

# Active Set and EM Algorithms for Log-Concave Densities Based on Complete and Censored Data

Lutz Dümbgen, André Hüsler and Kaspar Rufibach  
University of Bern

August 2007, revised March 2011

**Abstract.** We develop an active set algorithm for the maximum likelihood estimation of a log-concave density based on complete data. Building on this fast algorithm, we indicate an EM algorithm to treat arbitrarily censored or binned data.

## 1 Introduction

A probability density  $f$  on the real line is called log-concave if it may be written as

$$f(x) = \exp \phi(x)$$

for some concave function  $\phi : \mathbb{R} \rightarrow [-\infty, \infty)$ . The class of all log-concave densities provides an interesting nonparametric model consisting of unimodal densities and containing many standard parametric families; see Dümbgen and Rufibach (2009) for a more thorough overview.

This paper treats algorithmic aspects of maximum likelihood estimation for this particular class. In Section 2 we derive a general finite-dimensional optimization problem which is closely related to computing the maximum likelihood estimator of a log-concave probability density  $f$  based on independent, identically distributed observations. Section 3 is devoted to the latter optimization problem. At first we describe generally an active set algorithm, a useful tool from optimization theory (cf. Fletcher, 1987) with many potential applications in statistical computing. A key property of such algorithms is that they terminate after finitely many steps (in principle). Then we adapt this approach to our particular estimation problem, which yields an alternative to the iterative algorithms developed by Rufibach (2006, 2007) and Pal, Woodroffe and Meyer (2006). The resulting active set algorithm is similar in spirit to the vertex direction and support reduction algorithms described by Groeneboom, Jongbloed and Wellner (2008), who consider the special setting of mixture models.

In Section 4 we consider briefly the problem of estimating a probability distribution  $P$  on  $(0, \infty]$  based on censored or binned data. Censoring occurs quite frequently in biomedical applications, e.g.  $X$  being the time point when a person develops a certain disease or dies from a certain cause. Another field of application is quality control where  $X$  is the failure time of a certain object. A good reference for event time analysis is the monograph of Klein and Moeschberger (1997). Binning is typical in socioeconomic surveys, e.g. when persons or households are asked which of several given intervals their yearly income  $X$  falls into. We discuss maximum likelihood estimation of  $P$  under the assumption that it is absolutely continuous on  $(0, \infty)$  with log-concave probability density  $f$ . The resulting estimator is an alternative to those of Dümbgen et al. (2006). The latter authors restrict themselves to interval-censored data and considered the weaker constraints of  $f$  being non-increasing or unimodal. Introducing the stronger but still natural constraint of log-concavity allows us to treat arbitrarily censored data, similarly as Turnbull (1976). In Section 5 we indicate an expectation-maximization (EM) algorithm for the estimation of  $P$ , using the aforementioned active set algorithm as a building block. This approach is similar to Turnbull (1976) and Braun et al. (2005); the latter authors considered self-consistent kernel density estimators. For more information and references on EM and related algorithms in general we refer to Lange et al. (2000). A detailed description of our method for censored or binned data will be given elsewhere.

Section 6 contains most proofs and various auxiliary results.

## 2 The general log-likelihood function for complete data

**Independent, identically distributed observations.** Let  $X_1, X_2, \dots, X_n$  be independent random variables with log-concave probability density  $f = \exp \phi$  on  $\mathbb{R}$ . Then the normalized log-likelihood function is given by

$$\ell(\phi) := n^{-1} \sum_{i=1}^n \phi(X_i).$$

It may happen that due to rounding errors one observes  $\tilde{X}_i$  in place of  $X_i$ . In that case, let  $x_1 < x_2 < \dots < x_m$  be the different elements of  $\{\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n\}$  and define  $p_i := n^{-1} \#\{j : \tilde{X}_j = x_i\}$ . Then an appropriate surrogate for the normalized log-likelihood is

$$\ell(\phi) := \sum_{i=1}^m p_i \phi(x_i). \tag{1}$$

**The general log-likelihood function.** In what follows we consider the functional (1) for arbitrary given points  $x_1 < x_2 < \dots < x_m$  and probability weights  $p_1, p_2, \dots, p_m > 0$ , i.e.  $\sum_{i=1}^m p_i = 1$ . Suppose that we want to maximize  $\ell(\phi)$  over all functions  $\phi$  within a certain family  $\mathcal{F}$  of measurable functions from  $\mathbb{R}$  into  $[-\infty, \infty)$  satisfying the constraint  $\int \exp \phi(x) dx = 1$ . If  $\mathcal{F}$  is closed under addition of constants, i.e.  $\phi + c \in \mathcal{F}$  for arbitrary  $\phi \in \mathcal{F}$  and  $c \in \mathbb{R}$ , then one can easily show that maximizing  $\ell(\phi)$  over all  $\phi \in \mathcal{F}$  with  $\int \exp \phi(x) dx = 1$  is equivalent to maximizing

$$L(\phi) := \sum_{i=1}^m p_i \phi(x_i) - \int \exp \phi(x) dx$$

over the whole family  $\mathcal{F}$ ; see also Silverman (1982, Theorem 3.1).

**Restricting the set of candidate functions.** The preceding considerations apply in particular to the family  $\mathcal{F}$  of all concave functions. Now let  $\mathcal{G}$  be the set of all continuous functions  $\psi : [x_1, x_m] \rightarrow \mathbb{R}$  which are linear on each interval  $[x_k, x_{k+1}]$ ,  $1 \leq k < m$ , and we define  $\psi := -\infty$  on  $\mathbb{R} \setminus [x_1, x_m]$ . Moreover, let  $\mathcal{G}_{\text{conc}}$  be the set of all concave functions within  $\mathcal{G}$ . For any  $\phi \in \mathcal{F}$  with  $L(\phi) > -\infty$  let  $\psi$  be the unique function in  $\mathcal{G}_{\text{conc}}$  such that  $\psi = \phi$  on  $\{x_1, x_2, \dots, x_m\}$ . Then it follows from concavity of  $\phi$  that  $\psi \leq \phi$  pointwise, and  $L(\psi) \geq L(\phi)$ . Equality holds if, and only if,  $\psi = \phi$ . Thus maximizing  $L$  over the class  $\mathcal{F}$  is equivalent to its maximization over  $\mathcal{G}_{\text{conc}}$ .

**Properties of  $L(\cdot)$ .** For explicit calculations it is useful to rewrite  $L(\psi)$  as follows: Any function  $\psi \in \mathcal{G}$  may be identified with the vector  $\boldsymbol{\psi} := (\psi(x_i))_{i=1}^m \in \mathbb{R}^m$ . Likewise, any vector  $\boldsymbol{\psi} \in \mathbb{R}^m$  defines a function  $\psi \in \mathcal{G}$  via

$$\psi(x) := \left(1 - \frac{x - x_k}{\delta_k}\right) \psi_k + \frac{x - x_k}{\delta_k} \psi_{k+1} \quad \text{for } x \in [x_k, x_{k+1}], 1 \leq k < m,$$

where  $\delta_k := x_{k+1} - x_k$ . Then one may write

$$L(\psi) = L(\boldsymbol{\psi}) := \sum_{i=1}^m p_i \psi_i - \sum_{k=1}^{m-1} \delta_k J(\psi_k, \psi_{k+1})$$

with

$$J(r, s) := \int_0^1 \exp((1-t)r + ts) dt$$

for arbitrary  $r, s \in \mathbb{R}$ . The latter function  $J : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  is infinitely often differentiable and strictly convex. Hence  $L(\cdot)$  is an infinitely often differentiable and strictly concave functional on

$\mathbb{R}^m$ . In addition it is coercive in the sense that

$$L(\psi) \rightarrow -\infty \quad \text{as } \|\psi\| \rightarrow \infty. \quad (2)$$

This entails that both

$$\tilde{\psi} := \operatorname{argmax}_{\psi \in \mathcal{G}} L(\psi) \quad \text{and} \quad (3)$$

$$\hat{\psi} := \operatorname{argmax}_{\psi \in \mathcal{G}_{\text{conc}}} L(\psi) \quad (4)$$

are well defined and unique.

Let us discuss some further properties of  $L(\cdot)$  and its unrestricted maximizer  $\tilde{\psi}$ . To maximize  $L(\cdot)$  we need its Taylor expansion of second order. In fact, for functions  $\psi, v \in \mathcal{G}$ ,

$$\left. \frac{d}{dt} \right|_{t=0} L(\psi + tv) = \sum_{i=1}^m p_i v(x_i) - \int v(x) \exp \psi(x) dx, \quad (5)$$

$$\left. \frac{d^2}{dt^2} \right|_{t=0} L(\psi + tv) = - \int v(x)^2 \exp \psi(x) dx. \quad (6)$$

Note that the latter expression yields an alternative proof of  $L$ 's strict concavity. Explicit formulae for the gradient and hessian matrix of  $L$  as a functional on  $\mathbb{R}^m$  are given in Section 6, and with these tools one can easily compute  $\tilde{\psi}$  very precisely via Newton type algorithms. We end this section with a characterization and interesting properties of the maximizer  $\tilde{\psi}$ . In what follows let

$$J_{ab}(r, s) := \frac{\partial^{a+b}}{\partial r^a \partial s^b} J(r, s) = \int_0^1 (1-t)^a t^b \exp((1-t)r + ts) dt.$$

for nonnegative integers  $a$  and  $b$ .

**Theorem 2.1** *Let  $\psi \in \mathcal{G}$  with corresponding density  $f(x) := \exp \psi(x)$  and distribution function  $F(r) := \int_{x_1}^r f(x) dx$  on  $[x_1, x_m]$ . The function  $\psi$  maximizes  $L$  if, and only if, its distribution function  $F$  satisfies*

$$F(x_m) = 1 \quad \text{and} \quad \delta_k^{-1} \int_{x_k}^{x_{k+1}} F(x) dx = \sum_{i=1}^k p_i \quad \text{for } 1 \leq k < m.$$

In that case,

$$\int_{x_1}^{x_m} x f(x) dx = \sum_{i=1}^m p_i x_i$$

and

$$\int_{x_1}^{x_m} x^2 f(x) dx = \sum_{i=1}^m p_i x_i^2 - \sum_{k=1}^{m-1} \delta_k^3 J_{11}(\psi_k, \psi_{k+1}).$$

**Some auxiliary formulae.** For  $\psi \in \mathcal{G}$  with density  $f(x) := \exp \psi(x)$  and distribution function  $F(r) := \int_{x_1}^r f(x) dx$  on  $[x_1, x_m]$ , one can easily derive explicit expressions for  $F$  and the first two moments of  $f$  in terms of  $J(\cdot, \cdot)$  and its partial derivatives: For  $1 \leq k < m$ ,

$$F(x_{k+1}) = \sum_{i=1}^k \delta_i J(\psi_i, \psi_{i+1})$$

and

$$\delta_k^{-1} \int_{x_k}^{x_{k+1}} F(x) dx = F(x_k) + \delta_k J_{10}(\psi_k, \psi_{k+1}) \in (F(x_k), F(x_{k+1})).$$

Moreover, for any  $a \in \mathbb{R}$ ,

$$\begin{aligned} \int_{x_1}^{x_m} (x-a) f(x) dx &= \sum_{k=1}^{m-1} \delta_k ((x_k - a) J_{10}(\psi_k, \psi_{k+1}) + (x_{k+1} - a) J_{01}(\psi_k, \psi_{k+1})), \\ \int_{x_1}^{x_m} (x-a)^2 f(x) dx &= \sum_{k=1}^{m-1} \delta_k ((x_k - a)^2 J_{10}(\psi_k, \psi_{k+1}) + (x_{k+1} - a)^2 J_{01}(\psi_k, \psi_{k+1})) \\ &\quad - \sum_{k=1}^{m-1} \delta_k^3 J_{11}(\psi_k, \psi_{k+1}). \end{aligned}$$

### 3 An active set algorithm

#### 3.1 The general principle

We consider an arbitrary continuous and concave function  $L : \mathbb{R}^m \rightarrow [-\infty, \infty)$  which is coercive in the sense of (2) and continuously differentiable on the set  $\text{dom}(L) := \{\boldsymbol{\psi} \in \mathbb{R}^m : L(\boldsymbol{\psi}) > -\infty\}$ . Our goal is to maximize  $L$  on the closed convex set

$$\mathcal{K} := \left\{ \boldsymbol{\psi} \in \mathbb{R}^m : \mathbf{v}_i^\top \boldsymbol{\psi} \leq c_i \text{ for } i = 1, \dots, q \right\},$$

where  $\mathbf{v}_1, \dots, \mathbf{v}_q$  are nonzero vectors in  $\mathbb{R}^m$  and  $c_1, \dots, c_q$  real numbers such that  $\mathcal{K} \cap \text{dom}(L) \neq \emptyset$ . These assumptions entail that the set

$$\mathcal{K}_* := \operatorname{argmax}_{\boldsymbol{\psi} \in \mathcal{K}} L(\boldsymbol{\psi})$$

is a nonvoid and compact subset of  $\text{dom}(L)$ . For simplicity we shall assume that

$$\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_q \text{ are linearly independent,} \tag{7}$$

but see also the possible extensions indicated at the end of this section.

An essential tacit assumption is that for any index set  $A \subseteq \{1, \dots, q\}$  and the corresponding affine subspace

$$\mathcal{V}(A) := \left\{ \boldsymbol{\psi} \in \mathbb{R}^m : \mathbf{v}_a^\top \boldsymbol{\psi} = c_a \text{ for all } a \in A \right\}$$

of  $\mathbb{R}^m$ , we have an algorithm computing a point

$$\tilde{\psi}(A) \in \mathcal{V}_*(A) := \operatorname{argmax}_{\psi \in \mathcal{V}(A)} L(\psi),$$

provided that  $\mathcal{V}(A) \cap \operatorname{dom}(L) \neq \emptyset$ . Now the idea is to vary  $A$  suitably until, after finitely many steps,  $\tilde{\psi}(A)$  belongs to  $\mathcal{K}_*$ .

In what follows we attribute to any vector  $\psi \in \mathbb{R}^m$  the index set

$$A(\psi) := \left\{ i \in \{1, \dots, q\} : \mathbf{v}_i^\top \psi \geq c_i \right\}.$$

For  $\psi \in \mathcal{K}$  the set  $A(\psi)$  identifies the “active constraints” for  $\psi$ . The following theorem provides useful characterizations of  $\mathcal{K}_*$  and  $\mathcal{V}_*(A)$ .

**Theorem 3.1** *Let  $\mathbf{b}_1, \dots, \mathbf{b}_m$  be a basis of  $\mathbb{R}^m$  such that*

$$\mathbf{v}_i^\top \mathbf{b}_j \begin{cases} < 0 & \text{if } i = j \leq q, \\ = 0 & \text{else.} \end{cases}$$

(a) *A vector  $\psi \in \mathcal{K} \cap \operatorname{dom}(L)$  belongs to  $\mathcal{K}_*$  if, and only if,*

$$\mathbf{b}_i^\top \nabla L(\psi) \begin{cases} = 0 & \text{for all } i \in \{1, \dots, m\} \setminus A(\psi), \\ \leq 0 & \text{for all } i \in A(\psi). \end{cases} \quad (8)$$

(b) *For any given set  $A \subseteq \{1, \dots, q\}$ , a vector  $\psi \in \mathcal{V}(A) \cap \operatorname{dom}(L)$  belongs to  $\mathcal{V}_*(A)$  if, and only if,*

$$\mathbf{b}_i^\top \nabla L(\psi) = 0 \quad \text{for all } i \in \{1, \dots, m\} \setminus A. \quad (9)$$

The characterizations in this theorem entail that any vector  $\psi \in \mathcal{K}_*$  belongs to  $\mathcal{V}_*(A(\psi))$ . The active set algorithm performs one of the following two procedures alternately:

**Basic procedure 1: Replacing a feasible point with a “conditionally” optimal one.** Let  $\psi$  be an arbitrary vector in  $\mathcal{K} \cap \operatorname{dom}(L)$ . Our goal is to find a vector  $\psi_{\text{new}}$  such that

$$L(\psi_{\text{new}}) \geq L(\psi) \quad \text{and} \quad \psi_{\text{new}} \in \mathcal{K} \cap \mathcal{V}_*(A(\psi_{\text{new}})). \quad (10)$$

To this end, set  $A := A(\psi)$  and define the candidate vector  $\psi_{\text{cand}} := \tilde{\psi}(A)$ . By construction,  $L(\psi_{\text{cand}}) \geq L(\psi)$ . If  $L(\psi_{\text{cand}}) = L(\psi)$ , we set  $\psi_{\text{new}} := \psi$ . If  $L(\psi_{\text{cand}}) > L(\psi)$  and  $\psi_{\text{cand}} \in \mathcal{K}$ , we set  $\psi_{\text{new}} := \psi_{\text{cand}}$ . Here (10) is satisfied, because  $A(\psi_{\text{new}}) \supseteq A(\psi)$ , so that  $\mathcal{V}(A(\psi_{\text{new}})) \subseteq \mathcal{V}(A)$ . Finally, if  $L(\psi_{\text{cand}}) > L(\psi)$  but  $\psi_{\text{cand}} \notin \mathcal{K}$ , let

$$\begin{aligned} t = t(\psi, \psi_{\text{cand}}) &:= \max\{t \in (0, 1) : (1-t)\psi + t\psi_{\text{cand}} \in \mathcal{K}\} \\ &= \min\left\{ \frac{c_i - \mathbf{v}_i^\top \psi}{\mathbf{v}_i^\top \psi_{\text{cand}} - \mathbf{v}_i^\top \psi} : 1 \leq i \leq q, \mathbf{v}_i^\top \psi_{\text{cand}} > c_i \right\}. \end{aligned} \quad (11)$$

Then we replace  $\boldsymbol{\psi}$  with  $(1-t)\boldsymbol{\psi} + t\boldsymbol{\psi}_{\text{cand}}$ . Note that  $L(\boldsymbol{\psi})$  does not decrease in this step, due to concavity of  $L$ . Moreover, the set  $A(\boldsymbol{\psi})$  increases strictly. Hence, repeating the preceding manipulations at most  $q$  times yields finally a vector  $\boldsymbol{\psi}_{\text{new}}$  satisfying (10), because  $\mathcal{V}(\{1, \dots, q\})$  is clearly a subset of  $\mathcal{K}$ . With the new vector  $\boldsymbol{\psi}_{\text{new}}$  we perform the second basic procedure.

**Basic procedure 2: Altering the set of active constraints.** Let  $\boldsymbol{\psi} \in \mathcal{K} \cap \text{dom}(L) \cap \mathcal{V}_*(A)$  with  $A = A(\boldsymbol{\psi})$ . It follows from Theorem 3.1 that  $\boldsymbol{\psi}$  belongs to  $\mathcal{K}_*$  if, and only if,

$$\mathbf{b}_a^\top \nabla L(\boldsymbol{\psi}) \leq 0 \quad \text{for all } a \in A.$$

Now suppose that the latter condition is violated, and let  $a_o = a_o(\boldsymbol{\psi})$  be an index in  $A$  such that  $\mathbf{b}_{a_o}^\top \nabla L(\boldsymbol{\psi})$  is maximal. Then  $\boldsymbol{\psi} + t\mathbf{b}_{a_o} \in \mathcal{K}$  and  $A(\boldsymbol{\psi} + t\mathbf{b}_{a_o}) = A \setminus \{a_o\}$  for arbitrary  $t > 0$ , while  $L(\boldsymbol{\psi} + t\mathbf{b}_{a_o}) > L(\boldsymbol{\psi})$  for sufficiently small  $t > 0$ . Thus we consider the vector  $\boldsymbol{\psi}_{\text{cand}} := \tilde{\boldsymbol{\psi}}(A \setminus \{a_o\})$ , which satisfies necessarily the inequality  $L(\boldsymbol{\psi}_{\text{cand}}) > L(\boldsymbol{\psi})$ . It may fail to be in  $\mathcal{K}$ , but it satisfies the inequality

$$\mathbf{v}_{a_o}^\top \boldsymbol{\psi}_{\text{cand}} > c_{a_o}.$$

For  $\boldsymbol{\psi}_{\text{cand}} - \boldsymbol{\psi}$  may be written as  $\lambda_{a_o} \mathbf{b}_{a_o} + \sum_{i \notin A} \lambda_i \mathbf{b}_i$  with real coefficients  $\lambda_1, \dots, \lambda_m$ , and

$$0 < (\boldsymbol{\psi}_{\text{cand}} - \boldsymbol{\psi})^\top \nabla L(\boldsymbol{\psi}) = \lambda_{a_o} \mathbf{b}_{a_o}^\top \nabla L(\boldsymbol{\psi})$$

according to (9). Hence  $0 < \lambda_{a_o} = \mathbf{v}_{a_o}^\top (\boldsymbol{\psi}_{\text{cand}} - \boldsymbol{\psi}) = \mathbf{v}_{a_o}^\top \boldsymbol{\psi}_{\text{cand}} - c_{a_o}$ . If  $\boldsymbol{\psi}_{\text{cand}} \in \mathcal{K}$ , we repeat this procedure with  $\boldsymbol{\psi}_{\text{cand}}$  in place of  $\boldsymbol{\psi}$ . Otherwise, we replace  $\boldsymbol{\psi}$  with  $(1-t)\boldsymbol{\psi} + t\boldsymbol{\psi}_{\text{cand}}$ , where  $t = t(\boldsymbol{\psi}, \boldsymbol{\psi}_{\text{cand}}) > 0$  is defined in (11), which results in a strictly larger value of  $L(\boldsymbol{\psi})$ . Then we perform the first basic procedure.

**The complete algorithm and its validity.** Often one knows a vector  $\boldsymbol{\psi}_o \in \mathcal{K} \cap \text{dom}(L)$  in advance. Then the active set algorithm can be started with the first basic procedure and proceeds as indicated in Table 1. In other applications it is sometimes obvious that  $\mathcal{V}(\{1, \dots, q\})$ , which is clearly a subset of  $\mathcal{K}$ , contains a point in  $\text{dom}(L)$ . In that case the input vector  $\boldsymbol{\psi}_o$  is superfluous, and the first twelve lines in Table 1 may be simplified as indicated in Table 2. The latter approach with starting point  $\boldsymbol{\psi}_o = \tilde{\boldsymbol{\psi}}(\{1, \dots, q\})$  may be numerically unstable, presumably when this starting point is very far from the optimum. In the special settings of concave least squares regression or log-concave density estimation, a third variant turned out to be very reliable: We

start with  $A = \emptyset$  and  $\psi_o = \tilde{\psi}(A)$ . As long as  $\psi_o \notin \mathcal{K}$ , we replace  $A$  with the larger set  $A(\psi_o)$  and recompute  $\psi_o = \tilde{\psi}(A)$ ; see Table 3.

In Table 1, the lines marked with (\*) and (\*\*) correspond to the end of the first basic procedure. At this stage,  $\psi$  is a vector in  $\mathcal{K} \cap \text{dom}(L) \cap \mathcal{V}_*(A(\psi))$ . Moreover, whenever the point (\*\*) is reached, the value  $L(\psi)$  is strictly larger than previously and equal to the maximum of  $L$  over the set  $\mathcal{V}(A)$ . Since there are only finitely many different sets  $A \subseteq \{1, \dots, q\}$ , the algorithm terminates after finitely many steps, and the resulting  $\psi$  belongs to  $\mathcal{K}$  by virtue of Theorem 3.1.

When implementing these algorithms one has to be aware of numerical inaccuracies and errors, in particular, if the algorithm  $\tilde{\psi}(\cdot)$  yields only approximations of vectors in  $\mathcal{V}_*(\cdot)$ . In our specific applications we avoided endless loops by replacing the conditions “ $\mathbf{b}_a^\top \nabla L(\psi) < 0$ ” and “ $\mathbf{v}_i^\top \psi > c_i$ ” with “ $\mathbf{b}_a^\top \nabla L(\psi) < -\epsilon$ ” and “ $\mathbf{v}_i^\top \psi > c_i + \epsilon$ ”, respectively, for some small constant  $\epsilon > 0$ .

```

Algorithm  $\psi \leftarrow \text{ActiveSet1}(L, \tilde{\psi}(\cdot), \psi_o)$ 
 $\psi \leftarrow \psi_o$ 
 $A \leftarrow A(\psi)$ 
 $\psi_{\text{cand}} \leftarrow \tilde{\psi}(A)$ 
while  $\psi_{\text{cand}} \notin \mathcal{K}$  do
     $\psi \leftarrow (1 - t(\psi, \psi_{\text{cand}}))\psi + t(\psi, \psi_{\text{cand}})\psi_{\text{cand}}$ 
     $A \leftarrow A(\psi)$ 
     $\psi_{\text{cand}} \leftarrow \tilde{\psi}(A)$ 
end while
 $\psi \leftarrow \psi_{\text{cand}}$ 
 $A \leftarrow A(\psi)$  (*)
while  $\max_{a \in A} \mathbf{b}_a^\top \nabla L(\psi) > 0$  do
     $a \leftarrow \min(\arg\max_{a \in A} \mathbf{b}_a^\top \nabla L(\psi))$ 
     $A \leftarrow A \setminus \{a\}$ 
     $\psi_{\text{cand}} \leftarrow \tilde{\psi}(A)$ 
    while  $\psi_{\text{cand}} \notin \mathcal{K}$  do
         $\psi \leftarrow (1 - t(\psi, \psi_{\text{cand}}))\psi + t(\psi, \psi_{\text{cand}})\psi_{\text{cand}}$ 
         $A \leftarrow A(\psi)$ 
         $\psi_{\text{cand}} \leftarrow \tilde{\psi}(A)$ 
    end while
     $\psi \leftarrow \psi_{\text{cand}}$ 
     $A \leftarrow A(\psi)$  (**)
end while.

```

Table 1: Pseudo-code of an active set algorithm.

**Possible extension I.** The assumption of linearly independent vectors  $\mathbf{v}_1, \dots, \mathbf{v}_q$  has been made for convenience and could be relaxed of course. In particular, one can extend the previous consid-



```

Algorithm  $\psi \leftarrow \mathbf{ActiveSet2}(L, \tilde{\psi}(\cdot))$ 
 $\psi \leftarrow \tilde{\psi}(\{1, \dots, q\})$ 
 $A \leftarrow \{1, \dots, q\}$ 
while  $\max_{a \in A} \mathbf{b}_a^\top \nabla L(\psi) > 0$  do
    ...
end while.

```

Table 2: Pseudo-code of first modified active set algorithm.

```

Algorithm  $\psi \leftarrow \mathbf{ActiveSet3}(L, \tilde{\psi}(\cdot))$ 
 $\psi \leftarrow \tilde{\psi}(\emptyset)$ 
while  $\psi \notin \mathcal{K}$  do
     $A \leftarrow A(\psi)$ 
     $\psi \leftarrow \tilde{\psi}(A)$ 
end while
 $A \leftarrow A(\psi)$ 
while  $\max_{a \in A} \mathbf{b}_a^\top \nabla L(\psi) > 0$  do
    ...
end while.

```

Table 3: Pseudo-code of second modified active set algorithm.

erations easily to the situation where  $\mathcal{K}$  consists of all vectors  $\psi \in \mathbb{R}^m$  such that

$$c_{i,1} \leq \mathbf{v}_i^\top \psi \leq c_{i,2}$$

for  $1 \leq i \leq q$  with numbers  $-\infty \leq c_{i,1} < c_{i,2} < \infty$ .

**Possible extension II.** Again we drop assumption (7) but assume that  $c_1 = \dots = c_q = 0$ , so that  $\mathcal{K}$  is a closed convex cone. Suppose further that we know a finite set  $\mathcal{E}$  of generators of  $\mathcal{K}$ , i.e. every vector  $\psi \in \mathcal{K}$  may be written as

$$\psi = \sum_{e \in \mathcal{E}} \lambda_e e$$

with numbers  $\lambda_e \geq 0$ . In that case, a point  $\psi \in \mathcal{K} \cap \text{dom}(L)$  belongs to  $\mathcal{K}_*$  if, and only if,

$$\nabla L(\psi)^\top \psi = 0 \quad \text{and} \quad \max_{e \in \mathcal{E}} \nabla L(\psi)^\top e \leq 0. \quad (12)$$

Now we can modify our basic procedure 2 as follows: Let  $\psi \in \mathcal{K} \cap \text{dom}(L) \cap \mathcal{V}(A)$  with  $A := A(\psi)$ . If (12) is violated, let  $e(\psi) \in \mathcal{E}$  such that  $\nabla L(\psi)^\top e(\psi) > 0$ . Further let  $s(\psi), t(\psi) > 0$  such that  $\psi_{\text{new}} := s(\psi)\psi + t(\psi)e(\psi) \in \mathcal{K}$  satisfies  $L(\psi_{\text{new}}) > L(\psi)$ . Then we replace  $\psi$  with  $\psi_{\text{new}}$  and perform the first basic procedure.

### 3.2 The special case of fitting log-concave densities

Going back to our original problem, note that  $\psi \in \mathcal{G}$  lies within  $\mathcal{G}_{\text{conc}}$  if, and only if, the corresponding vector  $\boldsymbol{\psi}$  satisfies

$$\frac{\psi_{j+1} - \psi_j}{\delta_j} - \frac{\psi_j - \psi_{j-1}}{\delta_{j-1}} = \mathbf{v}_j^\top \boldsymbol{\psi} \leq 0 \quad \text{for } j = 2, \dots, m-1, \quad (13)$$

where  $\mathbf{v}_j = (v_{i,j})_{i=1}^m$  has exactly three nonzero components:

$$v_{j-1,j} := 1/\delta_{j-1}, \quad v_{j,j} := -(\delta_{j-1} + \delta_j)/(\delta_{j-1}\delta_j), \quad v_{j+1,j} := 1/\delta_j.$$

Note that we changed the notation slightly by numbering the  $m-2$  constraint vectors from 2 to  $m-1$ . This is convenient, because then  $\mathbf{v}_j^\top \boldsymbol{\psi} \neq 0$  is equivalent to the corresponding function  $\psi \in \mathcal{G}$  changing slope at  $x_j$ . Suitable basis vectors  $\mathbf{b}_i$  are given, for instance, by  $\mathbf{b}_1 := (1)_{i=1}^m$ ,  $\mathbf{b}_m := (x_i)_{i=1}^m$  and

$$\mathbf{b}_j = (\min(x_i - x_j, 0))_{i=1}^m, \quad 2 \leq j < m.$$

For this particular problem it is convenient to rephrase the active set method in terms of *inactive* constraints, i.e. true *knots* of functions in  $\mathcal{G}$ . Throughout let  $I = \{i(1), \dots, i(k)\}$  be a subset of  $\{1, 2, \dots, m\}$  with  $k \geq 2$  elements  $1 = i(1) < \dots < i(k) = m$ , and let  $\mathcal{G}(I)$  be the set of all functions  $\psi \in \mathcal{G}$  which are linear on all intervals  $[x_{i(s)}, x_{i(s+1)}]$ ,  $1 \leq s < k$ . This set corresponds to  $\mathcal{V}(A)$  with  $A := \{1, \dots, m\} \setminus I$ . A function  $\psi \in \mathcal{G}(I)$  is uniquely determined by the vector  $(\psi(x_{i(s)}))_{s=1}^k$ , and one may write

$$L(\psi) = \sum_{s=1}^k p_s(I) \psi(x_{i(s)}) - \sum_{s=1}^{k-1} (x_{i(s+1)} - x_{i(s)}) J(\psi(x_{i(s)}), \psi(x_{i(s+1)}))$$

with suitable probability weights  $p_1(I), \dots, p_k(I) > 0$ . Precisely, writing

$$\psi(x) = \frac{x_{i(s+1)} - x}{x_{i(s+1)} - x_{i(s)}} \psi(x_{i(s)}) + \frac{x - x_{i(s)}}{x_{i(s+1)} - x_{i(s)}} \psi(x_{i(s+1)})$$

for  $1 \leq s < k$  and  $x_{i(s)} \leq x \leq x_{i(s+1)}$  yields the explicit formulae

$$\begin{aligned} p_1(I) &= \sum_{i=1}^{i(2)-1} \frac{x_{i(2)} - x_i}{x_{i(2)} - x_1} p_i, \\ p_s(I) &= \sum_{i=i(s-1)+1}^{i(s+1)-1} \min\left(\frac{x_i - x_{i(s-1)}}{x_{i(s)} - x_{i(s-1)}}, \frac{x_{i(s+1)} - x_i}{x_{i(s+1)} - x_{i(s)}}\right) p_i \quad \text{for } 2 \leq s < k, \\ p_k(I) &= \sum_{i=i(k-1)+1}^m \frac{x_i - x_{i(k-1)}}{x_m - x_{i(k-1)}} p_i. \end{aligned}$$

Consequently, the computation of  $\tilde{\psi}$  or  $\tilde{\psi}^{(I)} := \operatorname{argmax}_{\psi \in \mathcal{G}(I)} L(\psi)$  are optimization problems of the same type.

Since the vectors  $\mathbf{b}_2, \dots, \mathbf{b}_m$  correspond to the functions  $\Delta_2, \dots, \Delta_m$  in  $\mathcal{G}$  with

$$\Delta_j(x) := \min(x - x_j, 0), \quad (14)$$

checking the inequality  $\mathbf{b}_a^\top \nabla L(\psi) \leq 0$  for  $a \in A$  amounts to checking whether the directional derivative

$$H_j(\psi) := \sum_{i=1}^m p_i \Delta_j(x_i) - \int_{x_1}^{x_m} \Delta_j(x) \exp \psi(x) dx \quad (15)$$

is nonpositive for all  $j \in \{1, \dots, m\} \setminus I$ . If  $\psi = \psi^{(I)}$  and  $j \notin I$ , the inequality  $H_j(\psi) > 0$  means that  $L(\psi)$  could be increased strictly by allowing an additional knot at  $x_j$ .

**Example 3.2** Figure 1 shows the empirical distribution function of  $n = 25$  simulated random variables from a Gumbel distribution, while the smooth distribution function is the estimator  $\hat{F}(r) := \int_{-\infty}^r \exp \hat{\psi}(x) dx$ . Figure 2 illustrates the computation of the log-density  $\hat{\psi}$  itself. Each picture shows the current function  $\psi$  together with the new candidate function  $\psi_{\text{cand}}$ . We followed the algorithm in Table 2, so the first (upper left) picture shows the starting point, a linear function  $\psi$  on  $[x_1, x_{25}]$ , together with  $\psi_{\text{cand}}$  having an additional knot in  $(x_1, x_{25})$ . Since  $\psi_{\text{cand}}$  is concave, it becomes the new function  $\psi$  shown in the second (upper right) plot. In the third (lower left) plot one sees the situation where adding another knot resulted in a non-concave function  $\psi_{\text{cand}}$ . So the current function  $\psi$  was replaced with a convex combination of  $\psi$  and  $\psi_{\text{cand}}$ . The latter new function  $\psi$  and the almost identical final fit  $\hat{\psi}$  are depicted in the fourth (lower right) plot.

## 4 Censored or binned data

In the current and the next section we consider independent random variables  $X_1, X_2, \dots, X_n$  with unknown distribution  $P$  on  $(0, \infty]$  having sub-probability density  $f = \exp \phi$  on  $(0, \infty)$ , where  $\phi$  is concave and upper semicontinuous. In many applications the observations  $X_i$  are not completely available. For instance, let the  $X_i$  be event times for  $n$  individuals in a biomedical study, where  $X_i = \infty$  means that the event in question does not happen at all. If the study ends at time  $c_i > 0$  from the  $i$ -th unit's viewpoint, whereas  $X_i > c_i$ , then we have a ‘‘right-censored’’ observation and know only that  $X_i$  is contained in the interval  $\tilde{X}_i = (c_i, \infty]$ . In other settings one has purely ‘‘interval-censored’’ data: For the  $i$ -th observation one knows only which of given intervals  $(0, t_{i,1}]$ ,  $(t_{i,1}, t_{i,2}]$ ,  $\dots$ ,  $(t_{i,m(i)}, \infty]$  contains  $X_i$ , where  $0 < t_{i,1} < \dots < t_{i,m(i)} < \infty$ . If

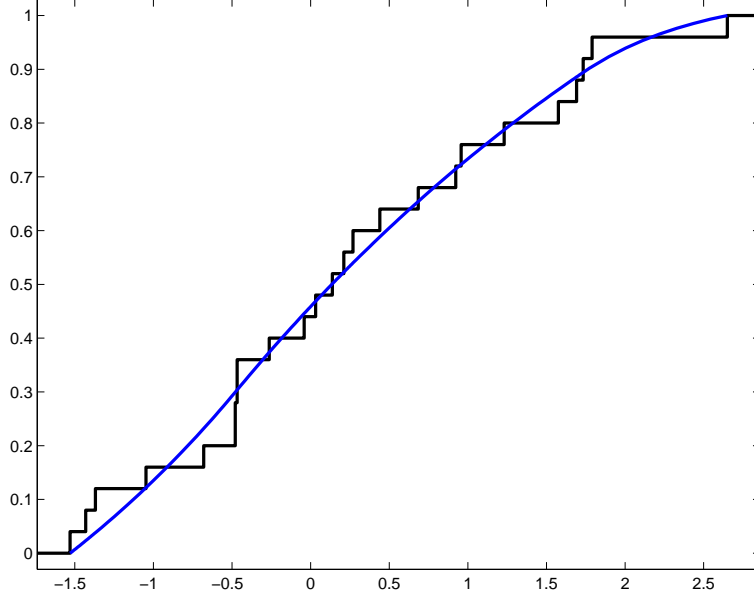


Figure 1: Estimated distribution functions for  $n = 25$  data points.

these candidate intervals are the same for all observations, one speaks of binned data. A related situation are rounded observations, e.g. when we observe  $\lceil X_i \rceil$  rather than  $X_i$ .

In all these settings we observe independent random intervals  $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n$ . More precisely, we assume that either  $\tilde{X}_i = (L_i, R_i] \ni X_i$  with  $0 \leq L_i < R_i \leq \infty$ , or  $\tilde{X}_i$  consists only of the one point  $L_i := R_i := X_i \in (0, \infty)$ . The normalized log-likelihood for this model reads

$$\begin{aligned} \bar{\ell}(\phi) := n^{-1} \sum_{i=1}^n & \left( 1\{L_i = R_i\} \phi(X_i) \right. & (16) \\ & \left. + 1\{L_i < R_i\} \log \left( \int_{L_i}^{R_i} \exp \phi(x) dx + 1\{R_i = \infty\} p_\infty \right) \right), \end{aligned}$$

where

$$p_\infty := 1 - \int_0^\infty \exp \phi(x) dx \in [0, 1].$$

## 5 An EM algorithm

Maximizing the log-likelihood function  $\bar{\ell}(\phi)$  for censored data is a non-trivial task and will be treated in detail elsewhere. Here we only indicate how this can be achieved in principle, assuming for simplicity that  $P(\{\infty\}) = 0$ , i.e.  $\int_0^\infty \exp \phi(x) dx = 1$  and  $p_\infty = 0$ . In this case, the log-likelihood simplifies to

$$\bar{\ell}(\phi) = n^{-1} \sum_{i=1}^n \left( 1\{L_i = R_i\} \phi(X_i) + 1\{L_i < R_i\} \log \left( \int_{L_i}^{R_i} \exp \phi(x) dx \right) \right).$$

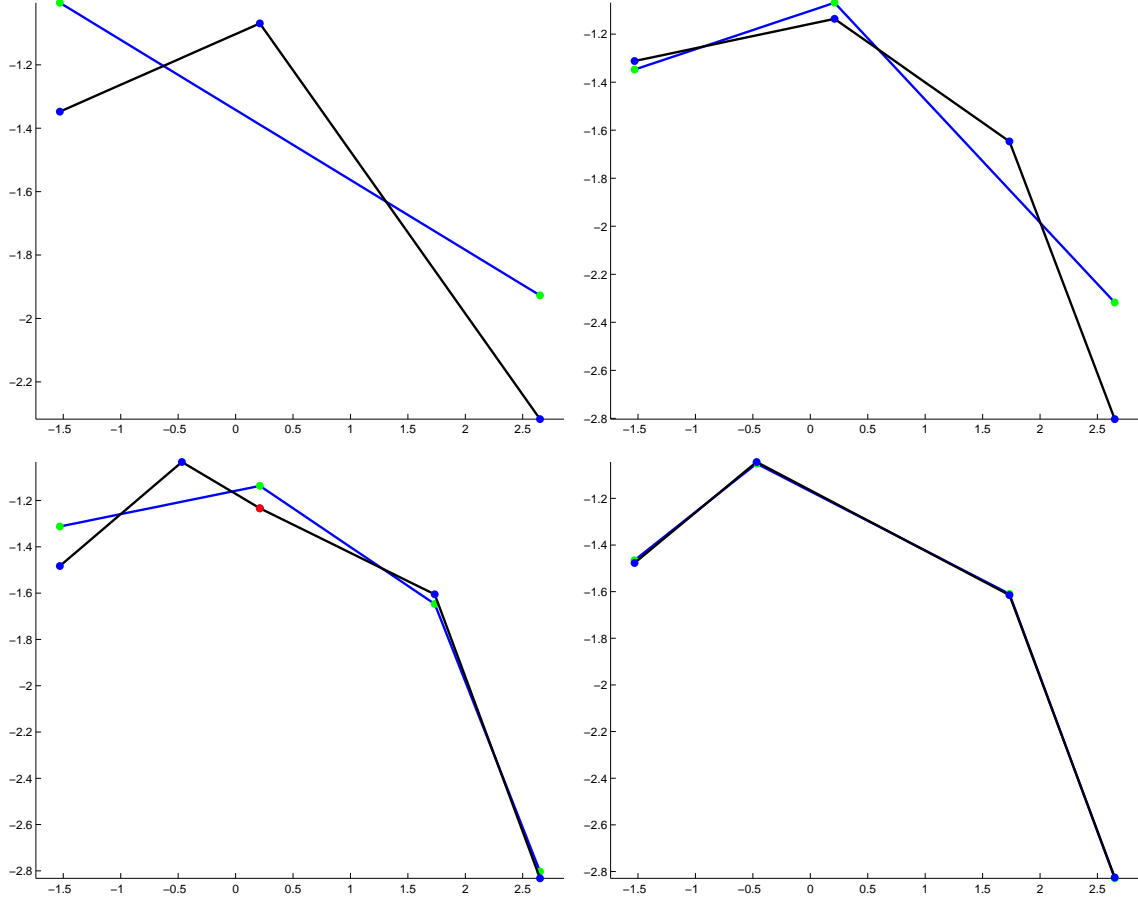


Figure 2: Estimating the log-density for  $n = 25$  data points.

Again one may get rid of the constraint  $\int_0^\infty \exp \phi(x) dx = 1$  by considering

$$\bar{L}(\phi) := \bar{\ell}(\phi) - \int_0^\infty \exp \phi(x) dx \quad (17)$$

for arbitrary concave and upper semicontinuous functions  $\phi : (0, \infty) \rightarrow [-\infty, \infty)$ .

A major problem is that  $\bar{\ell}(\phi)$  is not linear but convex in  $\phi$ . Namely, for  $v : (0, \infty) \rightarrow \mathbb{R}$  and  $0 \leq L < R \leq \infty$ ,

$$\frac{d^a}{dt^a} \Big|_{t=0} \log \left( \int_L^R \exp(\psi(x) + tv(x)) dx \right) = \begin{cases} \mathbb{E}_\phi(v(X) | X \in (L, R]) & \text{if } a = 1, \\ \text{Var}_\phi(v(X) | X \in (L, R]) & \text{if } a = 2. \end{cases} \quad (18)$$

Thus we propose to maximize  $\bar{\ell}(\phi)$  iteratively as follows: Starting from a function  $\phi$  with  $\bar{L}(\phi) > -\infty$ , we replace the target function  $\bar{L}(\phi_{\text{new}})$  with

$$\tilde{L}(\phi_{\text{new}} | \phi) := \frac{d}{dt} \Big|_{t=0} \bar{\ell}(\phi + t(\phi_{\text{new}} - \phi)) - \int_0^\infty \exp \phi_{\text{new}}(x) dx.$$

By means of (18), this may be written as

$$\tilde{L}(\phi_{\text{new}} | \phi) = \text{const}(\phi) + \int \phi_{\text{new}}(x) P(dx | \phi) - \int_0^\infty \exp \phi_{\text{new}}(x) dx, \quad (19)$$

where

$$P(\cdot | \phi) := n^{-1} \sum_{i=1}^n \left( 1\{L_i = R_i\} \delta_{X_i} + 1\{L_i < R_i\} \mathcal{L}_\phi(X | X \in (L_i, R_i]) \right),$$

a probability measure depending on the data and on  $\phi$ . In other words, for any Borel subset  $B$  of  $(0, \infty)$ ,

$$P(B | \phi) := n^{-1} \sum_{i=1}^n \left( 1\{L_i = R_i \in B\} + 1\{L_i < R_i\} \frac{\int_{B \cap (L_i, R_i)} \exp \phi(x) dx}{\int_{(L_i, R_i)} \exp \phi(x) dx} \right).$$

Note also that  $\tilde{L}(\phi_{\text{new}} | \phi)$  equals the conditional expectation of the complete-data log-likelihood  $L(\phi_{\text{new}})$ , given the available data and assuming the current  $\phi$  to be the true log-density:

$$\tilde{L}(\phi_{\text{new}} | \phi) = \mathbb{E}_\phi(L(\phi_{\text{new}}) | X_i \in \tilde{X}_i \text{ for } 1 \leq i \leq n),$$

where the  $\tilde{X}_i$  are treated temporarily as fixed.

After approximating the probability measure  $P(\cdot | \phi)$  by a discrete distribution with finite support, one can maximize  $\tilde{L}(\phi_{\text{new}} | \phi)$  over all concave functions  $\phi_{\text{new}}$  with the active-set algorithm presented in Section 3. Then we replace  $\phi$  with  $\phi_{\text{new}}$  and repeat this procedure until the change of  $\phi$  becomes negligible.

## 6 Auxiliary results and proofs

**Explicit formulae for  $J$  and some of its partial derivatives.** Recall the auxiliary function

$J(r, s) := \int_0^1 \exp((1-t)r + ts) dt$ . One may write explicitly

$$J(r, s) = J(s, r) = \begin{cases} (\exp(r) - \exp(s))/(r - s) & \text{if } r \neq s, \\ \exp(r) & \text{if } r = s, \end{cases}$$

or utilize the fact that  $J(r, s) = \exp(r)J(0, s - r)$  with  $J(0, 0) = 1$  and

$$J(0, y) = (\exp(y) - 1)/y = \sum_{k=0}^{\infty} \frac{y^k}{(k+1)!}.$$

To compute the partial derivatives  $J_{ab}(r, s)$  of  $J(r, s)$ , one may utilize the facts that  $J_{ab}(r, s) = J_{ba}(s, r) = \exp(r)J_{ab}(0, s - r)$ . Moreover, elementary calculations reveal that

$$\begin{aligned} J_{10}(0, y) &= (\exp(y) - 1 - y)/y^2 = \sum_{k=0}^{\infty} \frac{y^k}{(k+2)!}, \\ J_{20}(0, y) &= 2(\exp(y) - 1 - y - y^2/2)/y^3 = \sum_{k=0}^{\infty} \frac{2y^k}{(k+3)!}, \\ J_{11}(0, y) &= (y(\exp(y) + 1) - 2(\exp(y) - 1))/y^3 = \sum_{k=0}^{\infty} \frac{(k+1)y^k}{(k+3)!}. \end{aligned}$$

The Taylor series may be deduced as follows:

$$\begin{aligned}
J_{ab}(0, y) &= \int_0^1 (1-t)^a t^b e^{ty} dt \\
&= \sum_{k=0}^{\infty} \frac{y^k}{k!} \int_0^1 (1-t)^a t^{b+k} dt \\
&= \sum_{k=0}^{\infty} \frac{y^k}{k!} \frac{a!(b+k)!}{(k+a+b+1)!} \\
&= \sum_{k=0}^{\infty} \frac{a!(b+k)! y^k}{k!(k+a+b+1)!},
\end{aligned}$$

according to the general formula  $\int_0^1 (1-t)^k t^\ell dt = k!\ell!/(k+\ell+1)!$  for integers  $k, \ell \geq 0$ .

Numerical experiments revealed that a fourth degree Taylor approximation for  $J_{ab}(0, y)$  is advisable and works very well if

$$|y| \leq \begin{cases} 0.005 & (a = b = 0), \\ 0.01 & (a + b = 1), \\ 0.02 & (a + b = 2). \end{cases}$$

**Explicit formulae for the gradient and hessian matrix of  $L$ .** At  $\boldsymbol{\psi} \in \mathbb{R}^m$  these are given by

$$\begin{aligned}
\frac{\partial}{\partial \psi_k} L(\boldsymbol{\psi}) &= p_k - \begin{cases} \delta_1 J_{10}(\psi_1, \psi_2) & \text{if } k = 1, \\ \delta_{k-1} J_{01}(\psi_{k-1}, \psi_k) + \delta_k J_{10}(\psi_k, \psi_{k+1}) & \text{if } 2 \leq k < m, \\ \delta_{m-1} J_{01}(\psi_{m-1}, \psi_m) & \text{if } k = m, \end{cases} \\
-\frac{\partial^2}{\partial \psi_j \partial \psi_k} L(\boldsymbol{\psi}) &= \begin{cases} \delta_1 J_{20}(\psi_1, \psi_2) & \text{if } j = k = 1, \\ \delta_{k-1} J_{02}(\psi_{k-1}, \psi_k) + \delta_k J_{20}(\psi_k, \psi_{k+1}) & \text{if } 2 \leq j = k < m, \\ \delta_{m-1} J_{02}(\psi_{m-1}, \psi_m) & \text{if } j = k = m, \\ \delta_j J_{11}(\psi_j, \psi_k) & \text{if } 1 \leq j = k - 1 < m, \\ 0 & \text{if } |j - k| > 1. \end{cases}
\end{aligned}$$

**Proof of (2).** In what follows let  $\min(\boldsymbol{v})$  and  $\max(\boldsymbol{v})$  denote the minimum and maximum, respectively, of all components of a vector  $\boldsymbol{v}$ . Moreover let  $R(\boldsymbol{v}) := \max(\boldsymbol{v}) - \min(\boldsymbol{v})$ . Then with  $\boldsymbol{p} := (p_j)_{j=1}^m$  and  $\boldsymbol{\delta} = (\delta_k)_{k=1}^{m-1}$ , note first that

$$\begin{aligned}
L(\boldsymbol{\psi}) &\leq \max(\boldsymbol{\psi}) - (x_m - x_1) \exp(\min(\boldsymbol{\psi})) \\
&= R(\boldsymbol{\psi}) + \min(\boldsymbol{\psi}) - (x_m - x_1) \exp(\min(\boldsymbol{\psi})) \\
&\rightarrow -\infty \quad \text{as } \|\boldsymbol{\psi}\| \rightarrow \infty \text{ while } R(\boldsymbol{\psi}) \leq r_o
\end{aligned}$$

for any fixed  $r_o < \infty$ . Secondly, let  $\tilde{\psi}_j := \psi_j - \min(\boldsymbol{\psi})$ . Then  $\min(\tilde{\boldsymbol{\psi}}) = 0$ ,  $\max(\tilde{\boldsymbol{\psi}}) = R(\boldsymbol{\psi})$ , whence

$$\begin{aligned}
L(\boldsymbol{\psi}) &= \sum_{i=1}^m p_i \tilde{\psi}_i + \min(\boldsymbol{\psi}) - \exp(\min(\boldsymbol{\psi})) \int_{x_1}^{x_m} \exp(\tilde{\boldsymbol{\psi}}(x)) dx \\
&\leq (1 - \min(\boldsymbol{p})) R(\boldsymbol{\psi}) + \sup_{s \in \mathbb{R}} \left( s - \exp(s) \int_{x_1}^{x_m} \exp(\tilde{\boldsymbol{\psi}}(x)) dx \right) \\
&= (1 - \min(\boldsymbol{p})) R(\boldsymbol{\psi}) - \log \int_{x_1}^{x_m} \exp(\tilde{\boldsymbol{\psi}}(x)) dx - 1 \\
&= (1 - \min(\boldsymbol{p})) R(\boldsymbol{\psi}) - \log \left( \sum_{k=1}^{m-1} \delta_k J(\tilde{\boldsymbol{\psi}}_k, \tilde{\boldsymbol{\psi}}_{k+1}) \right) - 1 \\
&\leq (1 - \min(\boldsymbol{p})) R(\boldsymbol{\psi}) - \log \left( \min(\boldsymbol{\delta}) J(0, R(\boldsymbol{\psi})) \right) - 1 \\
&= (1 - \min(\boldsymbol{p})) R(\boldsymbol{\psi}) - \log J(0, R(\boldsymbol{\psi})) - \log(e \min(\boldsymbol{\delta})),
\end{aligned}$$

where we used the fact that  $\max_{s \in \mathbb{R}} (s - \exp(s)A) = -\log A - 1$  for any  $A > 0$ . Moreover, for  $r > 0$ ,

$$-\log J(0, r) = \log \left( \frac{r}{e^r - 1} \right) = -r + \log \left( \frac{r}{1 - e^{-r}} \right) \leq -r + \log(1 + r),$$

whence

$$L(\boldsymbol{\psi}) \leq -\min(\boldsymbol{p})R(\boldsymbol{\psi}) + \log(1 + R(\boldsymbol{\psi})) - \log(e \min(\boldsymbol{\delta})) \rightarrow -\infty \quad \text{as } R(\boldsymbol{\psi}) \rightarrow \infty. \quad \square$$

**Proof of Theorem 2.1.** It follows from strict concavity of  $L$  and (5) that the function  $\psi$  equals  $\check{\psi}$  if, and only if,

$$\sum_{i=1}^m p_i v(x_i) = \int_{x_1}^{x_m} v(x) f(x) dx \quad (20)$$

for any function  $v \in \mathcal{G}$ .

Note that any vector  $\boldsymbol{v} \in \mathbb{R}^m$  is a linear combination of the vectors  $\boldsymbol{v}^{(1)}, \boldsymbol{v}^{(2)}, \dots, \boldsymbol{v}^{(m)}$ , where

$$\boldsymbol{v}^{(k)} = (1\{i \leq k\})_{i=1}^m.$$

With the corresponding functions  $v^{(k)} \in \mathcal{G}$  we conclude that  $\psi$  maximizes  $L$  if, and only if,

$$\sum_{i=1}^k p_i = \int_{x_1}^{x_m} v^{(k)}(x) f(x) dx \quad (21)$$

for  $1 \leq k \leq m$ . Now the vector  $\boldsymbol{v}^{(m)}$  corresponds to the constant function  $v^{(m)} := 1$ , so that (21) with  $k = m$  is equivalent to  $F(x_m) = 1$ . In case of  $1 \leq k < m$ ,

$$v^{(k)}(x) := \begin{cases} 1 & \text{if } x \leq x_k, \\ (x_{k+1} - x)/\delta_k & \text{if } x_k \leq x \leq x_{k+1}, \\ 0 & \text{if } x \geq x_{k+1}, \end{cases}$$



and it follows from Fubini's theorem that

$$\begin{aligned}
\int_{x_1}^{x_m} v^{(k)}(x) f(x) dx &= \int_{x_1}^{x_m} \int_0^1 1\{u \leq v^{(k)}(x)\} du f(x) dx \\
&= \int_0^1 \int_{x_1}^{x_m} 1\{x \leq x_{k+1} - u\delta_k\} f(x) dx du \\
&= \int_0^1 F(x_{k+1} - u\delta_k) du \\
&= \delta_k^{-1} \int_{x_k}^{x_{k+1}} F(r) dr.
\end{aligned}$$

These considerations yield the characterization of the maximizer of  $L$ .

As for the first and second moments, equation (20) with  $v(x) := x$  yields the assertion that  $\sum_{i=1}^m p_i x_i$  equals  $\int_{x_1}^{x_m} x f(x) dx$ . Finally, let  $\mathbf{v} := (x_i^2)_{i=1}^m$  and  $v \in \mathcal{G}$  the corresponding piecewise linear function. Then

$$\begin{aligned}
\sum_{i=1}^m p_i x_i^2 - \int_{x_1}^{x_m} x^2 f(x) dx &= \int_{x_1}^{x_m} (v(x) - x^2) f(x) dx \\
&= \sum_{k=1}^{m-1} \int_{x_k}^{x_{k+1}} (x - x_k)(x_{k+1} - x) f(x) dx \\
&= \sum_{k=1}^{m-1} \delta_k^3 J_{11}(\psi_k, \psi_{k+1}). \quad \square
\end{aligned}$$

**Proof of Theorem 3.1.** It is well known from convex analysis that  $\psi \in \mathcal{K} \cap \text{dom}(L)$  belongs to  $\mathcal{K}_*$  if, and only if,  $\mathbf{v}^\top \nabla L(\psi) \leq 0$  for any vector  $\mathbf{v} \in \mathbb{R}^m$  such that  $\psi + t\mathbf{v} \in \mathcal{K}$  for some  $t > 0$ . By the special form of  $\mathcal{K}$ , the latter condition on  $\mathbf{v}$  is equivalent to  $\mathbf{v}_a^\top \mathbf{v} \geq 0$  for all  $a \in A(\psi)$ . In other words,  $\mathbf{v} = \sum_{i=1}^m \lambda_i \mathbf{b}_i$  with  $\lambda_a \geq 0$  for all  $a \in A(\psi)$ . Thus  $\psi \in \mathcal{K}$  belongs to  $\mathcal{K}_*$  if, and only if, it satisfies (8).

Similarly, a vector  $\psi \in \mathcal{V}(A) \cap \text{dom}(L)$  belongs to  $\mathcal{V}_*(A)$  if, and only if,  $\mathbf{v}^\top \nabla L(\psi) = 0$  for any vector  $\mathbf{v}$  in the linear space

$$\{\mathbf{v} \in \mathbb{R}^m : \mathbf{v}_a^\top \mathbf{v} = 0 \text{ for all } a \in A\} = \text{span}\{\mathbf{b}_i : i \in \{1, \dots, m\} \setminus A\}.$$

But this requirement is obviously equivalent to (9). □

**Acknowledgements.** This work was partially supported by the Swiss National Science Foundation. We are grateful to Charles Geyer for drawing our attention to active set methods and to Geurt Jongbloed for stimulating discussions about shape-constrained estimation.

**Software.** The methods of Rufibach (2006, 2007) as well as the active set method from Section 3 are available in the R package "logcondens" written by K. Rufibach and L. Dümbgen; see also Dümbgen and Rufibach (2011). Corresponding Matlab code is available from the first author's homepage on [www.stat.unibe.ch](http://www.stat.unibe.ch).

## References

- [1] W.J. BRAUN, T. DUCHESNE and J.E. STAFFORD (2005). Local likelihood estimation for interval censored data. *Canad. J. Statist.* **33**, 39-60.
- [2] L. DÜMBGEN, S. FREITAG-WOLF and G. JONGBLOED (2006). Estimating a unimodal distribution from interval-censored data. *J. Amer. Statist. Assoc.* **101**, 1094-1106.
- [3] L. DÜMBGEN and K. RUFIBACH (2009). Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli* **15**(1), 40-68.
- [4] L. DÜMBGEN and K. RUFIBACH (2011). logcondens: Computations related to univariate log-concave density estimation. *J. Statist. Software* **39**(6).
- [5] R. FLETCHER (1987). *Practical Methods of Optimization (2nd edition)*. Wiley, New York.
- [6] P. GROENEBOOM, G. JONGBLOED and J.A. WELLNER (2007). The support reduction algorithm for computing nonparametric function estimates in mixture models. *Scand. J. Statist.* **35**, 385-399.
- [7] J.P. KLEIN and M.L. MOESCHBERGER (1997). *Survival Analysis*. Springer Verlag.
- [8] K. LANGE, D.R. HUNTER and I. YANG (2000). Optimization transfer using surrogate objective functions (with discussion). *J. Comp. Graph. Statist.* **9**, 1-59.
- [9] J. PAL, M. WOODROOFE and M. MEYER (2006). Estimating a Polya frequency function. In: *Complex datasets and Inverse problems: Tomography, Networks and Beyond* (R. Liu, W. Strawderman, C.-H. Zhang, eds.), IMS Lecture Notes and Monograph Series **54**, pp. 239-249.
- [10] K. RUFIBACH (2006). *Log-Concave Density Estimation and Bump Hunting for I.I.D. Observations*. Dissertation, Universities of Bern and Göttingen.

- [11] K. RUFIBACH (2007). Computing maximum likelihood estimators of a log-concave density function. *J. Statist. Comp. Sim.* **77**, 561-574.
- [12] K. RUFIBACH and L. DÜMBGEN (2009). logcondens: Estimate a log-concave probability density from iid observations. *R package version 1.3.5*.
- [13] B.W. SILVERMAN (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795-810.
- [14] B.T. TURNBULL (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Royal Statist. Soc. B* **38**, 290-295.
- [15] J.A. WELLNER and Y. ZHAN (1997). A hybrid algorithm for computation of the non-parametric maximum likelihood estimator from censored data. *J. Amer. Statist. Assoc.* **92**, 945-959.