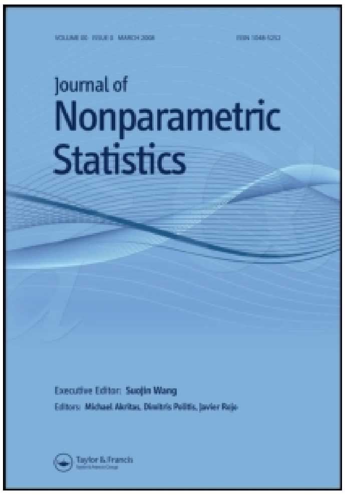

Journal of Nonparametric Statistics

Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

Computation of nonparametric convex hazard estimators via profile methods

Hanna K. Jankowski [a]; Jon A. Wellner [b]
[a] Department of Mathematics and Statistics, York University, Toronto, ON, Canada [b] Department of Statistics, University of Washington, Seattle, WA, USA

First Published:May2009

To cite this Article Jankowski, Hanna K. and Wellner, Jon A.(2009)'Computation of nonparametric convex hazard estimators via profile methods',Journal of Nonparametric Statistics,21:4,505 — 518

To link to this Article: DOI: 10.1080/10485250902745359

URL: http://dx.doi.org/10.1080/10485250902745359

Journal of Nonparametric Statistics
Vol. 21, No. 4, May 2009, 505–518

Computation of nonparametric convex hazard estimators via profile methods

Hanna K. Jankowski[a]* and Jon A. Wellner[b]

[a]*Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada;* [b]*Department of Statistics, University of Washington, Box 354322, Seattle WA 98195-4322, USA*

(*Received 10 November 2008; final version received 12 January 2009*)

This paper proposes a profile likelihood algorithm to compute the nonparametric maximum likelihood estimator of a convex hazard function. The maximisation is performed in two steps: First the support reduction algorithm is used to maximise the likelihood over all hazard functions with a given point of minimum (or antimode). Then it is shown that the profile (or partially maximised) likelihood is quasi-concave as a function of the antimode, so that a bisection algorithm can be applied to find the maximum of the profile likelihood, and hence also the global maximum. The new algorithm is illustrated using both artificial and real data, including lifetime data for Canadian males and females.

Keywords: shape constraint; least squares; maximum likelihood; support reduction algorithm; active set method; optimisation; survival analysis

AMS Subject Classification: 62G05; 62G99; 62P05; 62-09; 62N02

1. Introduction

Suppose we observe $X_1, \dots, X_n$ i.i.d. with density f. The X_is are assumed to represent lifetime data: failure of a material or machine, death, an earthquake, or infection by a disease. It is therefore natural to assume that f is concentrated on $[0, \infty)$. Of key interest to practitioners is the hazard (or failure) rate $h(t)$ given by the ratio $f(t)/(1 - F(t))$. Heuristically, $h(t)\mathrm{d}t$ is the probability that, given survival until time t, the event will occur in the next $\mathrm{d}t$ amount of time.

In reliability theory and demography it is quite natural to assume that the hazard rate is bathtub or U-shaped: that is, it is first decreasing and then increasing. (Throughout this paper we will say positive in lieu of 'non-negative' and strictly positive in lieu of 'positive'. A similar nomenclature will be used for negative functions). Heuristically, bathtub-shaped hazards correspond to lifetime distributions with high-initial hazard (or infant mortality), lower and often rather constant hazard during the middle of life, and then increasing hazard of failure (or wear out) as aging proceeds.

*Corresponding author. Email: hkj@mathstat.yorku.ca

ISSN 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/10485250902745359
http://www.informaworld.com

We will say that a bathtub-shaped function h has an *antimode* at a if it is nonincreasing on $[0, a]$ and nondecreasing on $[a, \infty]$. In particular, the antimode need not be a unique minimum.

Nonparametric estimators of hazard rates have received considerable interest in the literature, beginning with the work of Grenander [1] who considered the maximum likelihood estimator (MLE) of an increasing hazard rate. Bray *et al.* [2] extended this to the case of a general U-shaped rate function. It is well-known that these estimators result in a piecewise constant function and converge (under certain natural assumptions) at a rate of $n^{1/3}$ (cf. [3]). Here, we consider the nonparametric MLE with the additional assumption of convexity. This is often a natural assumption, and the resulting estimator is continuous (piecewise linear) and converges at the rate of $n^{2/5}$ (again under natural conditions).

Many other estimators of hazard functions (and solutions to the closely related problem of estimating the intensity of a Poisson process) with and without shape restrictions have been considered in the literature; see [4] for a nice review up to 2002. In recent years the focus has shifted to construction of 'adaptive' estimators over large scales of smoothness classes; see *e.g.* [5–7]. Virtually all of these other estimators require careful choice of penalty terms or tuning parameters, and computation of the adaptive estimators typically involves methods of combinatorial optimisation. Our estimators avoid the choices of tuning parameters or penalty terms by virtue of the shape constraint of convexity, and are therefore considerably more straightforward to compute.

To find the MLE, $\widehat{h}_n$, we need to maximise the likelihood

$$\mathcal{L}ik(h) = \prod_{i=1}^{n} h(X_i)\exp\{-H(X_i)\} = \prod_{i=1}^{n} h(X_{(i)})\exp\left\{-H(X_{(i)})\right\} \tag{1}$$

over the space of positive convex functions h. Here $\{X_{(i)}\}_{i=1}^{n}$ denotes the order statistics and $H(t) = \int_0^t h(s)\mathrm{d}s$. However, for a convex function fixed on $[0, X_{(n)})$, $\mathcal{L}ik(h)$ can be made arbitrarily large by increasing the value of $h(X_{(n)})$. We therefore find $\widehat{h}_n\colon [0, X_{(n)}) \mapsto \mathbb{R}_+$ by maximising the modified likelihood

$$\mathcal{L}^{\text{mod}}(h) = \prod_{i=1}^{n-1} h(X_{(i)})\exp\left\{-H(X_{(i)})\right\} \times \exp\left\{-H(X_{(n)})\right\}. \tag{2}$$

over $\mathcal{K}$, the space of nonnegative convex functions on $[0, X_{(n)})$. The full MLE is then found by additionally setting $\widehat{h}_n(x) = \infty$ for all $x \geq X_{(n)}$. Our goal is thus to find $\widehat{h}_n = \operatorname{argmin}_{h\in\mathcal{K}}\psi(h)$, for

$$\psi(h) = \int_0^\infty \left[\int_0^t h(s)\mathrm{d}s - \log h(t)\mathbb{I}_{t\neq X_{(n)}}\right]\mathrm{d}\mathbb{F}_n(t),$$

where $\mathbb{F}_n$ is the empirical cumulative distribution function (cdf) of the data.

Let $\mathcal{K}(a)$ denote the subspace of $\mathcal{K}$ of convex positive functions with an antimode at a, and consider the profile likelihood of a:

$$\mathcal{L}ik^{\text{mod}}(a) = \max_{h\in\mathcal{K}(a)} \mathcal{L}ik^{\text{mod}}(h).$$

A key result to our approach is that $-\log\, \mathcal{L}ik^{\text{mod}}(a)$ is itself bathtub-shaped in a. In the optimisation context, this property of ψ is also called *quasi-convex* [8]. This allows us to implement a two-step optimisation method: We apply the support reduction (SR) algorithm to maximise $\mathcal{L}ik(h)$ over $\mathcal{K}(a)$, and we maximise $\mathcal{L}ik^{\text{mod}}(a)$ with a bisection algorithm. We do not know of any other algorithm for this problem. Section 2 gives the details of the algorithm for the MLE, while Section 3 contains several examples.

If the true hazard function is convex, [9] show that the MLE is consistent, and that, under the additional assumption of strict convexity, it exhibits an $n^{2/5}$ *local* rate of convergence. This rate is also known to be optimal in a minimax sense. However, if the true hazard function has a second derivative equivalent to zero, we conjecture that the estimator will achieve a *global* rate of convergence of $n^{1/2}$. Using the algorithm we are able to provide evidence for this conjecture; see Figure 3 and the associated discussion in Section 3.

The algorithms described here are available through the R package convexHaz, [10]. Currently this contains only hazard estimation, but we hope to include right censoring and Poisson intensity estimation in future versions, as the techniques described here may be extended to those settings as well. Moreover, the package allows for the computation of a least-squares estimator of the hazard rate. The least-squares estimator is omitted here for the sake of brevity, but further information may be found in [9,11].

2. The algorithm

2.1. *Support reduction: minimising $\psi(h)$ over $\mathcal{K}(a)$*

The SR algorithm as developed in [12] is an extension of the vertex direction algorithm (cf. [13]). Within optimisation theory, the SR algorithm can be classified as an active set method. The algorithm is designed to handle nonparametric and semi-parametric M-estimation problems. Nonparametric solutions are infinite-dimensional; however, often it is known that the resulting estimator uses only a small number of dimensions. In these cases the SR algorithm works particularly well.

For a fixed antimode a, a positive convex function in $\mathcal{K}(a)$ may be decomposed as

$$h(t) = 1 \cdot \alpha + \int_0^a (\tau - t)_+ \mathrm{d}\nu(\tau) + \int_a^\infty (t - \eta)_+ \mathrm{d}\mu(\eta), \tag{3}$$

where ν and μ are positive measures, and $\alpha \geq 0$ is a constant (the positivity is what ensures that $h \in \mathcal{K}(a)$). In this representation, we call α, ν, and μ the mixing measure of h, and the support of these measures becomes the support of h. The total measure of a function h is then $\alpha + \nu[0, a] + \mu[a, X_{(n)}]$. Lemma A.1 shows that the support of $\widehat{h}_n$ is always finite for a fixed sample size. In fact, in practice the number of support points is considerably smaller than n. Thus, our decomposition will look like

$$h = \alpha \cdot e_0 + \sum_{i=1}^{k} \nu_i \cdot e_{1,\tau_i} + \sum_{j=1}^{m} \mu_j \cdot e_{2,\eta_j}, \tag{4}$$

where $e_0(t) = 1$, $e_{1,\tau}(t) = (\tau - t)_+$, and $e_{2,\eta}(t) = (t - \eta)_+$ denote the basis functions. We may equivalently characterise h in terms of the support, supp $= \{1\} \times \{\tau_1, \dots, \tau_k\} \times \{\eta_1, \dots, \eta_m\}$ and mixing measure, mix $= \{\alpha\} \times \{\nu_1, \dots, \nu_k\} \times \{\mu_1, \dots, \mu_m\}$.

We also note that the support points (or 'bend' points of the hazard function) will never fall on an observation point. Indeed more is known. There will be at most one support point between two successive observations points $X_{(i)}$ and $X_{(i+1)}$ for $i \leq n - 2$, except possibly when the two support points are τ_k and η_1 (that is, the last 'down' bend and the first 'up' bend). In this case there may be two support points between successive observation points. For a proof of this fact see [9].

Recently, Dümbgen *et al.* [14] have implemented an active set approach to develop an algorithm for the MLE of a log-concave density. In this case however, by using additional knowledge about the likelihood, one may reduce the problem to a finite-dimensional one. A similar simplification occurs in the estimation of a decreasing hazard, [1], although here a solution may be found exactly and no approximating algorithm is required. The reason that the simplification occurs is that the resulting estimator has jump points only at the observation points: the log-concave density estimator is piecewise linear with bend points at the observation points, and the decreasing hazard estimator has points of discontinuity only at the data points. In our situation, we know that our estimator is piecewise linear, but we do not know the location of the bend points.

The main idea behind the SR algorithm is as follows (for pseudocode see [12]). We wish to minimise the criterion function $\psi(h)$ over the space of $h \in \mathcal{K}(a)$. Given a current iterate $\widehat{h}$ with finite support $\widehat{supp}$ and mixing measure $\widehat{mix}$, we first find a new support point by finding the basis function e^* such that the directional derivative

$$\nabla\psi(\widehat{h})[e] \equiv \lim_{\varepsilon\to 0} \frac{\big(\psi(\widehat{h}+\varepsilon e) - \psi(\widehat{h})\big)}{\varepsilon}$$

is smallest. The support corresponding to e^* is added to $\widehat{supp}$ to yield $\widehat{supp}^*$, and then ψ is minimised over all h with support given by $\widehat{supp}^*$ to give the new mixing measure, $\widehat{mix}^*$. Note that the minimisation here is finite dimensional and is therefore easy to implement. This is the vertex direction part of the algorithm: the idea is to continually move in a direction that decreases the criterion function the most. The SR algorithm adds an additional step (called the SR step), which ensures that we remain in $\mathcal{K}(a)$, i.e. that the mixing measure is positive.

The idea behind the SR step is as follows. We have a previous estimate $\widehat{h} \in \mathcal{K}(a)$ (with $\widehat{supp}, \widehat{mix}$), and a current candidate $\widehat{h}^*$ (with $\widehat{supp}^*, \widehat{mix}^*$) that lies outside of $\mathcal{K}(a)$. We find a convex combination of $\widehat{h}$ and $\widehat{h}^*$ on the boundary of $\mathcal{K}(a)$ by finding the largest λ, λ^*, such that $(1-\lambda)\widehat{h} + \lambda\widehat{h}^*$ is in $\mathcal{K}(a)$, i.e. such that $(1-\lambda)\widehat{mix} + \lambda\widehat{mix}^* \geq 0$. This is equivalent to removing one of the support points of $\widehat{h}^*$. One can easily show that a newly proposed support point will never be removed here. Set $\widehat{supp}$ to the reduced support and calculate $\widehat{mix} = (1-\lambda^*)\widehat{mix} + \lambda^*\widehat{mix}^*$. ψ is again minimised over all h with support given by $\widehat{supp}^* = \widehat{supp}$ to give the new mixing measure, $\widehat{mix}^*$. If $\widehat{mix}^*$ is not positive, then the procedure is iterated.

Since $\psi(h)$ is convex as a function of h, $\widehat{h}$ is its minimiser over $\mathcal{K}(a)$ if and only if the directional derivative at $\widehat{h}$ is positive in any direction. The possible directions may be described via the basis functions $e_0, e_{1,\tau}$ for $\tau \in [0,a]$ and $e_{2,\eta}$ for $\eta \in [a,T]$. Hence the SR algorithm is iterated until

$$\nabla\psi(\widehat{h}) = \min\left\{\nabla\psi(\widehat{h})[e_0],\ \min_{\tau\in[0,a]} \nabla\psi(\widehat{h})[e_{1,\tau}],\ \min_{\eta\in[a,T]} \nabla\psi(\widehat{h})[e_{2,\eta}]\right\}$$

is larger than $-\varepsilon$, for some predetermined accuracy $\varepsilon > 0$.

To perform the SR algorithm as described above, one needs to minimise $\psi(h)$. However, even for a function h with fixed support, the logarithm in $\psi(h)$ causes the minimisation problem to be a general convex optimisation problem with no exact solution. Our method for dealing with this is to replace the general convex optimisation problem with a sequence of (simpler) quadratic optimisation problems, for which exact solutions are easily available. To achieve this, we minimise an approximate version of ψ instead of the true ψ. Suppose that the current iterate $\widehat{h}$ in the algorithm is close to the true minimiser of ψ, then instead of minimising ψ, we could equally well minimise the quadratic approximation to ψ. This *inner* minimisation is iterated via a line search strategy, until the directional derivatives of the true ψ are sufficiently large.

We define the approximate criterion function with respect to a fixed function, say $g \in \mathcal{K}(a)$, using the approximation $\log(1+x) \approx x - x^2/2$. Then $\psi(h)$ is equal to

$$\begin{aligned}
&\psi(g) + \int_0^\infty \left[H(t) - \log h(t)\mathbb{I}_{t\neq X_{(n)}}\right] \mathrm{d}\mathbb{F}_n(t) - \int_0^\infty \left[G(t) - \log g(t)\mathbb{I}_{t\neq X_{(n)}}\right] \mathrm{d}\mathbb{F}_n(t) \\
&\quad \approx \psi(g) + \int_0^\infty [H-G](t)\mathrm{d}\mathbb{F}_n(t) - \int_{[0,X_{(n-1)}]} \left(\frac{(h-g)(t)}{g(t)}\right) \mathrm{d}\mathbb{F}_n(t) \\
&\qquad + \frac{1}{2}\int_{[0,X_{(n-1)}]} \left(\frac{(h-g)(t)}{g(t)}\right)^2 \mathrm{d}\mathbb{F}_n(t).
\end{aligned}$$

As we minimise the approximation to ψ over h for a fixed g, we may remove all terms depending only on g to obtain

$$\psi^{\mathrm{mod}}(h|g) = \int_0^\infty H(t)d\mathbb{F}_n(t) - 2\int_{[0,X_{(n-1)}]} \frac{h(t)}{g(t)}\mathrm{d}\mathbb{F}_n(t) + \frac{1}{2}\int_{[0,X_{(n-1)}]} \left(\frac{h(t)}{g(t)}\right)^2 \mathrm{d}\mathbb{F}_n(t).$$

Algorithm to find the profile MLE:

```
Step 0. Obtain an initial estimate, $\widehat{h}$.
While $\nabla\psi(\widehat{h})$ is less than $-\varepsilon$ Repeat 1–3:
    Step 1 (inner loop). Given a current estimate $\widehat{h}$, find the next proposed $\widehat{h}_p$ by minimising the linearised criterion function $\psi^{\mathrm{mod}}(h|\widehat{h})$ using the SR algorithm.
        To find the starting value for the SR algorithm, do the following:
        A: Consider the support of $\widehat{h}$, and find the function $\widehat{h}_0$ with the same support which minimises the function $\psi^{\mathrm{mod}}(h|\widehat{h})$.
        While $\widehat{h}_0 \notin \mathcal{K}(a)$ Repeat B:
        B: Perform an SR step to obtain a new $\widehat{h}_0$, with a reduced support.
    Step 2. Find $\lambda^*$ in [0, 1] which minimises $\psi\left((1-\lambda)\widehat{h} + \lambda\widehat{h}_p\right)$.
    Step 3. The new $\widehat{h}$ is set to $(1-\lambda^*)\widehat{h} + \lambda^*\widehat{h}_p$.
```

Several computational issues arise in the implementation.

2.1.1. *Gridded implementation*

In practice, it is not possible to find the exact location of the minimum of the directional derivatives, as the gradient of the criterion function is far from smooth. Therefore, a natural approach is to minimise the gradient over a pre-specified, and sufficiently dense grid. This is not ideal, as there is no way to guarantee the behavior of the gradient outside of the grid.

In our implementation, we split $[0, X_{(n)}]$ into M intervals (resulting in $M+1$ grid points), and only check for the minimum at these locations. Naturally, the larger M is, the more accurate our answer. However, increasing M also increases computing time.

2.1.2. *A gridless alternative*

Increasing the gridded implementation from M =100 to $M = 1000$, say, typically does not have a drastic effect on the location of the support points. We therefore propose the following alternative to allow the algorithm to naturally fine-tune their location.

Suppose that the grid used in the algorithm is such that $G = \{\theta_1, \ldots, \theta_b\}$, and suppose also that the SR algorithm proposed the new support point $\theta_i \in G$. We then augment the grid to

$$G \cup \left\{ \frac{\theta_{i-1} + \theta_i}{2}, \frac{\theta_i + \theta_{i+1}}{2} \right\}.$$

This has no effect on the next step of the SR algorithm, but will impose a finer grid when the exit criterion $\nabla\varphi(\widehat{h}) \leq -\varepsilon$ is next checked.

Naturally, there are many other ways in which one could augment the grid at this time. We found that the proposed method was the most efficient, giving the best results without sacrificing the speed of the algorithm. The efficacy of the gridded vs. gridless modifications is studied in Figure 2.

2.1.3. *Nearly singular matrices*

In the inner loop we need to minimise a quadratic function in finitely many variables. Unfortunately, the system of equations is sometimes computationally singular. This most often happens just after a new support point has been added. If this occurs, we handle the problem by deleting a point of the support closest to the newly proposed support point. We find that this ad hoc solution works reasonably well in practice.

2.2. *The bisection algorithm*

The algorithm of the previous section finds the MLE of $\psi(h)$ over $\mathcal{K}(a)$ for fixed antimode a, and we now need to optimise over the possible values of a to find the overall MLE. The next result allows us to speed up the search for the optimal antimode by use of a bisection algorithm.

Proposition 2.1 *Let $a_0 \in [0, T]$ be such that the minimiser of $\psi(h)$ over h in $\mathcal{K}_+$ has an antimode at a_0. Suppose that $a_0 < a_1 < a_2$, then*

$$\min_{h\in\mathcal{K}_+} \psi(h) \equiv \min_{h\in\mathcal{K}_+(a_0)} \psi(h) \leq \min_{h\in\mathcal{K}_+(a_1)} \psi(h) \leq \min_{h\in\mathcal{K}_+(a_2)} \psi(h).$$

The inequalities also hold if $a_0 > a_1 > a_2$. That is, $\widetilde{\psi}(a) = \min_{h\in\mathcal{K}_+(a)} \psi(h)$ is bathtub-shaped in a.

Fix an accuracy parameter $\varepsilon > 0$. For a vector $\boldsymbol{x} = \{x_i\}_{i=1}^k$, let $\{x_{(1)}, \ldots, x_{(k)}\}$ denote the ordered elements of $\boldsymbol{x}$ (in increasing order) and let $\Delta\boldsymbol{x} = \sum_{i=1}^k (x_{(i)} - x_{(1)})^2$.

Bisection algorithm:

Step 0. Let $\boldsymbol{a} = \{a_i\}_{i=1}^5 = X_{(n)} * \{0, 0.25, 0.5, 0.75, 1\}$, and find $\widetilde{\boldsymbol{\psi}} = \{\widetilde{\psi}(a_i)\}_{i=1}^5$.
While $\Delta\widetilde{\boldsymbol{\psi}}$ is greater than ε Repeat 1-2:

Step 1. Write $\tilde{\boldsymbol{\psi}}$ as $\{\tilde{\psi}_i\}_{i=1}^5$. If $\tilde{\psi}_{(1)} = \tilde{\psi}_i$ for $i = 2, 3, 4$, set $\tilde{a}_1 = a_{i-1}, \tilde{a}_3 = a_i$ and $\tilde{a}_5 = a_{i+1}$. If $\tilde{\psi}_{(1)} = \tilde{\psi}_1$, set $\tilde{a}_1 = a_1, \tilde{a}_3 = a_2$ and $\tilde{a}_5 = a_3$. If $\tilde{\psi}_{(1)} = \tilde{\psi}_5$, set $\tilde{a}_1 = a_3, \tilde{a}_3 = a_4$ and $\tilde{a}_5 = a_5$. Fill in the remaining elements of $\tilde{\boldsymbol{a}}$: set $\tilde{a}_2 = (\tilde{a}_1 + \tilde{a}_3)/2$ and $\tilde{a}_4 = (\tilde{a}_3 + \tilde{a}_5)/2$.

Step 2. Let $\boldsymbol{a} = \tilde{\boldsymbol{a}}$. Find $\widetilde{\boldsymbol{\psi}} = \{\widetilde{\psi}(a_i)\}_{i=1}^5$. Note that three of the five entries have already been calculated.

Step 3. argmin $\widetilde{\psi}(a)$ is given by the a_i, which minimises the current $\tilde{\boldsymbol{\psi}}$.

2.2.1. *Estimating the antimode*

In [9], we show that if the true hazard function has a unique antimode at $\tilde{a}$, then the antimode of the MLE will converge to $\tilde{a}$ as the sample size tends to infinity. In fact, we conjecture that the rate of this convergence is $n^{1/5}$. However, our algorithms are not yet optimised to find the estimator of the antimode.

2.2.2. *Gridded implementation*

Proposition 2.1 shows that the theoretical value of $\widetilde{\psi}(a)$ is bathtub-shaped in a, and this is a key observation in the application of the bisection algorithm. In practice however, we use the gridded implementation to approximate the true $\widetilde{\psi}(a)$. Fortunately, we have found that this approximation does not invalidate the bathtub shape, and the same is true of the gridless implementation (see Figure 2).

2.2.3. *Convergence*

A key question in any algorithm is that of convergence. Using Lemma A.1 with Theorem 1 in [12] and Theorem 7.2.3 in [15], one can check conditions for the constrained minimisation step. However, these results would apply only to a theoretical implementation (and not our gridded, or even gridless modifications). Second, we address the bisection step. As $\tilde{\psi}(a)$ is not convex, this is not guaranteed to converge to the overall minimum. One alternative is to do a gridded search over $a \in [0, X_{(n)}]$, but this is much less efficient, and equally not guaranteed to find the minimum. However, we have run the algorithm hundreds of times on both simulated and real data sets, and have only observed nonconvergence in a small number of cases. For these cases, the reason that the algorithm failed lies in the problem of nearly singular matrices described above. We emphasise that the number of these cases was small, and convergence could always be achieved if the size of the grid M was not chosen to be too large.

3. Examples and simulations

3.1. *A simulated example*

To illustrate our proposed estimators, consider the distribution with density given by

$$f(t) = \frac{1+2b}{2A\sqrt{b^2 + (1+2b)t/A}}, \quad \text{on } 0 \le t \le A.$$

This distribution was proposed in [16] as a relatively simple model with bathtub-shaped hazards, which also has an adequate ability to model lifetime behavior. For simplicity, we will call this the HS distribution after the authors. The distribution has *convex* hazards for all values of b in the parameter space, $b > -1/2$. In Figure 1, we present an example of the MLE for a simulation from this distribution with a sample size of 100. Notice that the estimator blows up at zero, and also by definition at $X_{(n)}$. This behavior is typical of shape-constrained nonparametric estimators,

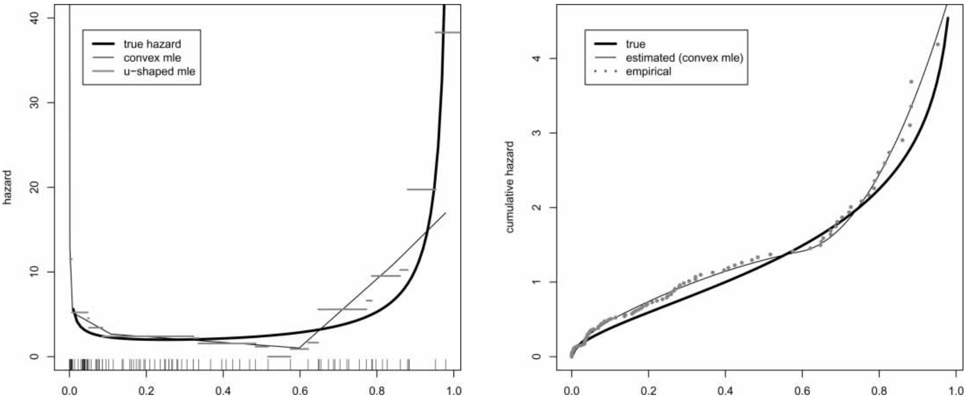

Figure 1. Estimation of the HS hazard with $b = 0$, $A = 1$ for a sample size of 100: the plot on the left shows the true hazard function (bold), the estimated convex hazard (black) and the estimated bathtub-shaped hazard (grey step-function), the actual observations are shown along the x-axis; the plot on the right shows the true cumulative hazard (bold), the empirical hazard function $\mathbb{H}_n$ (grey), and the estimated cumulative hazard found by integrating the convex MLE (black).

see for example Remark 4.5 in [9]. We also compare our convex estimators to the bathtub-shaped MLE without the convex restriction [2]. This estimator is also infinite beyond $X_{(n)}$, and is known to be inconsistent at zero. Both the U-shaped and convex MLE appear to be following a similar trend. Figure 3 also compares the cumulative hazards: of the true distribution, of the estimated convex MLE, and of the data, $\mathbb{H}_n$. Note that the estimated function follows the empirical one quite closely.

In Figure 2 we examine the gridless *vs.* gridded implementations for the HS distribution with $b = 0$, $A = 1$. First, we examine the 'bisection diagnostics' for different implementations for the same data as in Figure 3. The first plot of Figure 3 shows the values of the negative of the profile log-likelihood, $-\log\{\max_{h\in\mathcal{K}(a)} \mathcal{L}ik^{\text{mod}}(h)\}$, as a function of the antimode a for the values of the antimode checked by the bisection algorithm. Note that the bathtub shape of the negative of the logarithm of the profile likelihood is preserved by the different implementations. Also, the gridless implementation with gridsize $M = 100$ had a similar running time to that of the gridded implementation with $M = 2000$, but achieved greater accuracy. In the remaining plots we examine more closely the accuracy vs. running time of both implementations. The R package convexHaz was used to assess the efficacy of the methods for 25 samples of size 100 from the HS distribution with $b = 0$, $A = 1$. In each of the 25 samples, a convex MLE with antimode

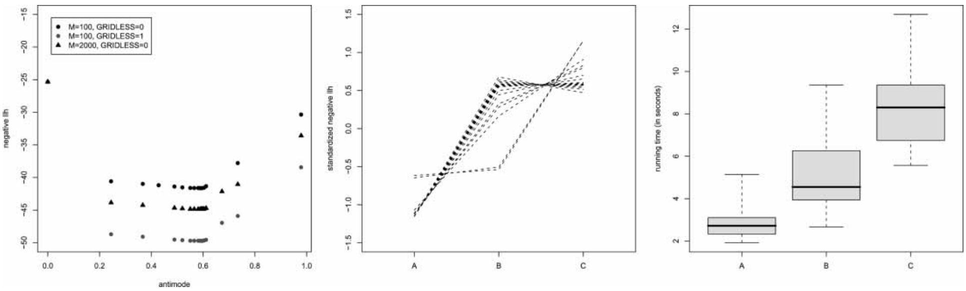

Figure 2. Gridded *vs.* gridless implementations: bisection diagnostics for the convex MLE of Figure 1 (left), and a comparison of the accuracy of 25 samples of size 100 from the HS distribution with $b = 0$, $A = 1$ (middle – standardised results for the negative log-likelihood, right – simulation times). The three configurations had (A) M =100, GRIDLESS = FALSE, (B) $M = 100$, GRIDLESS = TRUE, and (C) $M = 1000$, GRIDLESS = FALSE; M denotes the grid size, and if GRIDLESS = TRUE, the gridless implementation was used.

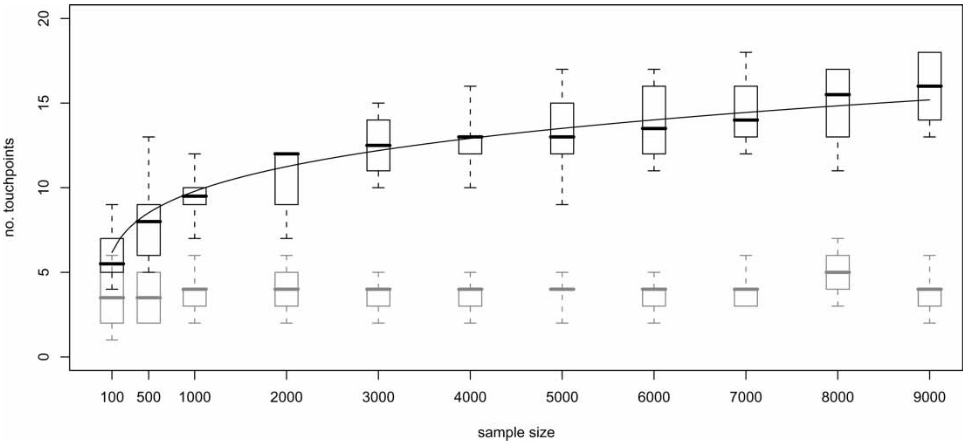

Figure 3. Support size as a function of sample size for the convex MLE: the figure shows the number of support points obtained for each sample size n when observations were drawn from either a Weibull distribution with cubic hazard (black) or an exponential distribution (grey). Each boxplot results from 10 different samples. For the Weibull distribution, we also show the line *no. touchpoints* $= Cn^{1/5}$, where the constant C was determined from the data.

at zero was fit to the data, using the function `srMLE` from the R package convexHaz. For a data set x, the functions implemented were (A) $M = 100$, GRIDLESS = FALSE, (B) $M = 100$, GRIDLESS = TRUE, and (C) $M = 1000$, GRIDLESS = FALSE. Standardised results for the negative log-likelihood for each data set across the three settings appear in the middle, and the boxplots of the running times in the right plot. We can see that 23 out of the 25 times the gridless $M = 100$ implementation has comparable results at a shorter running time than that of the gridded implementation with $M = 1000$. Using the gridless implementation does not always guarantee a faster and more accurate result for the MLE algorithm: The method does increase the accuracy of the step that uses the SR algorithm to minimise ψ^{mod}, but has varying results with the overall algorithm.

3.2. *Support size vs. sample size*

The size of the support of the MLE $\widehat{h}_n$ is in general considerably smaller than n. (For h with decomposition as in Equation (4), the size of the support is $k + m + 1_{\alpha\neq 0} << n$.) In fact, the behavior varies depending on the shape of the true hazard function.

Theorem 2.7 of [9] shows that the MLE converges locally at a rate of $n^{2/5}$. The result also gives information on the asymptotic number of support points of the estimators. That is, for a fixed location x_0, in a neighborhood of size $n^{-1/5}$ the number of support points is constant. This also implies that the total number of support points should grow as $Cn^{1/5}$. However, this holds only if the second derivative is strictly positive. We conjecture that for hazards with $h_0''(x) \equiv 0$, the rate of convergence will be $n^{1/2}$, and that this rate of convergence will be global. If this conjecture holds, then the growth of support points in sample size should be different for, say, the exponential distribution than for the Weibull distribution with cubic hazard. This is exactly what we see in Figure 3, where we plot the number of support points *vs.* sample size in the MLE for simulations from the Weibull distribution *vs.* the exponential distribution. Although the algorithm finds an approximation to the MLE, and hence the number of support points is also approximate, the simulation shows a clear difference in the asymptotic behavior between the two hazard rates.

3.3. *Two examples*

Next we consider the number of operating hours between successive failures of air conditioning equipment in 13 aircrafts. A total of 213 times were recorded. This data set was studied in [17] and again in [18]. We are interested in the overall hazard rate of the intervals between successive failures.

The analysis of [17] is summarised as follows. First an exponential fit to the data was considered. Although the null hypothesis of exponential times was not rejected by the Kolmogorov–Smirnov test, the data appeared to exhibit a decreasing hazard rate. Specifically, the empirical survival function lies first below, then above the fitted exponential one, indicating a lack of fit. Also, the intervals do not show a trend toward either longer or shorter intervals with increased use of the unit. On closer inspection, it appears that the exponential is a good-fit to the data, but that each airplane is following a different failure rate. This would correspond to the pooled intervals exhibiting a decreasing failure rate [17, Theorem 2]. The null hypothesis of a constant hazard rate (corresponding to the same exponential distribution for all 13 airplanes), was then tested against the alternative hypothesis of decreasing failure rate (corresponding to different exponential distributions for the different airplanes) via a test statistic because of [19]. The resulting test was significant, with a p-value of 0.007, and hence lead to the conclusion that the pooled distribution has a decreasing hazard rate.

Cox and Lewis [18] consider fitting time-dependent Poisson processes to the data, and ultimately settle on a mixture of homogeneous processes, in agreement with Proschan [17].

Figure 4 shows our fit of the nonparametric convex MLE. The MLE has an antimode at the 375 h mark, which appears to be in contradiction to the results of [19]. We investigated this further using resampling methods, and found that there is no sufficient evidence against the hypothesis of decreasing failure rate. Therefore, our ultimate estimator is the nonparametric convex and decreasing MLE to the data, also shown in Figure 4. We note that this estimator maximises the full likelihood (Equation (1)), and not the modified likelihood (Equation (2)). The tail of the survival functions explains why it is difficult to tell the difference between a decreasing curve, and a convex curve; this may be partially explained by how close the cdfs for both fits are (see the right plot in Figure 4), along with the fact that the difference appears in a region with very few observations.

Finally, we apply our estimators to a lifetime data set: the Canadian mortality table for the years 2000–2002 [20]. To generate our results, we took a random sample of size $n = 1000$ from the distribution given by the life tables. We also use a simplified version of the standard actuarial

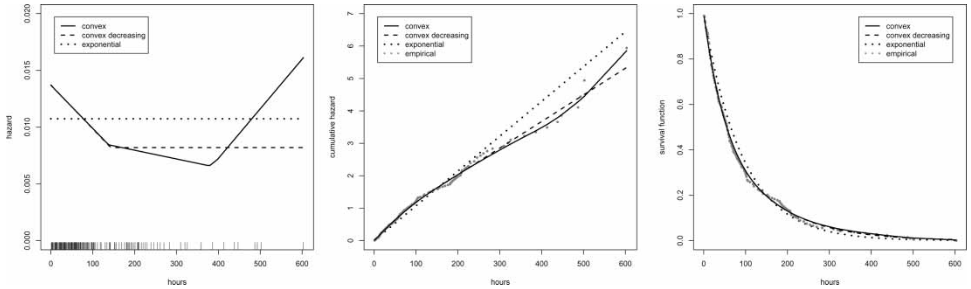

Figure 4. Maximum likelihood estimators for air conditioning data of [14]: all three plots show the convex fit (solid line), convex decreasing fit (dashed line) and exponential fit (dotted line). The rightmost plot shows the hazard functions, with the data plotted on the x-axis. The two left plots show the cumulative hazard and the survival function respectively; Here, the empirical functions are also added in grey. Note that differences in the nonparametric estimators appear at roughly the 300 h mark: only 12 of the 213 observations are larger than 300, and 7 of 213 are larger than 400.

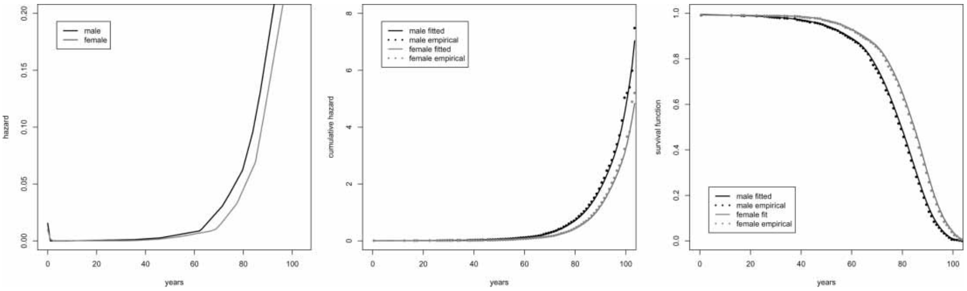

Figure 5. Maximum likelihood estimator for Canadian lifetime data: the leftmost plot shows the fitted hazard rates for males (black) and females (grey); fitted and empirical cumulative hazard rates and survival functions are shown in the two plots to the right.

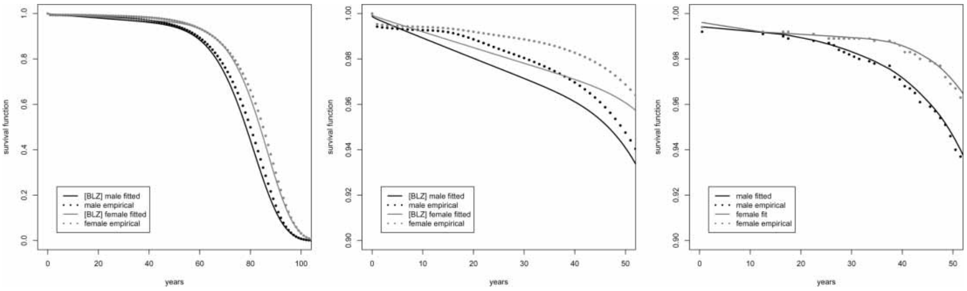

Figure 6. A comparison of the parametric model of [21] (denoted as [BLZ]) to the nonparametric convex MLE: we show the [BLZ] fitted and empirical survival functions (left, and middle) as well as the convex MLE fitted and empirical survival functions (right - this is a close-up of Figure 5). The [BLZ] model was fit directly to the life-table survival function, whereas the convex MLE was fit to a sample of size 1000 from this distribution.

assumption of uniform deaths for fractional ages. That is, we assume that all deaths occurred half-way through the year. The resulting MLE for both male and female lifetimes are given in Figure 5; fitted cumulative hazards and survival functions are also shown. A parametric approach for this data was considered in [21] (Figure 2(a)). Specifically, Bebbington *et al.* [21] fit a mixture of flexible and reduced additive Weibull survival functions. A comparison of the survival functions is provided in Figure 6.

Acknowledgements

The first author was partially supported by NSERC and York University. The second author was supported in part by NSF grants DMS-0503822, DMS-0804587, and NI-AID grant 2R01 AI291968-04.

References

[1] U. Grenander, *On the theory of mortality measurement II*, Skand. Aktuarietidskr. 39 (1956), pp. 125–153 (1957).

[2] T. Bray, G. Crawford, and F. Proschan, *Maximum likelihood estimation of a U-shaped failure rate function*, Mathematical Note 534, Mathematics Research Laboratory, Boeing Scientific Research Laboratories, Seattle, WA 1967. Available at http://www.stat.washington.edu/jaw/RESEARCH/OLD-PAPERS-OTHERS/UMLE.pdf

[3] B.L.S. Prakasa Rao, *Estimation of a unimodal density*, Sankhyā Ser. A 31 (1969), pp. 23–36.

[4] L. Reboul, *Estimation of a function under shape restrictions: Applications to reliability*, Ann. Statist. 33 (2005), pp. 1330–1356.

[5] Y. Baraud and L. Birgé, *Estimating the intensity of a random measure by histogram type estimators*, Probab. Theory Related Fields (To appear), published online with SpringerLink.

[6] E. Brunel and F. Comte, *Adaptive nonparametric regression estimation in presence of right censoring*, Math. Methods Statist. 15 (2006), pp. 233–255.

[7] P. Reynaud-Bouret, *Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities*, Probab. Theory Related Fields 126 (2003), pp. 103–153.

[8] J.M. Borwein and A.S. Lewis *Convex Analysis and Nonlinear Optimisation*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 3 Springer-Verlag, New York, 2000.

[9] H. Jankowski and J.A. Wellner, *Nonparametric estimation of a convex bathtub-shaped hazard function*, Tech. Rep. 521, Department of Statistics, University of Washington, 2007.

[10] H. Jankowski, X. Wang, H. McCauge, and J. Wellner, *convexHaz: R functions for convex hazard rate estimation*, 2008; software available at http://www.r-project.org.

[11] H. Jankowski and J.A. Wellner, *Computation of nonparametric convex hazard estimators via likelihood profile methods*, Tech. Rep. 542, Department of Statistics, University of Washington, 2008.

[12] P. Groeneboom, G. Jongbloed, and J.A. Wellner, *The support reduction algorithm for computing nonparametric function estimates in mixture models*, Scand. J. Statist. 35 (2008), pp. 385–399.

[13] D. Böhning, *Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process*, Ann. Statist. 10 (1982), pp. 1006–1008.

[14] L. Dümbgen, A. Hüsler, and K. Rufibach, *Active set and EM algorithms for log-concave densities based on complete and censored data*, University of Bern, 2007.

[15] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty *Nonlinear programming; Theory and Algorithms*, 3rd ed., Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006.

[16] E. Haupt and H. Schäbe, *The TTT transformation and a new bathtub distribution model*, J. Statist. Plann. Inference 60 (1997), pp. 229–240.

[17] F. Proschan, *Theoretical explanation of observed decreasing failure rate*, Technometrics 5 (1963), pp. 375–383.

[18] D.R. Cox and P.A.W. Lewis, *The Statistical Analysis of Series of Events*, Methuen and Co. Ltd., London, 1966.

[19] F. Proschan and R. Pyke, *Tests for monotone failure rate*, in Proceedings of the Fifth Berkeley Symposiom of Mathematical Statistics and Probability (Berkeley, CA, 1965/66), Vol. III: Physical Sciences, University of California Press, Berkeley, CA, 1967, pp. 293–312.

[20] *Statistics Canada, Complete life table, Canada, 2000 to 2002, females and males*, (2002). Available online: http://www.statcan.ca/english/freepub/84-537-XIE/tables/txttables/cam.txt (and caf.txt).

[21] M. Bebbington, C. Lai, and R. Zitikis, *Modeling human mortality using mixtures of bathtub shaped failure distributions*, J. Theoret. Biol. 245 (2007), pp. 528–538.

Appendix 1. Proofs

LEMMA A.1 *There exists a unique minimiser of the function $\psi(h)$ over $\mathcal{K}(a)$ and over $\mathcal{K}$. Moreover, in each case, the minimiser has support of at most size $n+1$, where n is the sample size. It follows that the minimiser has finite total measure.*

This result tells us that both the MLE and constrained MLE are well-defined. The proof is omitted, as it is a slight modification of the proof of Propositions 3.2 and 3.5 in [11]. The interested reader is directed to [13] for the details.

Proof of Proposition 2.1 Since $\mathcal{K}(a_1) \subset \mathcal{K}$, the first inequality is clearly true. It remains to prove the second.

First, by arguing along the lines of the proof of Lemma 2.1 of [11], one may show that the function $\hat{h}_{n,a}$ minimises $\psi(h)$ over $\mathcal{K}(a)$ if and only if it satisfies

$$\int_0^\infty \hat{H}_{n,a}(t)\mathrm{d}\mathbb{F}_n(t) = 1 - \frac{1}{n}. \tag{A1}$$

$$\int_0^\infty \frac{1}{\hat{h}_{n,a}(t)} \mathbb{I}_{t\neq X_{(n)}} \mathrm{d}\mathbb{F}_n(t) \leq \int_{[0,\infty)} t\,\mathrm{d}\mathbb{F}_n(t), \tag{A2}$$

$$\int_0^x \frac{x-t}{\hat{h}_{n,a}(t)} \mathbb{I}_{t\neq X_{(n)}} \mathrm{d}\mathbb{F}_n(t) \leq \frac{x^2}{2} - \int_{[0,x]} \frac{(x-t)^2}{2} \mathrm{d}\mathbb{F}_n(t) \tag{A3}$$

for all $x \in [0, a]$, with equality at all $\tau_1, \ldots, \tau_k$.

$$\int_x^\infty \frac{t-x}{\hat{h}_{n,a}(t)} \mathbb{I}_{t\neq X_{(n)}} \mathrm{d}\mathbb{F}_n(t) \leq \int_{[x,\infty)} \frac{(t-x)^2}{2} \mathrm{d}\mathbb{F}_n(t), \tag{A4}$$

for all $x \in [a, X_{(n)}]$, with equality at all $\eta_1, \ldots, \eta_m$.

Let $h_i = \operatorname{argmin}_{h\in\mathcal{K}(a_i)}\psi(h)$. A simple calculation shows that

$$\begin{aligned} 0 &\geq \psi(h_0) - \psi(h_1) \\ &\geq \int_0^\infty \left\{ H_0(t) - H_1(t) + \left(1 - \frac{h_0(t)}{h_1(t)}\right)\mathbb{I}_{t\neq X_{(n)}} \right\} d\mathbb{F}_n(t) \end{aligned}$$

Using the decomposition (3) of h_0 with mixing measures α_0, μ_0, and ν_0, this last display may be rewritten as

$$\begin{aligned} &\alpha_0\left\{\int_0^\infty \left(t - \frac{1}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right)\mathrm{d}\mathbb{F}_n(t)\right\} + \left\{1 - \frac{1}{n} - \int_0^\infty H_1(t)\mathrm{d}\mathbb{F}_n(t)\right\} \\ &\quad + \int_0^{a_0}\left\{\frac{x^2}{2} - \int_0^x \frac{(x-t)^2}{2}\mathrm{d}\mathbb{F}_n(t) - \int_0^x \frac{x-t}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\mathrm{d}\mathbb{F}_n(t)\right\}\mathrm{d}\nu_0(x) \\ &\quad + \int_{a_0}^\infty \left\{\int_x^\infty \frac{(t-x)^2}{2}\mathrm{d}\mathbb{F}_n(t) - \int_x^\infty \frac{t-x}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\mathrm{d}\mathbb{F}_n(t)\right\}\mathrm{d}\mu_0(x) \\ &\geq \int_{a_0}^{a_1}\left\{\int_x^\infty \frac{(t-x)^2}{2}\mathrm{d}\mathbb{F}_n(t) - \int_x^\infty \frac{t-x}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\mathrm{d}\mathbb{F}_n(t)\right\}\mathrm{d}\mu_0(x), \end{aligned}$$

where the last inequality follows from Equations (A1)–(A4). Now, if $\mu_0 \equiv 0$ on $[a_0, a_1]$ then h_0 has an antimode at a_1, and hence h_0 and h_1 must be equal. In this case, there is nothing to prove, and hence we assume that μ_0 has positive mass on $[a_0, a_1]$. It follows that there exists a $y \in [a_0, a_1]$ such that

$$\int_y^\infty \left\{\frac{(t-y)^2}{2} - \frac{t-y}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\} d\mathbb{F}_n(t) \leq 0.$$

Combining this with Equation (3) applied to h_1 on $[a_0, a_1]$, we obtain that

$$\frac{y^2}{2} - \int_0^y \left\{\frac{(y-t)^2}{2} + \frac{y-t}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\}\mathrm{d}\mathbb{F}_n(t) + \int_y^\infty \left\{-\frac{(t-y)^2}{2} - \frac{y-t}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\}\mathrm{d}\mathbb{F}_n(t)$$

is positive, which implies that

$$\int_0^\infty \left\{\frac{y^2}{2} - \frac{(t-y)^2}{2} + \frac{t-y}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\} d\mathbb{F}_n(t) \geq 0$$

for $y \in [a_0, a_1]$. Next, write this expression as a function of y

$$\begin{aligned} f(y) &= \int_0^\infty \left\{\frac{y^2}{2} - \frac{(t-y)^2}{2} + \frac{t-y}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\}\mathrm{d}\mathbb{F}_n(t) \\ &= \int_0^\infty \left\{\frac{t}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}} - \frac{t^2}{2}\right\}\mathrm{d}\mathbb{F}_n(t) + y\int_0^\infty \left\{t - \frac{1}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\}\mathrm{d}\mathbb{F}_n(t). \end{aligned}$$

By Equation (A2) this is an increasing function in y. Therefore, it follows that $f(y) \geq 0$ holds also for all $y \geq a_1$. Now, for h_1

$$\int_x^\infty \left\{\frac{(t-x)^2}{2} - \frac{t-x}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\}\mathrm{d}\mathbb{F}_n(t) \geq 0$$

holds by Equation (A4) for all $x \geq a_1$. Which, as $f(x) \geq 0$ for $x \geq a_1$ implies that

$$\int_0^\infty \left\{\frac{x^2}{2} - \frac{(t-x)^2}{2} + \frac{t-x}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\}\mathrm{d}\mathbb{F}_n(t) + \int_x^\infty \left\{\frac{(t-x)^2}{2} - \frac{t-x}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right\}\mathrm{d}\mathbb{F}_n(t)$$

is again greater than or equal to zero. This may in turn be rewritten as

$$\frac{x^2}{2} - \int_0^x \frac{(t-x)^2}{2} d\mathbb{F}_n(t) \geq \int_0^x \frac{x-t}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \quad \text{for all} x \geq a_1. \tag{A5}$$

Finally, calculating as above, we obtain the desired inequality

$$\begin{aligned}
\psi(h_2) - \psi(h_1) &\geq \int_0^\infty \left\{ H_2(t) - H_1(t) + \left(1 - \frac{h_2(t)}{h_1(t)}\right) \mathbb{I}_{t \neq X_{(n)}} \right\} d\mathbb{F}_n(t) \\
&= \alpha_2 \left\{ \int_0^\infty \left(t - \frac{1}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} \right) d\mathbb{F}_n(t) \right\} + \left\{ 1 - \frac{1}{n} - \int_0^\infty H_1(t) d\mathbb{F}_n(t) \right\} \\
&\quad + \int_0^{a_2} \left\{ \frac{x^2}{2} - \int_0^x \frac{(x-t)^2}{2} \mathrm{d}\mathbb{F}_n(t) - \int_0^x \frac{x-t}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} \mathrm{d}\mathbb{F}_n(t) \right\} \mathrm{d}\nu_2(x) \\
&\quad + \int_{a_2}^\infty \left\{ \int_x^\infty \frac{(t-x)^2}{2} \mathrm{d}\mathbb{F}_n(t) - \int_x^\infty \frac{t-x}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} \mathrm{d}\mathbb{F}_n(t) \right\} \mathrm{d}\mu_2(x) \\
&\geq \int_{a_1}^{a_2} \left\{ \int_0^x \frac{(t-x)^2}{2} d\mathbb{F}_n(t) - \int_0^x \frac{x-t}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} \mathrm{d}\mathbb{F}_n(t) \right\} \mathrm{d}\mu_0(x) \geq 0,
\end{aligned}$$

by Equations (A1)–(A4) and (A5). A similar argument proves the inequality $\min_{h \in \mathcal{K}(a_1)} \psi(h) \leq \min_{h \in \mathcal{K}(a_2)} \psi(h)$ for $a_2 < a_1 < a_0$. ■