau_{j_1+2})$ *unless there is an i such that* $L_i = R_i = \tau_{j_1}$, *and the interval* $(\tau_{j_2-2}, \tau_{j_2-1})$ *unless there is an i such that* $L_i = R_i = \tau_{j_2}$;
- *any interval* $(\tau_j, \tau_{j+1}) \subset \mathrm{dom}(\widehat{\varphi}) \setminus \bigcup_{i=1}^n [L_i, R_i]$.

Existence of maximizer

Theorem (existence)

Assume that there is ***no*** $i_o \in \{1, \ldots, n\}$ *with* $L_{i_o} = R_{i_o} = X_{i_o}$ *and*

$$\bigcap_{i=1}^{n} [L_i, R_i] = \{X_{i_o}\}.$$

Then

$$\operatorname*{arg\,max}_{\varphi \in \mathcal{G}_{\mathrm{conc}}} L(\varphi) \neq \emptyset,$$

where $\mathcal{G}_{\mathrm{conc}} \subset \Phi$ *consists of all piecewise linear functions with domain and changes of slope according to the shape theorem. We only allow slope changes on a fine grid* $t_1 < t_2 < \ldots < t_m$ *containing the finite* τ_j*.*

Counter-example

Suppose that there **is** an $i_o \in \{1, \ldots, n\}$ with $L_{i_o} = R_{i_o} = X_{i_o}$ and

$$\bigcap_{i=1}^{n} [L_i, R_i] = \{X_{i_o}\}.$$

Assuming that there are intervals to the left **and** to the right of $X_{i_o} =: \tau_{j_o}$ (otherwise idea is easily adapted), we can obtain an arbitrarily large log-likelihood, by defining φ as a triangle function with domain $[\tau_{j_o-1}, \tau_{j_o+1}]$ and peak in τ_{j_o}:

The integral of $\exp \circ \varphi$ over $[\tau_{j_o-1}, \tau_{j_o}]$ and $[\tau_{j_o}, \tau_{j_o+1}]$ can be kept constant, while $\varphi(\tau_{j_o})$ and hence $L(\varphi)$ go towards ∞.

Proof idea of existence theorem

Consider as a problem of maximizing $L(\varphi)$ over a convex cone in $[-\infty,\infty)^m \times [-\infty,0)$; $\varphi = (\varphi_1,\ldots,\varphi_m,\widetilde{\varphi}_{m+1})$, where $\varphi_i := \varphi(t_i)$ and $\widetilde{\varphi}_{m+1} := \varphi'(t_m+)$.

1. Log-likelihood is continuous on $\mathcal{G}_{\mathrm{conc}} \cap ([-\infty,k]^m \times [-\infty,-1/k])$ for every $k \in \mathbb{N}$, which is compact in extended Euclidean topology $\Longrightarrow$ $\max$ over this set exists.
2. Show $L(\varphi^{(k)}) \to -\infty$ for every sequence $(\varphi^{(k)})_k \in \mathcal{G}_{\mathrm{conc}}$ with $\max_{1\le i\le m+1} \varphi_i^{(k)} \to \infty$ by somewhat tedious calculations, where $\varphi_{m+1}^{(k)} := -\log(-\widetilde{\varphi}_{m+1}^{(k)}{}')$.

Assuming now that there is no maximizer in $\mathcal{G}_{\mathrm{conc}} \cap \big([-\infty,\infty)^m \times [-\infty,0)\big)$ leads to a contradiction.

Open problems

- uniqueness
- consistency and rates
- everything else . . .

Computation

Still maximization over cone in typically very high dimension (now m). But our log-likelihood function is not concave anymore!

We try an EM algorithm.

Remember:

$$\ell(\varphi) := \frac{1}{n}\sum_{i=1}^{n}\Big[1\{L_i = R_i\}\varphi(X_i) + 1\{L_i < R_i\}\log\Big(\int_{L_i}^{R_i}\exp\varphi(x)\,dx\Big)\Big].$$

“incomplete data log-likelihood” (normalized);

$$\ell^*(\varphi) := \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i)$$

“complete data log-likelihood” (normalized).

EM algorithm

E step: Given $\varphi_r \in \overline{\mathcal{G}}_{\mathrm{conc}}$ (log-densities in $\mathcal{G}_{\mathrm{conc}}$), compute

$$\begin{aligned}\tilde{\ell}^*(\varphi;\varphi_r) &:= \mathbb{E}_{\varphi_r}\big(\ell^*(\varphi)\,\big|\,X_i \in \widetilde{X}_i \text{ for all } i\big)\\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\varphi_r}\big(\varphi(X_i)\,\big|\,X_i \in \widetilde{X}_i\big)\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[1\{L_i = R_i\}\varphi(X_i) + 1\{L_i < R_i\}\frac{\mathbb{E}_{\varphi_r}\big(\varphi(X_i)\,1\{X_i \in \widetilde{X}_i\}\big)}{\mathbb{P}_{\varphi_r}\big(X_i \in \widetilde{X}_i\big)}\Big].\end{aligned}$$

Easy, since φ_r, φ piecewise linear.

M step: Maximize $\tilde{\ell}^*(\varphi;\varphi_r)$ over all $\varphi \in \overline{\mathcal{G}}_{\mathrm{conc}}$ with $\mathrm{dom}(\varphi) \subset \mathrm{dom}(\varphi_r)$. Domain will not get smaller. Call maximizer φ_{r+1}.

Reduction to Active Set Algorithm

$\mathbb{E}_{\varphi_r}\big(\varphi(X_i)\,1\{X_i \in [a,b]\}\big)$ is a linear combination of $\varphi(a)$ and $\varphi(b)$ if φ_r, φ are linear on $[a,b]$, $0 \le a \le b < \infty$.

$\mathbb{E}_{\varphi_r}\big(\varphi(X_i)\,1\{X_i \in [a,\infty)\}\big)$ is a linear combination of $\varphi(a)$ and $\varphi'(a+)$ if φ_r, φ are linear on $[a,\infty)$, $0 \le a < \infty$.

Therefore

$$\tilde{\ell}^*(\varphi;\varphi_r) = \sum_{j=1}^{m} w_j \varphi(t_j) + w_{m+1}\varphi'(t_m+)$$

with certain (computable) weights $w_1,\ldots,w_m > 0$; $w_{m+1} > 0$ if right endpoint ∞ appears in data intervals, $= 0$ otherwise.

Is weighted version of our exact data log-likelihood. We may use the original active set algorithm for m data points (with an extension for the right-hand slope)!

Domain reduction

EM algorithm starts on maximal domain $[\min_i L_i, \max_i R_i]$.

Problem: Domain is never reduced, but we may have started on too large domain. Algorithm forces φ_r towards $-\infty$ on extra domain (leading to convergence / numerical problems).

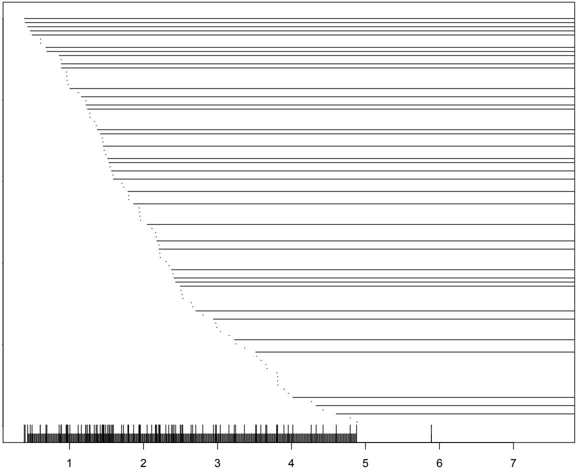
Solution: If $\int_{\tau_j}^{\tau_{j+1}} \exp \varphi(x)\, dx$ gets "too small" for some j (only possible at the very left or very right of current domain), the corresponding interval is removed from the domain and the EM algorithm is restarted on the reduced domain.

Three examples

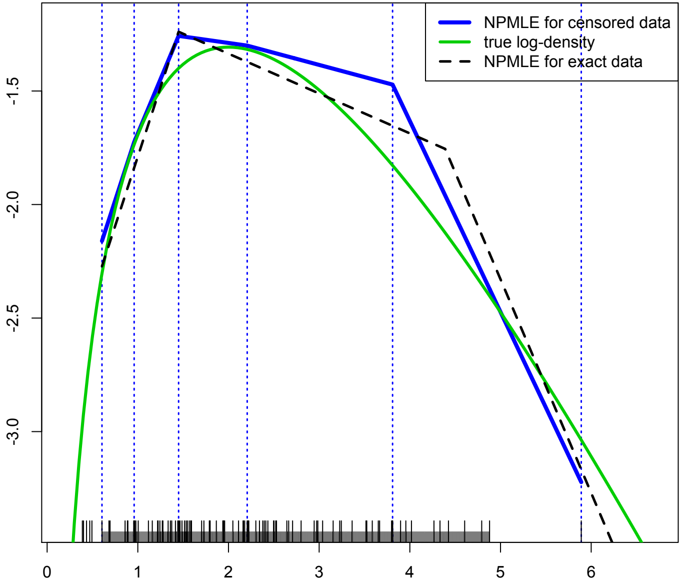
- (simulated) right-censored $\Gamma(3, 1)$-data
- (simulated) mixed-case interval-censored $\mathrm{Gumbel}(2, 1)$-data
- (real data) ovarian cancer cases

Example 1: Right-censoring

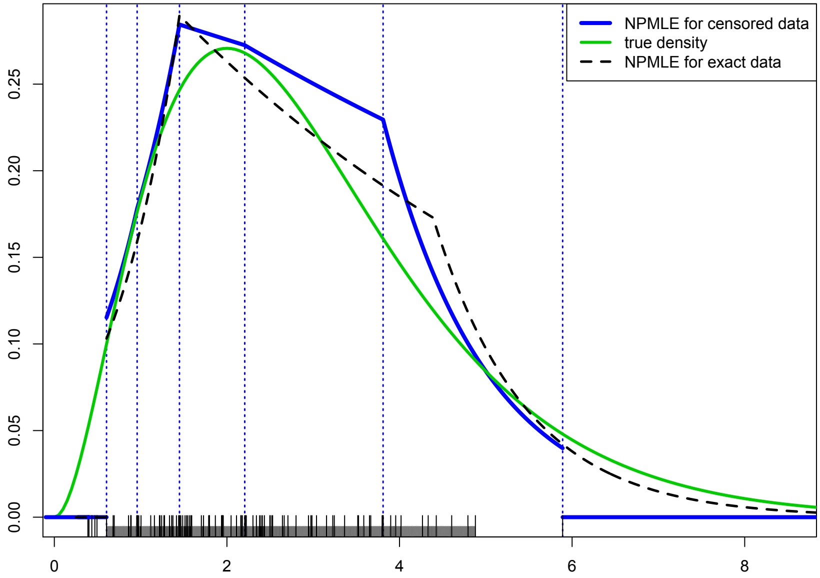
Consider the same 100 data points from $\Gamma(3, 1)$ as before, but now introduce censoring after individual $\Gamma(2, 1/2)$-distributed inspection times. Resulting data (37 censored):



Example 1: Log-density estimates

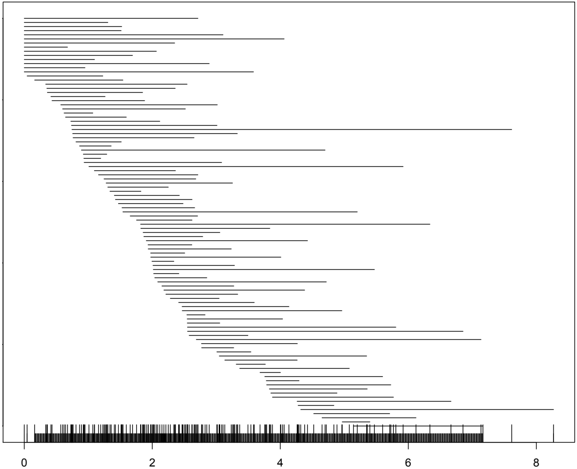


Example 1: Density estimates

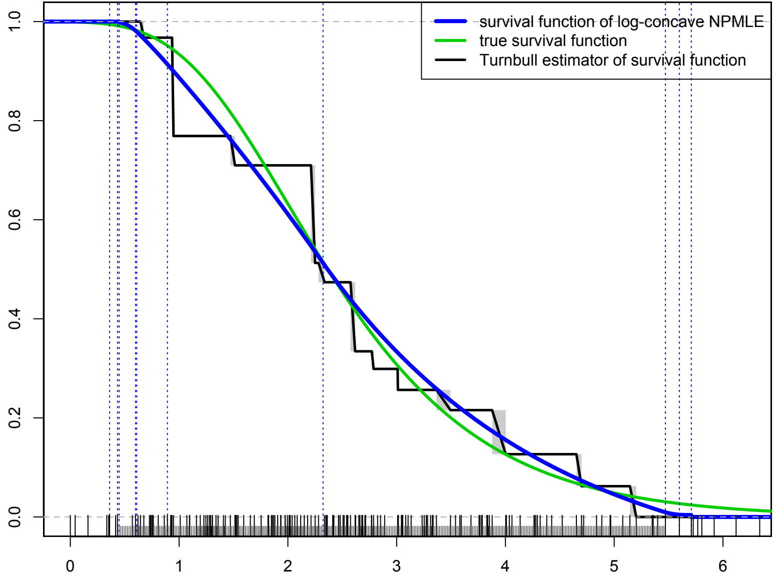


Example 2: Interval-censoring

100 data points from a $\mathrm{Gumbel}(2,1)$-distribution. Each individual is inspected according to a $\mathrm{Poisson}(1)$-process, i.e. each inspection takes place an $\mathrm{Exp}(1)$-distributed time after the last (all times independent)

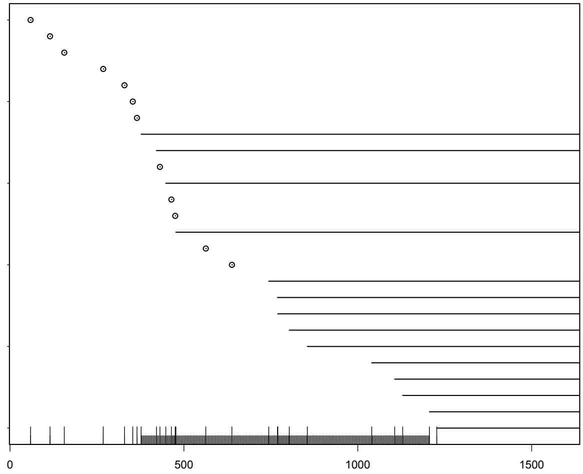


Example 2: Survival function estimates

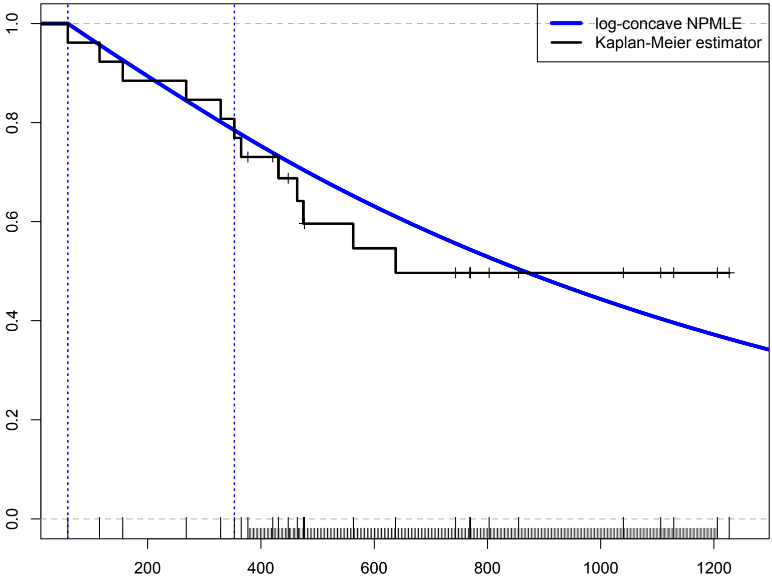


Example 3: Real data

We consider the dataset `ovarian` from the R-package `survival`. Survival times in days of 26 women with ovarian cancer (14 censored).



Example 3: Survival function estimates



Cure probability

What if we allow for the possibility that patients are cured, i.e. with a certain probability p_0 a woman will not die from cancer?

We model this by allowing subprobability densities of total mass $1 - p_0$ and adding a point mass of p_0 at time ∞.

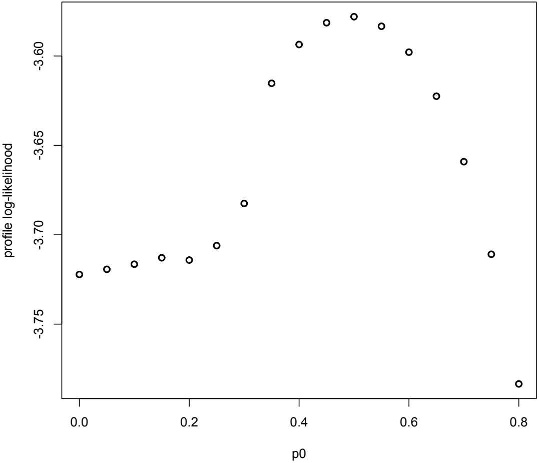
New log-likelihood:

$$\begin{aligned}\ell(\varphi, p_0) := \frac{1}{n}\sum_{i=1}^{n}\Bigl[& 1\{L_i = R_i\}\varphi(X_i) \\ & + 1\{L_i < R_i\}\log\Bigl(\int_{L_i}^{R_i} \exp\varphi(x)\,dx + p_0 1\{R_i = \infty\}\Bigr)\Bigr].\end{aligned}$$

For known p_0, nothing essential changes in our algorithm!

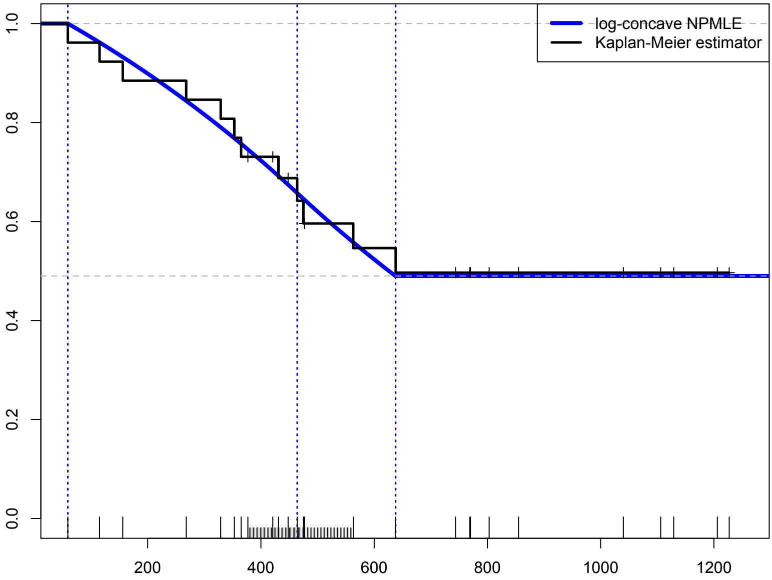
Profile likelihood maximization in p_0

Maximize $\ell_{\text{profile}}(p_0) := \max_{\varphi \in \overline{\mathcal{G}}_{\text{conc}}(p_0)} \ell(\varphi, p_0) \quad \rightsquigarrow \quad \widehat{p}_0$



In our example $\widehat{p}_0 \approx 0.49$. Compute $\widehat{\varphi} \in \arg\max_{\varphi \in \overline{\mathcal{G}}_{\text{conc}}(\widehat{p}_0)} \ell(\varphi, \widehat{p}_0)$.

Example 3: Survival function estimates



R package

Algorithm implemented in R package `logconcens`.

Give it a try!