COMPUTATION OF NONPARAMETRIC CONVEX HAZARD ESTIMATORS VIA PROFILE METHODS

TECHNICAL REPORT 542
DEPARTMENT OF STATISTICS, UNIVERSITY OF WASHINGTON

HANNA K. JANKOWSKI, JON A. WELLNER

ABSTRACT. In this paper we develop an algorithm to find the maximum likelihood estimator of a convex hazard function. The maximization is done in two steps. First, we use the support reduction algorithm of [GJW1] to find the profile likelihood over a constrained space. We next show that (-1) times the profile likelihood is bathtub-shaped in the parameters, and use a bisection algorithm to find the overall maximizer. We use the same approach to find a least squares estimator of a convex hazard rate. Simulations and data examples are also given.

1. INTRODUCTION

Suppose we observe $X_1, \ldots, X_n$ i.i.d. with density f. The X_i's 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, $h(t)dt$ is the probability that, given survival until time t, the event will occur in the next dt amount of time. The hazard function is also known as the force of mortality in actuarial science, or the intensity function in extreme value theory.

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[1] and then increasing. 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. The observed failure rate is then a mixture of these three types of failure, as seen in Figure 1. We will say that a bathtub shaped function h has an *antimode* at a if it

Date: September 15, 2008.

Key words and phrases. nonparametric estimation, profile likelihood, convex hazard rate, antimode, support reduction, active set method.

[1]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.

is non-increasing on $[0, a]$ and non-decreasing 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 [Gre] who considered the maximum likelihood estimator (MLE) of an increasing hazard rate. The MLE in this case may be found exactly by using either a graphical representation (via derivatives of the concave majorant of the time on test statistic) or the pool-adjacent violators algorithm. Later, [BCP] extended this work 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. [PR]).

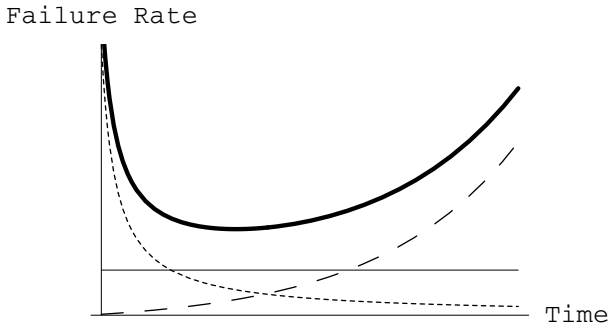

FIGURE 1. Example when the observed failure (bold) is equal to the mixture of the infant mortality (short-dashed), constant, and wear-out (long-dashed) failure rates.

To find the MLE, $\hat{h}_n$, we first consider the likelihood written in term of the hazard rate

$$\mathcal{L}ik(h) = \prod_{i=1}^{n} h(X_i) \exp\left\{ - \int_0^{X_i} h(t)dt \right\}. \tag{1.1}$$

The goal is to find the convex positive function h which maximizes $\mathcal{L}ik(h)$. However, $\mathcal{L}ik(h)$ can be made arbitrarily large by increasing the value of $h(X_{(n)})$. We therefore maximize the modified likelihood

$$\mathcal{L}^{mod}(h) = \prod_{i=1}^{n-1} h(X_i) \exp\left\{-H(X_i)\right\} \times \exp\left\{-H(X_{(n)})\right\}. \tag{1.2}$$

and set $\widehat{h}_n(X_{(n)})$ to be arbitrarily large (i.e. $\widehat{h}_n(X_{(n)}) = \infty$) to find the MLE. That is, the MLE on $[0, X_{(n)})$ is found by maximizing $\mathcal{L}^{mod}(h)$, and it is set to $+\infty$ for all $x \geq X_{(n)}$. This is the same approach as taken in the monotone case, see e.g. [Gre] page

142, or [RWD], page 338. Let $\mathcal{K}$ denote the space of non-negative convex functions with domain $[0, X_{(n)})$. Then the MLE may be re-written as

$$\hat{h}_n = \text{argmin}_{h\in\mathcal{K}}\psi(h),$$

for

$$\psi(h) = \int_0^\infty \left[\int_0^t h(s)ds - \log h(t)\mathbb{I}_{t\neq X_{(n)}}\right] d\mathbb{F}_n(t) \tag{1.3}$$

where $\mathbb{F}_n$ is the empirical cumulative distribution function. In Section 3, we provide a detailed characterization of the maximum likelihood estimator. It may be defined via a set of equalities and inequalities; However, there is currently no explicit solution of these inequalities, and computational techniques are necessary.

Let $\mathcal{K}(a)$ denote the subspace of $\mathcal{K}$ of convex positive function with an antimode at a. Consider then the profile likelihood of a:

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

We show in Section 3.2 that $-\log \mathcal{L}ik^{mod}(a)$ is itself bathtub-shaped in a, and therefore propose the following, two-step, optimization method

- [bis] maximize $\mathcal{L}ik^{mod}(a)$ with a bisection algorithm.
- [SR] use the support reduction algorithm developed in [GJW1] to maximize $\mathcal{L}ik(h)$ over $\mathcal{K}(a)$.

That is, we find $\hat{h}_n$, via two minimizations:

$$\underbrace{\min_{a}}_{\text{[bis]}} \underbrace{\min_{h\in\mathcal{K}(a)}}_{\text{[SR]}} \left\{\int_0^\infty \left[\int_0^t h(s)ds - \log h(t)\mathbb{I}_{t\neq X_{(n)}}\right] d\mathbb{F}_n(t)\right\}.$$

For a fixed antimode a, a positive convex function in $\mathcal{K}(a)$ may be decomposed in terms of its *mixing measure* and *support* as

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

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 ν, μ and α the mixing measure of h, and the support of these measures becomes the support of h. Note that by definition, the support of ν is contained in $[0, a]$ and the support of μ is contained in $[a, X_{(n)}]$. The *total measure* of a function h is then $\alpha + \nu[0, a] + \mu[a, X_{(n)}]$. We also use the term *basis* functions for $1, (\tau - t)_+$ for τ in $[0, a]$ and $(t - \eta)_+$ for η in $[a, X_{(n)}]$. Thus, if $h(t) = 0.5 + 3\cdot(2 - t)_+$ we will say that it has support $supp = \{1\} \times \{2\} \times \emptyset$ and mixing measure $mix = \{0.5\} \times \{\delta_3\} \times \emptyset$. Proposition 3.1 shows that the support of $\hat{h}_n$ is always finite for a fixed sample size. In fact, in practice the number of support points is considerably smaller than n.

The support reduction (SR) algorithm as developed by [GJW1], is an extension of the *vertex direction algorithm*, previously developed by [Fed, Sim, Wyn] (see also, [Böh1, Böh2, LK]). Within optimization theory, the SR algorithm can be classified as an active set method. The algorithm is designed to handle nonparametric and semiparametric 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 support reduction algorithm works particularly well.

The main idea behind the support reduction algorithm is as follows. We wish to minimize a criterion function $\psi(h) = -\log \mathcal{L}ik^{mod}(h)$ over the space of $h \in \mathcal{K}(a)$ with decomposition (1.4). Given a current iterate $\hat{h}$ 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

$$\lim_{\varepsilon \to 0} \frac{\psi\left(\hat{h} + \varepsilon e\right) - \psi\left(\hat{h}\right)}{\varepsilon}$$

is smallest. The support corresponding to e^* is added to $\widehat{supp}$ to yield $\widehat{supp}^*$, and then ψ is minimized over all h with support given by $\widehat{supp}^*$ to give the new mixing measure. This is the *vertex direction* part of the algorithm: the idea is to continually move in a direction which decreases the criterion function the most. The support reduction algorithm adds an additional step, which insures that throughout the algorithm we remain in the constrained space; in this case this is equivalent to requiring that the mixing measure remain positive.

Section 3 gives the details of the implementation for the MLE. The section also contains all of the necessary theoretical results. However, as the practical implementation of the algorithm for the MLE requires several technical adjustments, we also discuss here the least squares estimator (LSE) for which the exposition is more straightforward.

Let $\mathbb{H}_n$ denote the cumulative empirical hazard function

$$\mathbb{H}_n(t) = \int_0^t \frac{1}{1 - \mathbb{F}_n(s-)} d\mathbb{F}_n(s),$$

and fix a $0 < T < \infty$. The LSE is defined as

$$\tilde{h}_n = \text{argmin}_{h \in \mathcal{K}_T} \left\{ \frac{1}{2} \int_0^T h^2(t) dt - \int_0^T h(t) d\mathbb{H}_n(t) \right\},$$

where $\mathcal{K}_T$ denotes the space of positive convex functions on $[0, T]$. This can be motivated as follows. Suppose that $\mathbb{H}_n$ is absolutely continuous with respect to Lebesgue measure so that $d\mathbb{H}_n(t) = h_n^*(t)dt$ makes sense. Then the LSE could be found as the

minimizer of

$$\begin{aligned}
&\frac{1}{2}\int_0^T (h_n(t) - h_n^*(t))^2 dt \\
&= \frac{1}{2}\int_0^T h^2(t)dt - \int_0^T h(t)h_n^*(t)dt + \frac{1}{2}\int_0^T (h_n^*)^2(t)dt \\
&= \frac{1}{2}\int_0^T h^2(t)dt - \int_0^T h(t)d\mathbb{H}_n(t) + C(X_1, \ldots, X_n),
\end{aligned}$$

where the term $C(X_1, \ldots, X_n)$ depends on the data *only*, and does not impact the minimization.

Since the implementation of our scheme is simpler for the LSE, this is given in detail in Section 2. Examples for both estimators are considered in Section 4. Technical proofs are collected in the Appendix.

The asymptotic theory of both the maximum likelihood and least squares estimators was studied in-depth in [JW], and we direct the interested reader there for the theory, and also for a more thorough review of the history of the problem. For the purposes of this report, it sufficient to note that both the MLE and LSE are consistent for the true hazard function, and, under the assumptions of strict convexity, exhibit an $n^{2/5}$ local rate of convergence. In [JW] we conjecture that if the true hazard function has a second derivative equivalent to zero, then both estimators will achieve a global rate of convergence of $n^{1/2}$. Using the algorithm we are able to provide further evidence for this conjecture, see Figure 6 in Section 4 and the associated discussion.

We also note that in [JW], hazard estimation under right censoring, and the estimation of a convex Poisson intensity are studied. The techniques described here may be extended to those settings as well.

The algorithms described here are available through the R package convexHaz, [JWMW]. Currently this includes the MLE and LSE for hazard estimation, but we hope to include right censoring and Poisson intensity estimation in future versions.

2. Least Squares Estimator

To find the LSE, we need to find the minimizer of

$$\varphi(h) = \frac{1}{2}\int_0^T h^2 dt - \int_0^T h d\mathbb{H}_n,$$

over the space $\mathcal{K}_T$, the space of nonnegative convex hazard functions on $[0, T]$. Proposition 3.1 in [JW] shows that the minimizer of φ exists and is unique, and hence the problem is well-defined. We denote the antimode of $\tilde{h}_n$ by a_0.

Let $\mathcal{K}_T(a)$ denote the class of nonnegative convex functions on $[0, T]$ with antimode occurring at a. Proposition 2.6 shows that $\widetilde{\varphi}(a) = \min_{h \in \mathcal{K}_T(a)} \varphi(h)$ is bathtub shaped as a function of a. Clearly, its minimum lies at the "target", namely the antimode a_0 of $\tilde{h}_n$.

As for the MLE, we perform the minimization in two steps. The bisection algorithm is used to search over $a \in [0, T]$ to find a_0, and the support reduction algorithm is used to find $\widetilde{\varphi}(a)$ for fixed a. The SR algorithm also finds the unique function which minimizes $\varphi(h)$ over $\mathcal{K}_T(a)$, and hence, when $a = a_0$, this yields the LSE. Proposition 2.2 shows that the constrained LSE exists and is unique. Also, it has finite support, and hence the support reduction is quite efficient.

The rest of this section is organized as follows. In Section 2.1 we give the details of the support reduction algorithm for the constrained least squares problem. Section 2.2 discusses briefly the bisection algorithm we use. Lastly, necessary technical results are collected in Section 2.3.

2.1. **Support Reduction: minimizing $\varphi(h)$ over $\mathcal{K}_T(a)$.** To describe the support reduction algorithm, we begin by labelling the basis functions

$$\begin{aligned} e_0(t) &\equiv 1, \\ e_{1,\tau}(t) &= (\tau - t)_+, \\ e_{2,\eta}(t) &= (t - \eta)_+. \end{aligned}$$

Now, any hazard function $\tilde{h}$ with finite support may be expressed in terms of the basis functions

$$\tilde{h} = \tilde{\alpha} \cdot e_0 + \sum_{i=1}^{k} \tilde{\nu}_i \cdot e_{1,\tau_i} + \sum_{j=1}^{m} \tilde{\mu}_j \cdot e_{2,\eta_j}.$$

In the description of the algorithm, it is useful to work with the equivalent characterization of $\tilde{h}$ in terms of the support $\widetilde{supp} = \{1\} \times \{\tau_1, \ldots, \tau_k\} \times \{\eta_1, \ldots, \eta_m\}$ and associated mixing measure $\widetilde{mix} = \{\tilde{\alpha}\} \times \{\tilde{\nu}_1, \ldots, \tilde{\nu}_k\} \times \{\tilde{\mu}_1, \ldots, \tilde{\mu}_m\}$.

Next, we calculate the directional derivatives of the criterion function φ. This is done using integration by parts. Let $\mathbb{Y}_n(t) = \int_0^t \mathbb{H}_n(s)ds$ and for any function h, let $H(t) = \int_0^t h(s)ds$ denote its integral and $\mathcal{H}(t) = \int_0^t H(s)ds$ denote its double integral. Then the directional derivatives are

$$\begin{aligned} \nabla_0\varphi(h) &= \lim_{\epsilon \to 0} \frac{\varphi(h + \epsilon\, e_0) - \varphi(h)}{\epsilon} \\ &= (H - \mathbb{H}_n)(T), \\ \nabla_1\varphi(h)[\tau] &= \lim_{\epsilon \to 0} \frac{\varphi(h + \epsilon\, e_{1,\tau}) - \varphi(h)}{\epsilon} \\ &= (\mathcal{H} - \mathbb{Y}_n)(\tau), \\ \nabla_2\varphi(h)[\eta] &= \lim_{\epsilon \to 0} \frac{\varphi(h + \epsilon\, e_{2,\tau}) - \varphi(h)}{\epsilon} \\ &= (T - \eta)(H - \mathbb{H}_n)(T) - (\mathcal{H} - \mathbb{Y}_n)(T) + (\mathcal{H} - \mathbb{Y}_n)(\eta). \end{aligned} \tag{2.1}$$

Since the criterion function φ is convex as a function of h, we know that we have found its minimizer $\tilde{h}_n$ over $\mathcal{K}_T$ if the directional derivative at $\tilde{h}_n$ is positive in *any* possible direction. The same is true when minimizing over $\mathcal{K}_T(a)$, and in this case, 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 $\eta \in [a, T]$. To this end, let $\nabla\varphi(\tilde{h})$ denote the minimum of the directional derivatives

$$\nabla\varphi(\tilde{h}) = \min\left\{\nabla_0\varphi(\tilde{h}), \min_{\tau\in[0,m_0]} \nabla_1\varphi(\tilde{h})[\tau], \min_{\eta\in[m_0,T]} \nabla_2\varphi(\tilde{h})[\eta]\right\},$$

at a current estimate $\tilde{h}$.

The support reduction algorithm now proceeds as follows. Choose an accuracy constant $\varepsilon > 0$.

SUPPORT REDUCTION ALGORITHM FOR THE CONSTRAINED LSE OF A CONVEX HAZARD:

STEP 0. Obtain an initial estimate, $\tilde{h}$. A simple option is $\tilde{h} = \tilde{\alpha}$, where $\tilde{\alpha} = \mathbb{H}_n(T)/T$ (this minimizes the criterion function for constant hazards).

WHILE $\nabla\varphi(\tilde{h})$ is less than $-\varepsilon$ **REPEAT 1-3:**

STEP 1. Given a current estimate $\tilde{h}$, find the best direction to move in. To do this, check

$$\nabla_0\varphi(\tilde{h}), \min_{\tau\in[0,a]} \nabla_1\varphi(\tilde{h})[\tau], \min_{\eta\in[a,T]} \nabla_2\varphi(\tilde{h})[\eta].$$

The argument which minimizes the smallest of these three quantities identifies the new direction, and we denote it as e^*. For example, if $\min_{\tau\in[0,a]} \nabla_1\varphi(\tilde{h})[\tau]$ is smallest, with minimum occurring at τ_{k+1}, then $e^* = e_{1,\tau_{k+1}}$.

Add the support of e^* to $\widetilde{supp}$, obtaining $\widetilde{supp}^*$.

STEP 2. Find the mixing measure $\widetilde{mix}^*$ which minimizes the criterion function $\varphi(h)$ over all functions with support $\widetilde{supp}^*$. Note that this a straightforward finite-dimensional minimization of a quadratic.

Let $\tilde{h}^*$ denote the function with support $\widetilde{supp}^*$ and mixing measure $\widetilde{mix}^*$. Unfortunately, there is no guarantee that $\tilde{h}^*$ is in $\mathcal{K}_T(a)$, since there may be negative weights in $\widetilde{mix}^*$. We therefore find the largest λ such that $(1-\lambda)\tilde{h} + \lambda\tilde{h}^*$ is in $\mathcal{K}_T(a)$. This is the support reduction step. Further details are given in Remark 2.1.

Suppose that $\widetilde{supp}^*$ is given by $\{\tilde{\alpha}^*\}\times\{\tilde{\nu}_1^*, \ldots, \tilde{\nu}_l^*\}\times\{\tilde{\mu}_1^*, \ldots \tilde{\mu}_m^*\}$. We may also express $\widetilde{supp}$ as $\{\tilde{\alpha}\} \times \{\tilde{\nu}_1, \ldots, \tilde{\nu}_l\} \times \{\tilde{\mu}_1, \ldots \tilde{\mu}_m\}$, where one element of the mixing measure (corresponding to the newly added support point) is zero.

WHILE $\min\{\tilde{\alpha}^*, \tilde{\nu}_1^*, \ldots, \tilde{\nu}_l^*, \tilde{\mu}_1^*, \ldots \tilde{\mu}_m^*\} < 0$ **REPEAT A-B:**

STEP A. Let

$$\gamma^* = \min \left\{ \frac{\tilde{\alpha}^*}{\tilde{\alpha}}, \frac{\tilde{\nu}_i^*}{\tilde{\nu}_i}, \frac{\tilde{\mu}_j^*}{\tilde{\mu}_j} \right\}_{i=1,\ldots,l; j=1,\ldots,m}.$$

Set $\lambda^* = (1-\gamma^*)^{-1}$ and calculate $\widetilde{mix} = (1-\lambda^*)\widetilde{mix} + \lambda^* \widetilde{mix}^*$, with component-wise addition. This creates the comparison iterate in case the support reduction step needs to be repeated. However, one of the entries of $\widetilde{mix}$ is equal to zero. Delete the corresponding support point and associated element of the mixing measure from $\widetilde{mix}$ and $\widetilde{supp}^*$.

STEP B. Find the mixing measure $\widetilde{mix}^*$ which minimizes the criterion function $\varphi(h)$ over all functions with support $\widetilde{supp}^*$.

STEP 3. Set the current estimate $\tilde{h} = \tilde{h}^*$.

Remark 2.1. *How does the support reduction step work? For simplicity, consider the case when* $\widetilde{mix}^*$ *is given by* $\emptyset \times \emptyset \times \{\tilde{\mu}_1^*, \ldots \tilde{\mu}_m^*\}$, *with* $\widetilde{mix} = \emptyset \times \emptyset \times \{\tilde{\mu}_1, \ldots \tilde{\mu}_m\}$. *To find the largest* λ *such that* $(1-\lambda)\tilde{h} + \lambda\tilde{h}^*$ *is in* $\mathcal{K}_T(a)$, *we need to find the largest* λ *such that* $(1-\lambda)\tilde{\mu}_i + \lambda\tilde{\mu}_i^* \geq 0$ *for* $i = 1, \ldots, m$. *Choosing the largest such* λ *gives the most weight to the newly proposed* $\tilde{h}^*$, *and hence this is the desirable choice.*

Since the desired inequality holds whenever $\tilde{\mu}_j^* \geq 0$, *we restrict the search to the negative weights. For these, we have that* $(1-\lambda)\tilde{\mu}_i + \lambda\tilde{\mu}_i^* \geq 0$ *if and only if* $\lambda \leq \tilde{\mu}_i/(\tilde{\mu}_i - \tilde{\mu}_i^*)$, *and hence the largest such* λ *is given by*

$$\lambda^* = \min_{i:\tilde{\mu}_i^* < 0} \left\{ \frac{\tilde{\mu}_i}{\tilde{\mu}_i - \tilde{\mu}_i^*} \right\} = \frac{1}{1 - \min_{i=1,\ldots,m}(\tilde{\mu}_i^*/\tilde{\mu}_i)}.$$

Clearly, there exists an i_0 *such that* $(1-\lambda)\tilde{\mu}_{i_0} + \lambda\tilde{\mu}_{i_0}^* = 0$.

In addition, note that Proposition 2.7 guarantees that a newly proposed point will not be removed by this step.

Figure 2 shows an example of the support reduction algorithm for a random sample of size 100 from a Weibull distribution with hazard function $h(t) = 2t$. The algorithm converged after 11 iterations. We set $a = 0, T = 1.5$, and $\varepsilon = 10^{-8}$.

2.1.1. *Practical considerations.* There are several computational issues that arise in the implementation of the algorithm.

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 minimize the gradient over

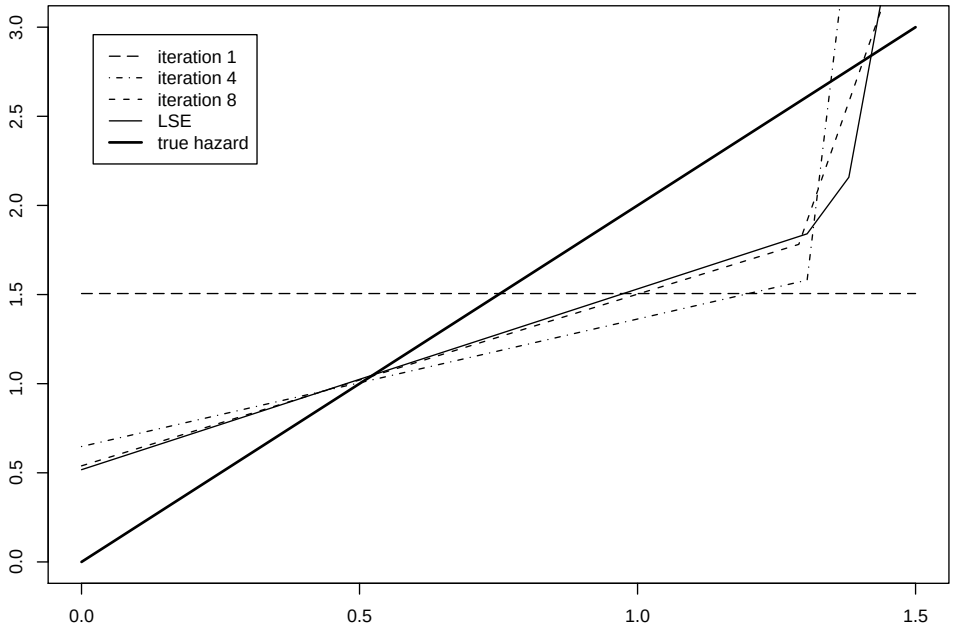

FIGURE 2. Example of convergence in the support reduction algorithm: The example was done using the R package convexHaz with $T = 1.5$ and the antimode set at zero. The LSE is obtained as the 11th iterate.

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 $[0, T]$ into M intervals (resulting in $M + 1$ grid points), and only checking for the minimum at these locations. Naturally, the larger M is, the more accurate our answer; However, increasing M also increases computing time. For example, to obtain the results in Figure 2, M was set to 100 (the default). For a sample size of 100, the computing time takes seconds. However, as we ultimately iterate the support reduction algorithm in the bisection step, it is important to keep computing time to a minimum.

A gridless alternative. Suppose that the grid used in the algorithm is such that $G = \{\theta_1, \ldots, \theta_{M+1}\}$, and suppose also that in Step 1 the algorithm proposed the new support point $\theta \in G$. 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 Step 2 of the algorithm, but will impose a finer grid when the criterion $\nabla\varphi(\tilde{h}) \leq -\varepsilon$ 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. A comparison of the gridded vs. gridless implementations is provided in Section 4.

2.2. **The Bisection Algorithm.** We next briefly describe the bisection algorithm used in the outer minimization step. Proposition 2.8 shows that the "profile" criterion function $\tilde{\varphi}(a) = \min_{h \in \mathcal{K}(a)} \varphi(h)$ is itself bathtub shaped in a. Although $\tilde{\varphi}(a)$ is not convex, and therefore the bisection algorithm is not guaranteed to converge to the overall minimum, we have found that it works quite well in practice (see for example Figure 5). The alternative is of course to do a gridded search over $a \in [0, T]$, but this is much less efficient, and equally not guaranteed to find the minimum.

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 $\boldsymbol{a} = \{a_i\}_{i=1}^5 = \{0, T/4, T/2, 3T/4, T\}$, and find $\widetilde{\boldsymbol{\varphi}} = \{\widetilde{\varphi}(a_i)\}_{i=1}^5$.

WHILE $\Delta\widetilde{\boldsymbol{\varphi}}$ is greater than ε **REPEAT 1-2:**

STEP 1. Write $\tilde{\boldsymbol{\varphi}}$ as $\{\tilde{\varphi}_i\}_{i=1}^5$. If $\tilde{\varphi}_{(1)} = \tilde{\varphi}_i$ 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 $\tilde{\varphi}_{(1)} = \tilde{\varphi}_1$, set $\tilde{a}_1 = a_1, \tilde{a}_3 = a_2$ and $\tilde{a}_5 = a_3$. If $\tilde{\varphi}_{(1)} = \tilde{\varphi}_5$, 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 $\widetilde{\boldsymbol{\varphi}} = \{\widetilde{\varphi}(a_i)\}_{i=1}^5$. Note that three of the five entries have already been calculated.

STEP 3. $\operatorname{argmin} \widetilde{\varphi}(a)$ is given by the a_i which minimizes the current $\tilde{\boldsymbol{\varphi}}$.

Clearly, many choices for $\Delta\boldsymbol{x}$ exist. Although our choice of the squared difference is actually quite mild, we find that it works quite well in practice. It appears to do a solid job of finding the minimum, while keeping the number of iterations low.

Estimating the antimode. In [JW] we show that if the true hazard function has a unique antimode at $\tilde{a}$, then the antimode of the LSE, $\tilde{h}_n$, will converge to $\tilde{a}$ as the sample size, n, tends to infinity. In fact, we conjecture that the rate of this convergence is $n^{1/5}$. However, we would like to point out that our methods are not optimized to find the estimator of the antimode.

2.2.1. *Practical Considerations.* **Gridded implementation.** Proposition 2.8 shows that the theoretical value of $\widetilde{\varphi}(a)$ 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 $\widetilde{\varphi}(a)$. Fortunately, we have found that this approximation does not invalidate the bathtub shape, and the same is true of the gridless implementation. An example is given in Figure 5.

2.3. **Some Technical Results.** In this section we collect the technical results for least squares estimation. First of all, we check that the constrained least squares estimator is well-defined; this is done in Theorem 2.2. Next, we look at the true characterization of both the overall and constrained LSEs.

In [GJW1], conditions on convergence of the support reduction algorithm are given, and we check that these hold in our case in Proposition 2.6. Proposition 2.7 shows that the support reduction step will never remove a newly proposed point, thus avoiding a potential infinite loop. Lastly, in Proposition 2.8 we show that the profile least squares criterion, $\tilde{\varphi}(a)$ is bathtub shaped.

Lemma 2.2. *There exists a unique minimizer of the function $\varphi(h)$ over $\mathcal{K}_T(a)$. Moreover, the minimizer has support of at most size $n+1$, where n is the sample size. It follows that the minimizer has finite total measure.*

Remark 2.3. *For any function h, we use the notation $H(t) = \int_0^t h(s)ds$ and $\mathcal{H}(t) = \int_0^t H(s)ds$.*

Lemma 2.4 (Characterization of LSE)**.** *The function $\tilde{h}_n$ minimizes $\varphi(h)$ over $\mathcal{K}_T$ if and only if it satisfies*

$$\tilde{H}_n(T) = \mathbb{H}_n(T), \tag{2.2}$$

$$\tilde{\mathcal{H}}_n(T) = \mathbb{Y}_n(T), \tag{2.3}$$

$$\tilde{\mathcal{H}}_n(t) \geq \mathbb{Y}_n(t) \text{ for all } t \in [0,T], \tag{2.4}$$

$$\int_0^T (\tilde{\mathcal{H}}_n - \mathbb{Y}_n)(t) d\, \tilde{h}_n'(t) = 0. \tag{2.5}$$

The last statement is the same as: $\tilde{\mathcal{H}}_n(x) = \mathbb{Y}_n(x)$ for all changes of slope x of $\tilde{h}_n$.

This result is proven in [JW].

Lemma 2.5 (Characterization of constrained LSE)**.** *Let* $\mathbb{Y}_n(t) = \int_0^t \mathbb{H}_n(s)ds$. *The function* $\tilde{h}_{n,a}$ *minimizes* $\varphi(h)$ *over* $\mathcal{K}_T(a)$ *if and only if it satisfies*

$$\tilde{H}_{n,a}(T) \geq \mathbb{H}_n(T), \tag{2.6}$$

$$\tilde{\mathcal{H}}_{n,a}(t) - \mathbb{Y}_n(t) \geq 0, \text{ for all } t \in [0,a] \tag{2.7}$$

with equality at all $\tau_1, \ldots, \tau_k$.

$$(T-t)[\tilde{H}_{n,a} - \mathbb{H}_n](T) - \int_t^T [\tilde{H}_{n,a} - \mathbb{H}_n](s)ds \geq 0 \text{ for all } t \in [a,T], \tag{2.8}$$

with equality at all $\eta_1, \ldots, \eta_m$.

$$\tilde{h}_{n,a}(T)[\tilde{H}_{n,a} - \mathbb{H}_n](T) - \tilde{h}'_{n,a}(T)[\tilde{\mathcal{H}}_{n,a} - \mathbb{Y}_n](T) + \int_0^T (\tilde{\mathcal{H}}_{n,a} - \mathbb{Y}_n)(t)d\ \tilde{h}'_{n,a}(t) = 0. \tag{2.9}$$

Essentially, the characterization says that $\tilde{h}_{n,a}$ is the unique minimizer if and only if the directional derivatives (2.1) are positive in all directions, and equal to zero in the direction of $\tilde{h}_{n,a}$. These are the standard Fenchel conditions. Since the space is infinitely dimensional though, some work is necessary to make the statement rigorous.

In [GJW1], three assumptions on the criterion function are described as sufficient so that the support algorithm converges to the true minimizer. We next show that the LSE criterion $\varphi(\cdot)$ satisfies these.

Proposition 2.6. *The criterion function for the LSE*

$$\varphi(h) = \frac{1}{2}\int_0^T h^2 dt - \int_0^T h d\mathbb{H}_n,$$

satisfies the conditions

- A1. φ *is convex on* $\mathcal{K}_T(a)$ *and* $\varphi(h + t(g-h))$ *is continuously differentiable as a function of* t, *for* $t \in (0,1)$.
- A2. *The directional derivative* $\nabla\varphi(h)[g]$ *is linear in* g.
- A3. *For any specific function* $h_0 \in \mathcal{K}_T(a)$ *with* $\varphi(h_0) < \infty$, *there exists an* $\bar{\varepsilon} \in (0,1]$ *such that for all* $h \in \mathcal{K}_T(a)$ *with* $\varphi(h) < \varphi(h_0)$ *and any basis function* e, *the following implication holds*

$$\nabla\varphi(h)[e-h] \leq -\delta \quad \Rightarrow \quad \varphi(h + \varepsilon(e-h)) - \varphi(h) \leq -\frac{1}{2}\varepsilon\delta \quad \text{for all } \varepsilon \in (0, \bar{\varepsilon}].$$

Let h_k *denote a sequence generated by the support reduction algorithm. Also, suppose that in each iteration the new support point, corresponding to a basis function* e, *is chosen so that*

$$\nabla\varphi(h_k)(e) \leq \frac{1}{2}\inf_{e' : basis} \nabla\varphi(h_k)(e'). \tag{2.10}$$

Then, $\varphi(h_k) \to \varphi(\tilde{h}_n)$.

The conditions are not too difficult to understand: the first allows us to take derivatives, the second restricts the exit condition to the basis functions, and the third ensures that each iteration gets closer to the true minimizer. Note that the last condition, (2.10), holds trivially within the algorithm if the basis functions considered in the infimum consist of all those with support in the finite set of points corresponding to the grid. However, this does not correspond to the overall minimizer of φ_n. To find the overall minimizer, one needs to consider all of the possible support points within the continuum $[0, a]$ and $[a, X_{(n)}]$. Therefore, the algorithm finds an approximation to the least squares estimator.

Since φ_n is convex, the minimizer, $\tilde{h}_n$, is the unique hazard function which satisfies the Fenchel conditions

$$\begin{aligned} &\text{(i).} \quad \nabla\varphi_n(\tilde{h}_n)[e] \geq 0, \text{ for all basis functions } e, \\ \text{and} \quad &\text{(ii).} \quad \nabla\varphi_n(\tilde{h}_n)[\tilde{h}_n] = 0. \end{aligned}$$

This is easily justified heuristically (a convex function has zero derivative at the minimum and has an increasing derivative in all other directions), and is also not difficult to prove (cf. Lemma 1 in [GJW1]). Let h_K denote the result of the last iteration of the support reduction algorithm. The gridded implementation of the algorithm achieves

$$\begin{aligned} H_K(T) - \mathbb{H}_n(T) = \nabla\varphi_n(h_K)[e_0] &\geq -\varepsilon, \\ \nabla\varphi_n(h_K)[e_{1,\tau}] &\geq -\varepsilon, \text{ for all } \tau \leq a \text{ in the grid} \\ \nabla\varphi_n(h_K)[e_{2,\eta}] &\geq -\varepsilon, \text{ for all } \eta \geq a \text{ in the grid.} \end{aligned}$$

It is natural to ask to what extent this implies the full Fenchel conditions. Suppose that τ_1 is in the grid, and τ_2 does not. Then,

$$\begin{aligned} \nabla\varphi_n(h_K)[e_{1,\tau_2}] &= \nabla\varphi_n(h_K)[e_{1,\tau_1}] + \nabla\varphi_n(h_K)[e_{1,\tau_2} - e_{1,\tau_1}] \\ &\geq -\varepsilon - ||e_{1,\tau_2} - e_{1,\tau_1}||_\infty \{H_K(T) + \mathbb{H}_n(T)\} \\ &= -\varepsilon - |\tau_2 - \tau_1| \{H_K(T) + \mathbb{H}_n(T)\} \end{aligned}$$

Similarly,

$$\nabla\varphi_n(h_K)[e_{1,\eta_2}] \geq -\varepsilon - |\eta_2 - \eta_1| \{H_K(T) + \mathbb{H}_n(T)\}.$$

Also, $\nabla\varphi_n(h_K)[h_K] = 0$ is always satisfied by nature of the algorithm. Therefore, if we could bound $H_K(T)$ from above, we would obtain a bound on the error caused by the gridded approximation. Unfortunately, this bound is only possible if $e_0 \equiv 1$ is in the characterization, as then we automatically obtain $H_K(T) = \mathbb{H}_n(T)$.

The next result guarantees that when a new support point is added by the algorithm, then it will always have positive weight assigned to it. This ensures that the algorithm does not enter an infinite loop.

Proposition 2.7. *Suppose that τ_{k+1} is the new support point added. Then, using the notation used in the description of the algorithm, $\tilde{\nu}^*_{k+1} > 0$. If η_{m+1} is the new support point added, or if 1 is the new support point added, then $\tilde{\mu}^*_{m+1} > 0$ or $\tilde{\alpha}^* > 0$.*

Proposition 2.8. *Let $a_0 \in [0, T]$ be such that the minimizer of $\varphi(h)$ over h in $\mathcal{K}_T$ has an antimode at a_0. Suppose that $a_0 < a_1 < a_2$. Then*

$$\min_{h\in\mathcal{K}_T} \varphi(h) \equiv \min_{h\in\mathcal{K}_T(a_0)} \varphi(h) \leq \min_{h\in\mathcal{K}_T(a_1)} \varphi(h) \leq \min_{h\in\mathcal{K}_T(a_2)} \varphi(h).$$

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

3. Maximum Likelihood Estimator

In this section we describe how to obtain the maximum likelihood estimator, $\hat{h}_n$. As we discuss in the introduction, the approach is similar to that for the LSE, and therefore we only describe how to find the profile likelihood (the bisection algorithm being the same). Additional technical difficulties arise for the MLE. In this case, it is practically not possible to perform step 2 of the algorithm: given the support points, one cannot minimize the criterion function in order to find the new mixing measure. In the following section we outline a method which may be used to overcome this difficulty. The method is the same as used in [GJW3, Bal].

3.1. **Support Reduction: minimizing $\psi(h)$ over $\mathcal{K}(a)$.** The idea of the modification is to minimize an approximate version of ψ instead of the true ψ (defined in (1.3)). Suppose that the current iterate $\hat{h}$ in the algorithm is close to the true minimizer of ψ, then instead of minimizing ψ, we could equally well minimize the quadratic approximation to ψ. This *inner* minimization is iterated, taking the result as the new $\hat{h}$ after each loop, until the directional derivatives of the true ψ are sufficiently large.

To give the details, we define the approximate criterion function with respect to a fixed function, which we will call g. Notice that for the approximation to be close, we assume that g is close to the true minimizer of ψ. We use here the approximation $\log(1 + x) \approx x - x^2/2$. Fix a function $g \in \mathcal{K}(a)$. Then $\psi(h)$ is equal to

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

The goal will be to minimize the above approximation in h for a fixed g, and hence we remove all terms depending only on g to obtain

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

The same basis functions apply here as for the LSE algorithm. We therefore define two sets of directional derivatives.

First we look at the directional derivatives for the true criterion function.

$$\begin{aligned}
\nabla_0\psi(h) &\equiv \lim_{\epsilon\to 0} \frac{\psi(h+\epsilon e_0) - \psi(h)}{\epsilon} \\
&= \int_0^\infty \left(t - \frac{1}{h(t)}\mathbb{I}_{t\neq X_{(n)}}\right) d\mathbb{F}_n(t), \\
\nabla_1\psi(h)[\tau] &\equiv \lim_{\epsilon\to 0} \frac{\psi(h+\epsilon e_{1,\tau}) - \psi(h)}{\epsilon} \\
&= \int_0^\infty \left(t\wedge\tau\left(2\tau - \frac{t\wedge\tau}{2}\right) - \frac{(\tau-t)_+}{h(t)}\right) d\mathbb{F}_n(t), \\
\nabla_2\psi(h)[\eta] &\equiv \lim_{\epsilon\to 0} \frac{\psi(h+\epsilon e_{2,\tau}) - \psi(h)}{\epsilon} \\
&= \int_\eta^\infty \left(\frac{1}{2}(t-\eta)^2 - \frac{(t-\eta)}{h(t)}\mathbb{I}_{t\neq X_{(n)}}\right) d\mathbb{F}_n(t).
\end{aligned}$$

We also define the three directional derivatives for the quadratic approximation.

$$\begin{aligned}
\nabla_0\psi^{mod}(h|g) &= \int_0^\infty \left\{t - \left(\frac{2}{g(t)} - \frac{h(t)}{g^2(t)}\right)\mathbb{I}_{t\neq X_{(n)}}\right\} d\mathbb{F}_n(t), \\
\nabla_1\psi^{mod}(h|g)[\tau] &= \int_0^\infty \left\{t\wedge\tau\left(2\tau - \frac{t\wedge\tau}{2}\right) - 2\frac{(\tau-t)_+}{g(t)} + \frac{(\tau-t)_+h(t)}{g^2(t)}\right\} d\mathbb{F}_n(t), \\
\nabla_2\psi^{mod}(h|g)[\eta] &= \int_\eta^\infty \left\{\frac{1}{2}(t-\eta)^2 - \left(\frac{2(t-\eta)}{g(t)} + \frac{(t-\eta)h(t)}{g^2(t)}\right)\mathbb{I}_{t\neq X_{(n)}}\right\} d\mathbb{F}_n(t).
\end{aligned}$$

Note that both $\nabla_2\psi(h)[\eta] \geq 0$ and $\nabla_2\psi^{mod}(h|g)[\eta] \geq 0$ for $\eta > X_{(n-1)}$. Also, the term $\mathbb{I}_{t\neq X_{(n)}}$ is not part of the derivatives in the direction of the decreasing elbows, as the term has no impact for $\tau \geq X_{(n)}$.

We may now describe the algorithm. Choose an accuracy variable $\varepsilon > 0$, and define

$$\nabla\psi(\tilde{h}) = \min\left\{\nabla_0\psi(\tilde{h}), \min_{\tau\in[0,a]} \nabla_1\psi(\tilde{h})[\tau], \min_{\eta\in[a,T]} \nabla_2\psi(\tilde{h})[\eta]\right\}.$$

SUPPORT REDUCTION ALGORITHM TO FIND THE PROFILE MLE OF A CONVEX HAZARD:

STEP 0. Obtain an initial estimate, $\hat{h}$. For the MLE there are two natural choices: (1) set $\hat{h} = \hat{\alpha} = 1/\bar{X}$, or (2) find the LSE, $\tilde{h}_{n,a}$, for some choice of $T < X_{(n)}$, extend it linearly beyond T, and set $\hat{h} = \tilde{h}_{n,a}$.

WHILE $\nabla\psi(\tilde{h})$ is less than $-\varepsilon$ **REPEAT 1-3:**

STEP 1 (INNER LOOP). Given a current estimate $\hat{h}$, we find the next proposed $\hat{h}_p$ by minimizing the linearized criterion function $\psi^{mod}(h|\hat{h})$. This minimization step is done using the support reduction algorithm. That is, let $\hat{h}_p = \text{argmin}_h \psi^{mod}(h|\hat{h})$.

To find the starting value for the support reduction algorithm, preform the following:

STEP A: Consider the support of $\hat{h}$, and find the function $\hat{h}_0$ with the same support which minimizes the $\psi_n^{mod}(h|\hat{h})$.

WHILE $\hat{h}_0 \notin \mathcal{K}(a)$ **REPEAT B:**

STEP B: Perform a support reduction step to obtain a new $\hat{h}_0$, with a reduced support.

Step 1 yields a new proposed estimate, $\hat{h}_p$. However, since we are minimizing the approximation of the criterion function and not the function itself, there is no guarantee that we have actually improved our estimate. Therefore we perform the next step.

STEP 2 (ARMIJO STEP). Find λ in $[0, 1]$ which minimizes

$$\psi\left((1-\lambda)\hat{h} + \lambda\hat{h}_p\right).$$

STEP 3. The new $\hat{h}$ is set to $(1-\lambda^*)\hat{h} + \lambda^*\hat{h}_p$.

3.1.1. *Practical Considerations.* The same observations apply to the MLE as to the LSE. Additional issues are as follows.

Armijo step. In practice, this is implemented using, again, a gridded approach for λ. In practice we have found that for all outer iteration other than the first, $\lambda^* = 1$. To speed up the calculations, we only use the gridded approach on iteration one, and use the faster

$$\lambda = 1$$
$$\texttt{while } \psi\left((1-\lambda)\hat{h} + \lambda\hat{h}_p\right) - \psi\left(\hat{h}\right) \geq 0 \texttt{ set } \lambda = 0.9\lambda$$

otherwise.

Starting point. Because of the inner loop in this approach, the choice of starting point is more important here. Setting the initial value of $\hat{h}$ in the algorithm to an

estimate obtained by using the LSE can be used to reduce the computing time for the MLE. Naturally, this also requires computing the least squares estimator.

Vector Calculations. Most of the computation required for the algorithm may be re-written in terms of vector operations, which greatly speeds up the algorithm. This is especially important when calculating the directional derivatives in the inner loop.

For example, the vector `DD2` $= \{\nabla_2 \psi^{mod}(h|g)[\eta]\}_{\eta \in \texttt{GRID}}$ may be found as follows. We denote `GRID` $=\{\eta_j\}_{j=1,\dots,m}$, and let $\{x_{(i)}\}_{i=1,\dots,n}$ denote the *ordered* data. Note that below we use `*` and `/` to denote *component-wise* multiplication and division of matrices. Matrix multiplication is denoted by `%*%`.

`u.vec` $=\frac{1}{n}\{1\}_{j=1,\dots,n}$
`u.vec.r` $=\frac{1}{n}\{\mathbb{I}_{j\neq n}\}_{j=1,\dots,n}$
`h.mat` $=\{h(x_{(i)})\}_{i=1,\dots,n;j=1,\dots,m}$
`g.mat` $=\{g(x_{(i)})\}_{i=1,\dots,n;j=1,\dots,m}$
`e2.mat` $=\{e_{2,\eta_j}(x_{(i)})\}_{i=1,\dots,n;j=1,\dots,m}$
`E2.mat` $=\{\int_0^{x_{(i)}} e_{2,\eta_j}(s)ds\}_{i=1,\dots,n;j=1,\dots,m} = \frac{1}{2}\{e_{2,\eta_j}(x_{(i)})^2\}_{i=1,\dots,n;j=1,\dots,m}$

```
DD2.mat = - 2*e2.mat/g.mat + e2.mat*h.mat/(g.mat*g.mat)
DD2 = u.vec %*% E2.mat + u.vec.r %*% DD2.mat
```

A similar implementation may be used in all calculations.

Nearly singular matrices. As for the LSE, in the inner loop we need to minimize a quadratic function in finitely many variables. This is easily done using a built-in function designed to solve systems of linear equations: `solve()` in R or `LinearSolve[]` in Mathematica, for example. Unfortunately, for the MLE algorithm, 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 adhoc solution works reasonably well in practice. Another solution would be to change the starting point of the algorithm, but this is much slower.

The finer the grid in the gridded implementation, the more often this problem arises. We therefore recommend not setting too fine a grid for the MLE. For this reason also, we do not recommend implementing a *gridless* version of the MLE.

3.2. **Some theoretical results.** We again collect all technical results in a separate section. All of the results are analogues of the results of Section 2.3, and we will therefore provide only minimal explanations. Note that there is no need to rework Proposition 2.7, as the proof does not depend on the particular shape of the criterion function. Similarly, the proof of Lemma 3.1 proceeds along the same lines as the proof of Proposition 3.4 in [JW] and will therefore be omitted.

Lemma 3.1. *There exists a unique minimizer, $\hat{h}_{n,a}$, of the function $\psi(h)$ over $\mathcal{K}_+(a)$. Moreover, the minimizer has support of at most size $n+1$, where n is the sample size.*

It follows that the minimizer has finite total measure. Also, the minimizer is always strictly positive on all observation points. That is,

$$\hat{h}_{n,a}(X_i) > 0 \ \forall i = 1, \ldots, n.$$

Lemma 3.2. *A function $\hat{h}_n$ minimizes ψ over $\mathcal{K}$ (and hence is the MLE) if and only if:*

$$\int_0^x \frac{x-t}{\hat{h}_n(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \quad \leq \quad \frac{x^2}{2} - \int_0^x \frac{(x-t)^2}{2} d\mathbb{F}_n(t) = \int_0^x \int_0^t \mathbb{S}_n(s) ds dt, \quad (3.1)$$

for all $x \geq 0$ with equality at τ_i for $i = 1, \ldots, k$;

$$\int_x^\infty \frac{t-x}{\hat{h}_n(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \quad \leq \quad \int_x^\infty \frac{(t-x)^2}{2} d\mathbb{F}_n(t) = \int_x^\infty \int_t^\infty \mathbb{S}_n(s) ds dt, \quad (3.2)$$

for all $x \geq 0$ with equality at η_j for $j = 1, \ldots, m$;

$$\int_0^\infty \frac{1}{\hat{h}_n(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \quad \leq \quad \int_0^\infty t d\mathbb{F}_n(t) = \int_0^\infty \mathbb{S}_n(t) dt, \quad (3.3)$$

$$\int_0^\infty \hat{H}_n(t) d\mathbb{F}_n(t) \quad = \quad 1 - 1/n. \quad (3.4)$$

A proof appears in [JW]. Using a similar approach we may also prove a characterization for the constrained minimum.

Lemma 3.3. *The function $\hat{h}_{n,a}$ minimizes $\psi(h)$ over $\mathcal{K}(a)$ if and only if it satisfies*

$$\int_0^\infty \frac{1}{\hat{h}_{n,a}(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \ \leq \ \int_0^\infty \mathbb{S}_n(t) dt, \quad (3.5)$$

$$\int_0^x \frac{x-t}{\hat{h}_{n,a}(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \ \leq \ \int_0^x \int_0^t \mathbb{S}_n(s) ds dt \quad (3.6)$$

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

$$\int_x^\infty \frac{t-x}{\hat{h}_{n,a}(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \ \leq \ \int_x^\infty \int_t^\infty \mathbb{S}_n(s) ds dt, \quad (3.7)$$

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

$$\int_0^\infty \hat{H}_{n,a}(t) d\mathbb{F}_n(t) = 1 - 1/n. \quad (3.8)$$

Proposition 3.4. *The criterion function for the MLE*

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

satisfies the conditions

A1. *ψ is convex on $\mathcal{K}(a)$ and $\psi(h + t(g - h))$ is continuously differentiable as a function of t, for $t \in (0, 1)$.*

A2. *The directional derivative* $\nabla\psi(h)[g]$ *is linear in* g.

A3. *For any specific function* $h_0 \in \mathcal{K}(a)$ *with* $\psi(h_0) < \infty$, *there exists an* $\bar{\varepsilon} \in (0,1]$ *such that for all* $h \in \mathcal{K}_+(a)$ *with* $\psi(h) < \psi(h_0)$ *and any basis function* e, *the following implication holds*

$$\nabla\psi(h)[e-h] \le -\delta \quad \Rightarrow \quad \psi(h+\varepsilon(e-h)) - \psi(h) \le -\frac{1}{2}\varepsilon\delta \quad \text{for all } \varepsilon \in (0,\bar{\varepsilon}].$$

It follows that the support reduction algorithm converges to the true minimizer of ψ *over* $\mathcal{K}(a)$.

Let h_k *denote a sequence generated by the support reduction algorithm. Also, suppose that in each iteration the new support point, corresponding to a basis function* e, *is chosen so that*

$$\nabla\psi(h_k)[e] \le \frac{1}{2}\inf_{e' : basis} \nabla\psi(h_k)[e'].$$

Then, $\psi(h_k) \to \psi(\hat{h}_n)$.

Note that unlike the least squares case, the algorithm for the MLE is a further approximation to the support reduction algorithm. For this reason, we refrain from the error calculations we presented there.

Proposition 3.5. *Let* $a_0 \in [0,T]$ *be such that the minimizer of* $\psi(h)$ *over* h *in* $\mathcal{K}_+$ *has an antimode at* a_0. *Suppose that* $a_0 < a_1 < a_2$, *then*

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

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

4. Examples and Simulations

4.1. **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 [HS] 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 H-S distribution after the authors. Notably, the distribution has *convex* hazards for all values of b in the parameter space, $b > -1/2$. In Figure 3, we present an example of the LSE and MLE for a simulation from this distribution with a sample size of 100. For the LSE estimator we set T to be 0.8. Notice that both the MLE and LSE blow up at zero and at the ends, $T, X_{(n)}$ (this occurs by definition for the convex MLE, since $\widehat{h}_n(x) = \infty$ for $x \ge X_{(n)}$). This behavior is to typical of shape-constrained nonparametric estimators, see for example

Remark 4.5 in [JW]. We also compare our convex estimators to the U-shaped MLE [BCP]. The U-shaped MLE also lacks consistency at $0, X_{(n)}$ (in fact, as in the convex case, it is arbitrarily large for all $x \geq X_{(n)}$).

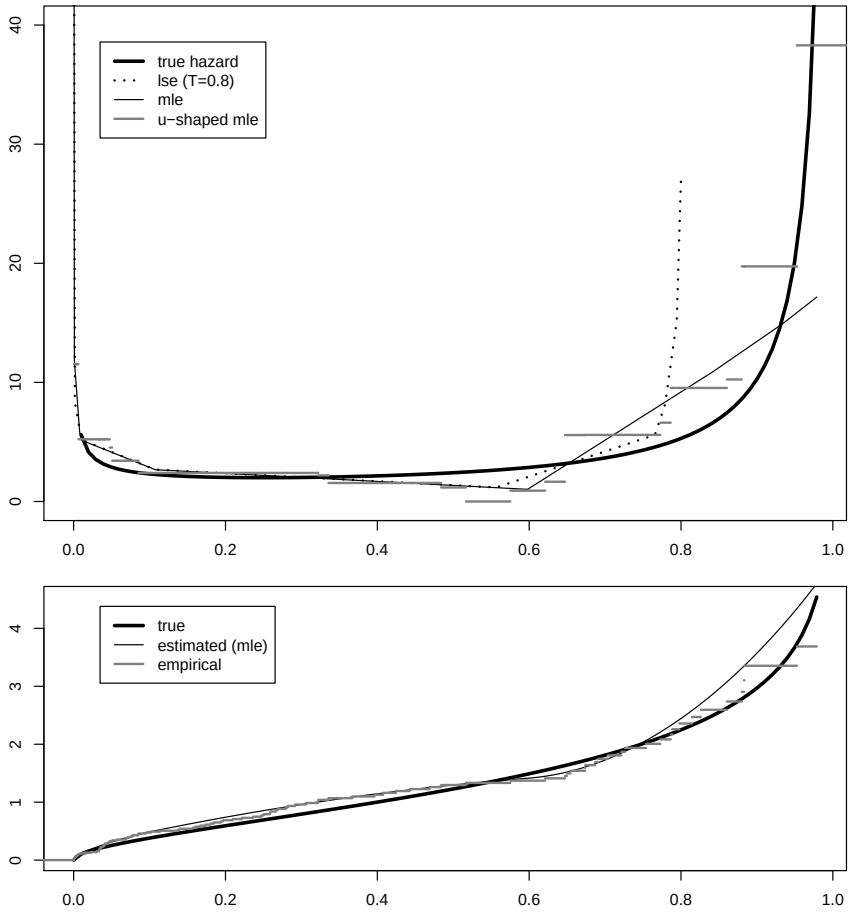

FIGURE 3. Estimation of the H-S hazard with $b = 0, A = 1$ for a sample size of 100: LSE with T set to 0.8, convex MLE and U-shaped MLE (top); cumulative hazard of the convex MLE compared to the empirical $\mathbb{H}_n$ and true function (bottom).

Note that both the U-shaped and convex maximum likelihood estimators appear to be following a similar trend, except that one is continuous and the other a step function. Figure 3 also looks at the cumulative hazards: of the true distribution, of the estimated convex MLE, and of the data, $\mathbb{H}_n$. Notice that the estimated function

follows the empirical one quite closely. This will be true in general: the estimator can only be as good as the data. This trend continues, although to a lesser degree, even if we set the antimode to an "incorrect" value, see Figure 4.

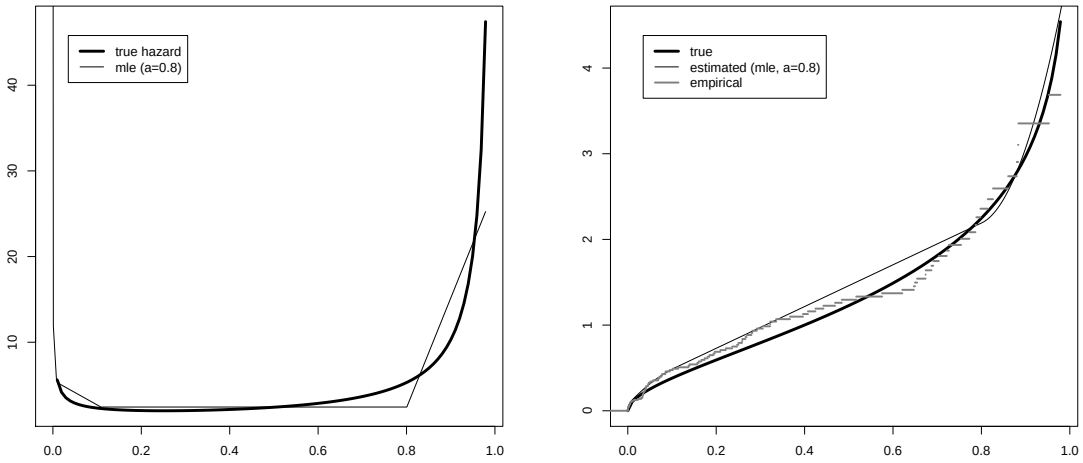

FIGURE 4. Constrained maximum likelihood estimator of the H-S hazard with antimode set at $a = 0.8$: true and estimated hazard functions (left); true, estimated and empirical cumulative hazard functions (right).

For this example we also examine the effects of the parameter settings in the algorithm: grid size and the gridless implementation. The R package, convexHaz was used to find the MLE by calling on the function convexMLE(x, M), where M indicates the size of the grid used. For the LSE we look at the function convexLSE(x, M, GRIDLESS), modifying M, and setting GRIDLESS to 1 if a gridless implementation is desired. The results are shown in Figure 5. We note that the bathtub shape of the negative of the logarithm of the profile likelihood is preserved by the different implementations; this is also true for the LSE.

4.2. **Testing a conjecture.** In [JW] we show that both the maximum likelihood and least squares estimators converge *locally* at a rate of $n^{2/5}$. Indeed, the following stronger result holds.

Theorem 4.1. *Suppose that h_0 is convex and $x_0 > 0$ is a point which satisfies $h_0(x_0) > 0$, $h_0''(x_0) > 0$, and that $h_0''(\cdot)$ is continuous in a neighborhood of x_0 (also, $x_0 < T$ for the LSE). Then for $\bar{h}_n = \widehat{h}_n$ or $\widetilde{h}_n$,*

$$n^{1/5}(\bar{h}_n'(x_0 + n^{-1/5}t) - h_0'(x_0)) \Rightarrow \widetilde{\mathcal{I}}^{(3)}(t)$$

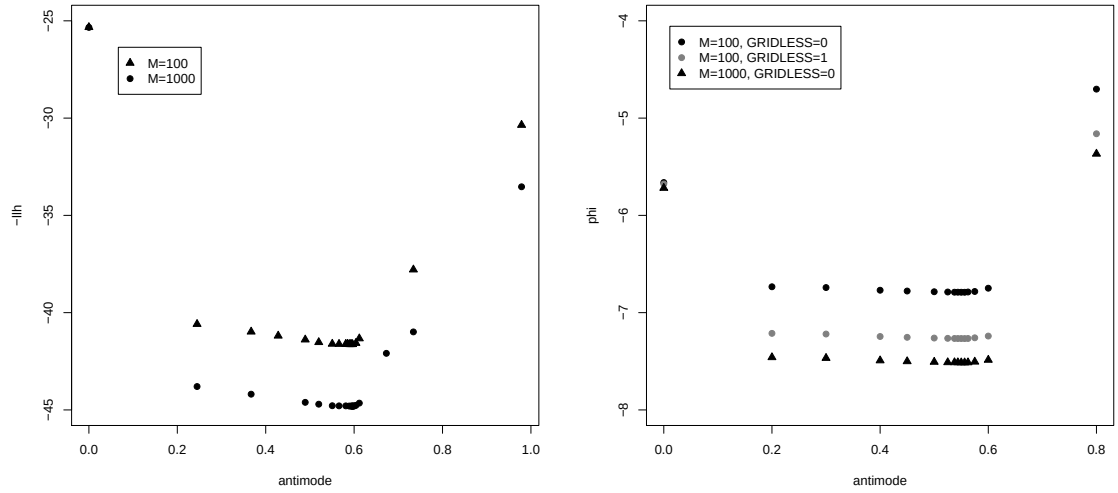

FIGURE 5. Bisection diagnostics for the estimators in Figure 3. For both the MLE (left) and the LSE (right) we can see that the larger the M is, the better the algorithm behaves. For the LSE, note also how the gridless implementation with $M = 100$ achieves results relatively close to those reached by $M = 1000$ but with GRIDLESS set to zero. However, the running time for $M = 100$, GRIDLESS=1 was about $1/6th$ of the running time of $M = 1000$, GRIDLESS=0.

where $\mathcal{I}^{(3)}$ *is the third derivative of the invelope of* $Y(t) \equiv k_1 \int_0^t W(s)ds + k_2 t^4$, *with*

$$k_1 = \sqrt{\frac{h_0(x_0)}{1 - F_0(x_0)}}, \qquad k_2 = \frac{1}{24} h_0''(x_0).$$

The invelope process $\mathcal{I}$ of Y is defined below. It was shown in [GJW2] that it exists and is almost surely uniquely defined. Moreover, with probability one, $\mathcal{I}$ is three times differentiable at $t = 0$.

Definition 4.2. *Let* $W(s)$ *denote a standard two-sided Brownian motion, with* $W(0) = 0$, *and define* $Y(t) = \int_0^t W(s)ds + t^4$. *The function* $\{\mathcal{I}(t) : t \in \mathbb{R}\}$, *the invelope of the process* $\{Y(t) : t \in \mathbb{R}\}$, *is defined as follows:*

- *The function* $\mathcal{I}$ *is above the function* Y: $\mathcal{I}(t) \geq Y(t)$ *for all* $t \in \mathbb{R}$. (4.1)
- *The function* $\mathcal{I}$ *has a convex second derivative.* (4.2)
- *The function* $\mathcal{I}$ *satisfies* $\int_{\mathbb{R}} \{\mathcal{I}(t) - Y(t)\} d\mathcal{I}^{(3)}(t) = 0$. (4.3)

Theorem 4.1 is proved in [JW]. The idea of the proof is to take the characterizations of the MLE (Lemma 3.2) and the LSE (Lemma 2.4), describe appropriate *local* versions of these, and show that as $n \to \infty$ the characterizations become equivalent

to those of the invelope. The result 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. However, the theorem holds only if the second derivative is positive at x_0. 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 6, where we look at the number of support points vs. sample size in the least squares estimator for simulations from the Weibull distribution versus the Exponential distribution. Although the algorithm finds an approximation to the least squares estimator, and hence the number of support points is also approximate, the simulation shows a clear difference in the asymptotic behavior between the two distributions. Similar behavior should be seen for the maximum likelihood estimator. We consider the LSE here, because the algorithm has a faster implementation.

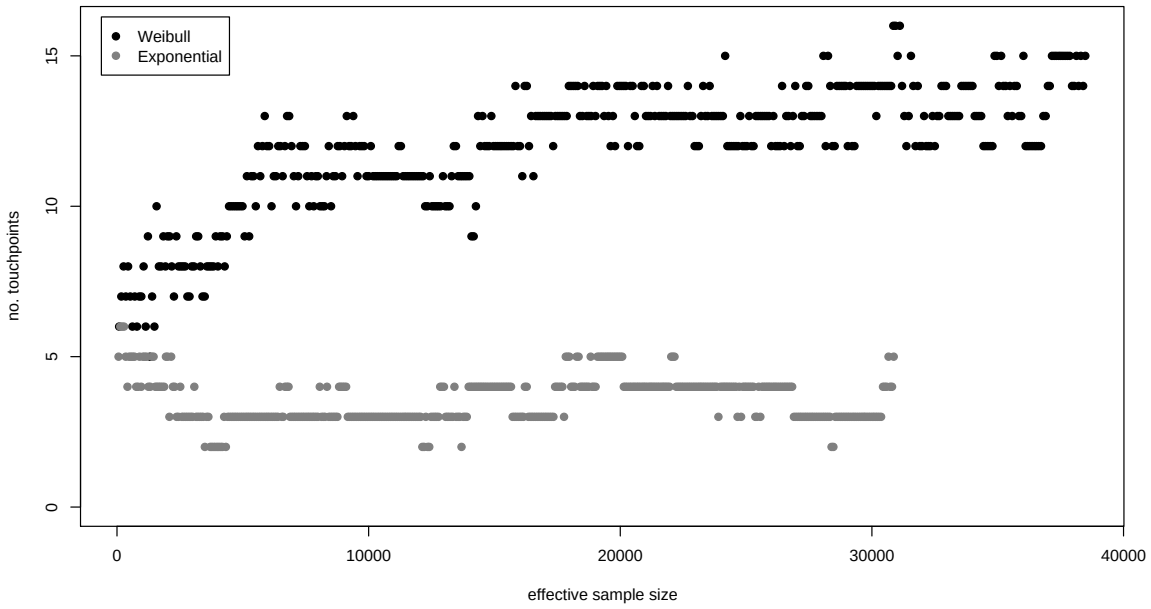

FIGURE 6. Support size as a function of sample size for the LSE: observations for the exponential distribution are grey, and observations for the Weibull distribution with cubic hazard are black. The least squares estimator with $T = 1.2$ was used, and we plot the support size against the effective sample size: the number of data points below 1.2. The data supports the conjecture that the asymptotic behavior for flat or linear hazards is different than that of strictly convex hazards.

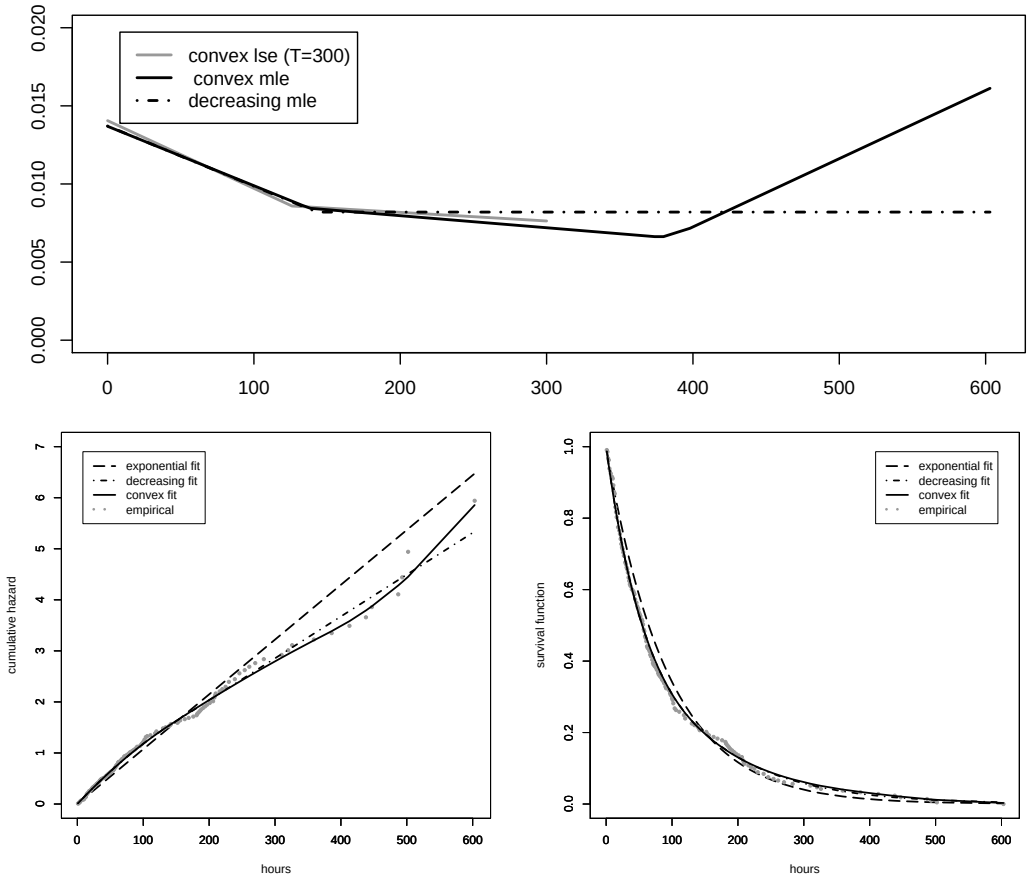

FIGURE 7. Maximum likelihood and least squares estimators for air conditioning data of [Pro]: the MLE (both convex and convex decreasing) and LSE, with $T = 300$ hours (top), and a comparison of the fitted cumulative hazard functions and cumulative distributions with their empirical counterparts (bottom left and bottom right, respectively). Differences in the nonparametric estimators appear at roughly the 300 hour mark: only 12 of the 213 observations are larger than 300, and 6 of 213 are larger than 400.

4.3. **Two examples.** Next we consider the number of operating hours between successive failures of airconditioning equipment in 13 aircraft. A total of 213 times were

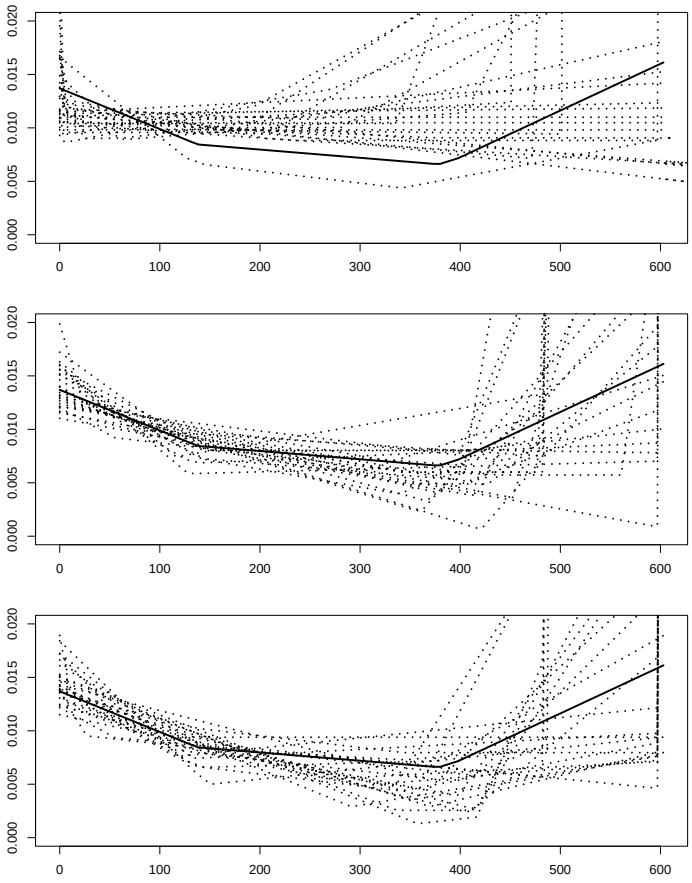

FIGURE 8. All plots show the convex MLE in bold, along with: (top) Plot of 25 convex MLE fits to a sample with exponential distribution (mean equal to that of the air conditioning data sample); (middle) Plot of 25 convex MLE fits to a sample from a distribution which has hazard the same as the convex decreasing MLE; (bottom) Plot of 25 convex MLE fits to a bootstrapped sample from the air conditioning data set. In both the bottom and middle plots, the band of sampled curves appears to be centered at the convex MLE, with the band giving an indication of the variance. It is apparently 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 Figure 7), along with the fact that the difference appears in a region with very few observations. The top plot is designed to test the hypothesis of exponential fit. There is evidence that this is false, as the bold curve lies in an extreme region of the band.

recorded. This data set was studied in [Pro] and again in [CL]. We are interested in the overall hazard rate of the intervals between successive failures.

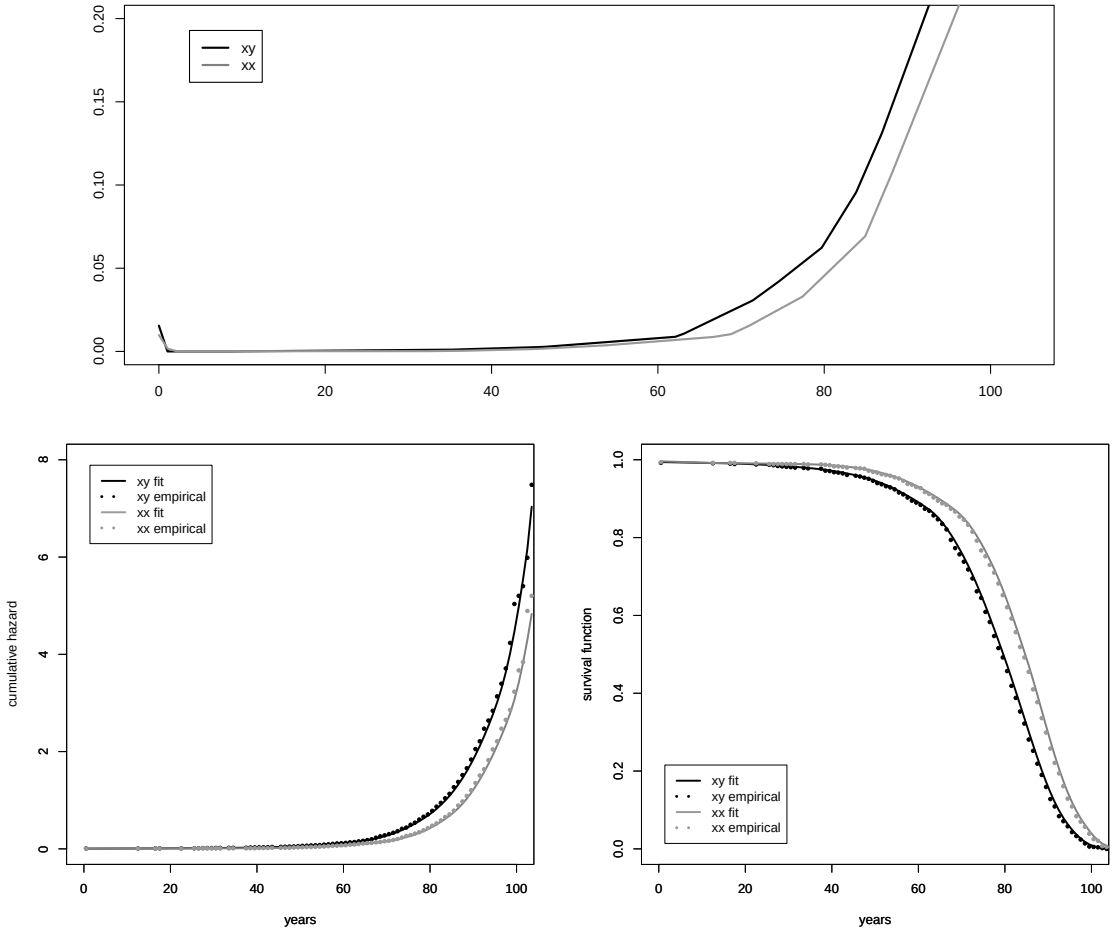

FIGURE 9. Maximum likelihood estimator for Canadian lifetime data: the fitted hazard rates (top), cumulative hazard rates and survival functions (bottom left and right, respectively).

The analysis of [Pro] is summarized as follows. First an exponential fit to the data is considered. Although the null hypothesis of exponential times is not rejected by the Kolmogorov-Smirnov test, the data does 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 towards 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 (Theorem 2, [Pro]). 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 due to [PP]. The resulting test was significant, with a p-value of .007, and hence lead to the conclusion that the pooled distribution has decreasing hazard rate.

[CL] consider fitting time-dependent Poisson processes to the data, and ultimately settle on a mixture of homogeneous processes, in agreement with [Pro].

Figure 7 shows our fit of the nonparametric convex MLE and LSE (with T set to 300 hours for the LSE). The MLE has an antimode at the 375 hour mark, which appears to be in contradiction to the results of [PP]. We investigate this further using resampling methods in Figure 8, and find that there is not 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 7. We note that this estimator uses the full likelihood, (1.1), and not the modified likelihood (1.2).

Lastly, we apply our estimators to a lifetime data set: the Canadian mortality table for the years 2000 to 2002 [Can]. To generate our results, we took a random sample of size $n = 1000$ from the distribution given by the lifetables. We also use a simplified version of the standard actuarial assumption of uniform deaths for fractional ages. That is, we assume that all deaths occurred half-way through the year. The resulting maximum likelihood estimators for both male and female lifetimes are given in Figure 9, fitted cumulative hazards and survival functions are also shown. A parametric approach for this data was considered in [BLZ] (Figure 2a). Specifically, [BLZ] fit a mixture of flexible and reduced additive Weibull survival functions. A comparison of the survival functions is provided in Figure 10.

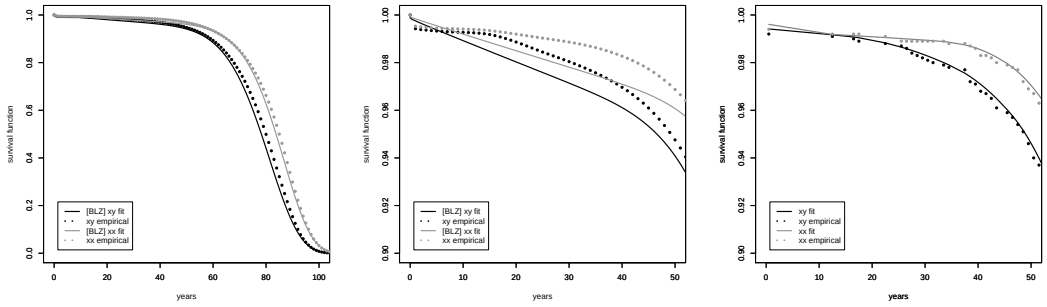

FIGURE 10. Fitted survival functions of [BLZ] (left and middle), and the survival functions from the convex MLE (right). The [BLZ] model was fit directly to the life table survival function, whereas the convex MLE was fit to a sample of size 1000 from this distribution.

References

[Bal] Fadoua Balabdaoui. *Nonparametric Estimation of a k-monotone density: a new asymptotic distribution theory.* Ph.d., University of Washington, 2004.

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

[BLZ] M Bebbington, C.D. Lai, and R Zitikis. Modeling human mortality using mixtures of bathtub shaped failure distributions. *Journal of Theoretical Biology*, 245(3):528–538, 2007.

[Böh1] Dankmar Böhning. Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.*, 10(3):1006–1008, 1982.

[Böh2] Dankmar Böhning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Statist. Plann. Inference*, 47(1-2):5–28, 1995. Statistical modelling (Leuven, 1993).

[Can] 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/caf.txt, http://www.statcan.ca/english/freepub/84-537-XIE/tables/txttables/cam.txt.

[CL] D. R. Cox and P. A. W. Lewis. *The statistical analysis of series of events*. Methuen & Co. Ltd., London, 1966.

[Fed] V. V. Fedorov. *Theory of optimal experiments*. Academic Press, New York, 1972. Translated from the Russian and edited by W. J. Studden and E. M. Klimko, Probability and Mathematical Statistics, No. 12.

[GJW1] Piet Groeneboom, G. Jongbloed, and Jon A. Wellner. The support reduction algorithm for computing nonparametric function estimates in mixture models. *Scand. J. Statist.*, 2008. To appear.

[GJW2] Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. A canonical process for estimation of convex functions: the "invelope" of integrated Brownian motion $+t^4$. *Ann. Statist.*, 29(6):1620–1652, 2001.

[GJW3] Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.*, 29(6):1653–1698, 2001.

[Gre] Ulf Grenander. On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, 39:125–153 (1957), 1956.

[HS] Elvira Haupt and Hendrik Schäbe. The TTT transformation and a new bathtub distribution model. *J. Statist. Plann. Inference*, 60(2):229–240, 1997.

[JW] Hanna Jankowski and Jon A. Wellner. Nonparametric estimation of a convex bathtub-shaped hazard function. Technical Report 521, University of Washington, Department of Statistics, 2007.

[JWMW] H. Jankowski, X. Wang, H. McCauge, and J. Wellner. *convexHaz: R functions for convex hazard rate estimation*, 2008. R package version 0.0.

[LK] Mary L. Lesperance and John D. Kalbfleisch. Mixture models for matched pairs. *Canad. J. Statist.*, 22(1):65–74, 1994.

[PP] Frank Proschan and Ronald Pyke. Tests for monotone failure rate. In *Proc. Fifth Berkeley Sympos. Mathematical Statistics and Probability (Berkeley, Calif., 1965/66), Vol. III: Physical Sciences*, pages 293–312. Univ. California Press, Berkeley, Calif., 1967.

[PR] B. L. S. Prakasa Rao. Estimation of a unimodal density. *Sankhyā Ser. A*, 31:23–36, 1969.

[Pro] F. Proschan. Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5(3):375–383, 1063.

[RWD] Tim Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1988.

[Sim] Léopold Simar. Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.*, 4(6):1200–1209, 1976.

[Wyn] Henry P. Wynn. The sequential generation of D-optimum experimental designs. *Ann. Math. Statist.*, 41:1655–1664, 1970.

Appendix

Sketch of proof of Lemma 2.2. The key idea is to show that the minimizer of $\varphi(h)$ over $\mathcal{K}_T(a)$ must lie in the compact set

$$\{h : h \in \mathcal{K}_T(a), 0 \leq h \leq B\}$$

for some constant B. The proof of this is similar to that of Proposition 3.1 in [JW], and we therefore omit the details. Since the function φ is strictly convex on $\mathcal{K}_T(a)$ it now follows that there exists a unique minimizer.

Next, again as in Proposition 3.1 in [JW], we may show that the minimizer has at most $n+1$ changes of slope. Since each change of slope corresponds to an element of the support, the result follows. □

Proof of Lemma 2.5. For two functions h, g, we calculate

$$\begin{aligned}
\varphi(g) - \varphi(h) &= \frac{1}{2}\int_0^T (g-h)^2 dt + \int_0^T (g-h) d[H - \mathbb{H}_n](t) \\
&\geq \int_0^T (g-h) d[H - \mathbb{H}_n](t) \\
&= (g-h)(T)[H - \mathbb{H}_n](T) - (g-h)'(T)[\mathcal{H} - \mathbb{Y}_n](T) \\
&\quad + \int_0^T [\mathcal{H} - \mathbb{Y}_n](t) d(g-h)', \qquad \text{(A-1)}
\end{aligned}$$

by integration by parts. To show that the characterization is necessary, notice that by the above, we have that

$$\begin{aligned}
\nabla\varphi(h)[\gamma] &\equiv \lim_{\varepsilon \to 0} \frac{\varphi(h + \varepsilon\gamma) - \varphi(h)}{\varepsilon} \\
&= \gamma(T)[H - \mathbb{H}_n](T) - \gamma'(T)[\mathcal{H} - \mathbb{Y}_n](T) \\
&\quad + \int_0^T [\mathcal{H} - \mathbb{Y}_n](t) d\gamma',
\end{aligned}$$

for any functions h and γ.

Let $\tilde{h}_{n,a}$ denote the true minimizer of $\mathcal{K}_T(a)$. Now, if γ is such that $\tilde{h}_{n,a} + \varepsilon\gamma$ is in $\mathcal{K}_T(a)$ (for sufficiently small ε), then we have that $\nabla\varphi(\tilde{h}_{n,a})[\gamma] \geq 0$ (since $\varphi(\tilde{h}_{n,a} + \varepsilon\gamma) \geq \varphi(\tilde{h}_{n,a})$). This is true for the choice of γ equal to 1, $(\tau - t)_+$ for $\tau \in [0, a]$, and for $(t - \eta)_+$ for $\eta \in [a, T]$. These choices of γ yield the inequalities in conditions (2.6),

(2.7) and (2.8). Next, if γ is such that $\tilde{h}_{n,a} \pm \varepsilon\gamma$ is in $\mathcal{K}_T(a)$ (again, for sufficiently small ε), then it follows that $\nabla\varphi(\tilde{h}_{n,a})[\gamma] = 0$. Choosing $\gamma = (\tau_i - t)_+, (t - \eta_j)_+$ for $i = 1, \ldots, k$ and $j = 1, \ldots, m$, as well as $\gamma = \tilde{h}_{n,a}$ gives the remaining equalities.

Next, using (A-1), we find that for any h and $\tilde{h}_{n,a} \in \mathcal{K}(a)$, with $\tilde{h}_{n,a}$ satisfying the conditions of the lemma, $\varphi(h) - \varphi(\tilde{h}_{n,a})$ is bounded below by

$$\begin{aligned}
&(h - \tilde{h}_{n,a})(T)[\tilde{H}_{n,a} - \mathbb{H}_n](T) - (h - \tilde{h}_{n,a})'(T)[\tilde{\mathcal{H}}_{n,a} - \mathbb{Y}_n](T) \\
&\qquad + \int_0^T [\tilde{\mathcal{H}}_{n,a} - \mathbb{Y}_n](t) d(h - \tilde{h}_{n,a})' \\
&\overset{(2.9)}{=} h(T)[\tilde{H}_{n,a} - \mathbb{H}_n](T) - h'(T)[\tilde{\mathcal{H}}_{n,a} - \mathbb{Y}_n](T) + \int_0^T [\tilde{\mathcal{H}}_{n,a} - \mathbb{Y}_n](t) dh'.
\end{aligned}$$

Therefore

$$\begin{aligned}
\varphi(h) - \varphi(\tilde{h}_{n,a}) \;\geq\; & \alpha[\tilde{H}_{n,a} - \mathbb{H}_n](T) + \int_0^a [\tilde{\mathcal{H}}_{n,a} - \mathbb{Y}_n](\tau) d\nu(\tau) \\
& + \int_a^T \left\{ (T - \eta)[\tilde{H}_{n,a} - \mathbb{H}_n](T) - \int_\eta^T [\tilde{H}_{n,a} - \mathbb{H}_n](s) ds \right\} d\mu(\eta),
\end{aligned} \tag{A-2}$$

using that any function in $\mathcal{K}_T(a)$ may be decomposed as

$$h(t) \;=\; 1 \cdot \alpha + \int_0^a (\tau - t)_+ d\nu(\tau) + \int_a^T (t - \eta)_+ d\mu(\eta).$$

Lastly, (A-2) is nonnegative by the inequalities in (2.6)-(2.8). Thus the conditions are sufficient. □

Proof of Proposition 2.6. A simple calculation shows that for any h and g

$$\varphi(h + \varepsilon g) - \varphi(h) \;=\; \frac{1}{2}\varepsilon^2 \int_0^T g^2 dt + \varepsilon \int_0^T g d(H - \mathbb{H}_n) \tag{A-3}$$

$$\;=\; \frac{1}{2}\varepsilon^2 \int_0^T g^2 dt + \varepsilon \nabla\varphi(h)[g] \tag{A-4}$$

and the first two conditions follow easily. Since the LSE has finite support measure, we may assume that all of the functions h have total measure $(\alpha + \nu[0, a] + \mu[a, T])$ bounded by some fixed, large, number, S. We then have that

$$\sup_{e:\text{basis}} \int_0^T e^2 dt \leq \max\left\{ T, \frac{T^3}{3} \right\} \equiv M,$$

and it follows that

$$\int_0^T h^2 dt \;\leq\; 2(S^2 + 1)M,$$

by applying Jensen's inequality. Using the result of the above display in A-3, we obtain that

$$\varphi(h+\varepsilon(e-h))-\varphi(h) \;\leq\; \frac{1}{2}2(S^2+1)M\varepsilon^2-\delta\varepsilon,$$

if $\nabla\varphi(h)[g]\leq-\delta$. Therefore, if $\bar{\varepsilon}$ is chosen sufficiently small, this is bounded above by $-\delta\varepsilon/2$ as required.

The last part of the proposition follows from Theorem 1 in [GJW1]. □

Proof of Proposition 2.7. We provide the details only in the first case in the statement of the proposition. Suppose that the current estimate of $\tilde{h}$ is given by

$$\tilde{h}=\tilde{\alpha}\cdot e_0+\sum_{i=1}^{k}\tilde{\nu}_i e_{1,\tau_i}+\sum_{j=1}^{m}\tilde{\mu}_j e_{2,\eta_j}.$$

We will assume that $\tilde{\alpha}>0$ to simplify the exposition. Note that $\tilde{h}$ must be the minimizer of φ over the class of hazard functions with support given by $\widetilde{supp}=\{1\}\times\{\tau_1,\ldots,\tau_k\}\times\{\eta_1,\ldots,\eta_m\}$. It therefore follows that

$$\nabla\varphi(\tilde{h})[\gamma]=0,$$

for $\gamma=e_0,e_{1,\tau_1},\ldots,e_{1,\tau_k},e_{2,\eta_1},\ldots,e_{2,\tau_m}$.

The new mixing measure $\widetilde{mix}^*$ is obtained by minimizing φ over the class of hazard functions with support given by $\widetilde{supp}^*=\{1\}\cup\{\tau_1,\ldots,\tau_k,\tau_{k+1}\}\cup\{\eta_1,\ldots,\eta_m\}$, which is a larger class of functions than $\widetilde{supp}$. Therefore,

$$\varphi(\tilde{h}^*)<\varphi(\tilde{h}).$$

By the convexity of φ, for any $\epsilon>0$,

$$\begin{aligned}\varphi(\epsilon\tilde{h}^*+(1-\epsilon)\tilde{h})-\varphi(\tilde{h}) &\leq \epsilon\varphi(\tilde{h}^*)+(1-\epsilon)\varphi(\tilde{h})-\varphi(\tilde{h})\\ &= \epsilon(\varphi(\tilde{h}^*)-\varphi(\tilde{h}))<0.\end{aligned}$$

Then

$$\begin{aligned}\nabla\varphi(\tilde{h})[\tilde{h}^*-\tilde{h}] &= \lim_{\varepsilon\to0}\frac{\varphi(\epsilon\tilde{h}^*+(1-\epsilon)\tilde{h})-\varphi(\tilde{h})}{\varepsilon}\\ &= [\tilde{\alpha}^*-\tilde{\alpha}]\nabla\varphi(h)[e_0]\\ &\quad+\sum_{i=1}^{k}[\tilde{\nu}_i^*-\tilde{\nu}_i]\nabla\varphi(\tilde{h})[e_{1,\tau_i}]+\tilde{\nu}_{k+1}^*\,\nabla\varphi(h)[e_{1,\tau_{k+1}}]\\ &\quad+\sum_{j=1}^{m}[\tilde{\mu}_j^*-\tilde{\mu}_j]\nabla\varphi(\tilde{h})[e_{2,\eta_j}]\\ &= \tilde{\nu}_{k+1}^*\,\nabla\varphi(\tilde{h})[e_{1,\tau_{k+1}}]<0.\end{aligned}$$

Since $\nabla\varphi(\tilde{h})[e_{1,\tau_{k+1}}]<0$, we conclude that $\tilde{\nu}_{k+1}^*>0$. □

Proof of Proposition 2.8. Since $K_T(a_1) \subset K_T$, the first inequality is clearly true. It remains to prove the second.

Invoking Lemma 2.2, let h_i denote the unique minimizer of $\varphi(h)$ over $\mathcal{K}_T(a_i)$ for $i = 0, 1, 2$. Then, as in (A-2), we obtain that

$$\begin{aligned}
\varphi(h_2) - \varphi(h_1) &\geq (h_2 - h_1)(T)[H_1 - \mathbb{H}_n](T) - (h_2 - h_1)'(T)[\mathcal{H}_1 - \mathbb{Y}_n](T) \\
&\quad + \int_0^T [\mathcal{H}_1 - \mathbb{Y}_n](t) d(h_2 - h_1)' \\
&\overset{(2.9)}{=} h_2(T)[H_1 - \mathbb{H}_n](T) - h_2'(T)[\mathcal{H}_1 - \mathbb{Y}_n](T) + \int_0^T [\mathcal{H}_1 - \mathbb{Y}_n](t) dh_2' \\
&= \alpha_2 [H_1 - \mathbb{H}_n](T) + \int_0^{a_2} [\mathcal{H}_1 - \mathbb{Y}_n](\tau) d\nu_2(\tau) \\
&\quad + \int_{a_2}^T \left\{ (T - \eta)[H_1 - \mathbb{H}_n](T) - \int_\eta^T [H_1 - \mathbb{H}_n](s) ds \right\} d\mu_2(\eta),
\end{aligned}$$

by writing

$$h_2(t) = \alpha_2 \cdot 1 + \int_{[0,a_2]} (\tau - t)_+ d\nu_2(\tau) + \int_{[a_2,T]} (t - \eta)_+ d\mu_2(\eta).$$

Since h_1 must satisfy (2.6)-(2.8), we obtain that

$$\begin{aligned}
\varphi(h_2) - \varphi(h_1) &\geq \alpha_2 [H_1 - \mathbb{H}_n](T) + \int_0^{a_2} [\mathcal{H}_1 - \mathbb{Y}_n](\tau) d\nu_2(\tau) \\
&\quad + \int_{a_2}^T \left\{ (T - \eta)[H_1 - \mathbb{H}_n](T) - \int_\eta^T [H_1 - \mathbb{H}_n](s) ds \right\} d\mu_2(\eta), \\
&\overset{(2.6)}{\geq} \int_0^{a_2} [\mathcal{H}_1 - \mathbb{Y}_n](\tau) d\nu_2(\tau) \\
&\quad + \int_{a_2}^T \left\{ (T - \eta)[H_1 - \mathbb{H}_n](T) - \int_\eta^T [H_1 - \mathbb{H}_n](s) ds \right\} d\mu_2(\eta), \\
&\overset{(2.8)}{\geq} \int_0^{a_2} [\mathcal{H}_1 - \mathbb{Y}_n](\tau) d\nu_2(\tau) \\
&\overset{(2.7)}{\geq} \int_{a_1}^{a_2} [\mathcal{H}_1 - \mathbb{Y}_n](\tau) d\nu_2(\tau).
\end{aligned}$$

Therefore, if we could show that $[\mathcal{H}_1 - \mathbb{Y}_n](\tau) \geq 0$ for all $\tau \in [a_1, a_2]$, then we would be done. This is what we shall prove next. Using a similar argument as above we

show that

$$\begin{aligned} 0 &\geq \varphi(h_0) - \varphi(h_1) \\ &= \frac{1}{2}\int_0^T (h_0 - h_1)^2 dt + \int_0^T (h_0 - h_1) d(H_1 - \mathbb{H}_n) \\ &\geq \int_0^T h_0 d(H_1 - \mathbb{H}_n), \end{aligned}$$

by using (2.9). Integration by parts as above, plus (2.6)-(2.8) show that

$$0 \geq \int_{a_0}^{a_1} \{(T-\eta)(H_1 - \mathbb{H}_n)(T) - (\mathcal{H}_1 - \mathbb{Y}_n)(T)\}\, d\mu_0(\eta).$$

Now, if $\mu_0 \equiv 0$ on $[a_0, a_1]$ then h_0 has an antimode at a_1 and 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 $t \in [a_0, a_1]$ such that

$$(T-t)(H_1 - \mathbb{H}_n)(T) - (\mathcal{H}_1 - \mathbb{Y}_n)(T) \leq 0.$$

By (2.6), the function on the left-hand side of the display is decreasing in t, and hence the inequality must also hold for all $t \geq a_1$. Now, by (2.8), we have that for all $\eta \geq a_1$,

$$(T-\eta)(H_1 - \mathbb{H}_n)(T) - (\mathcal{H}_1 - \mathbb{Y}_n)(T) + (\mathcal{H}_1 - \mathbb{Y}_n)(\eta) \geq 0.$$

Since, the first two terms on the left-hand side are negative, it follows that the second term must be positive. That is, we have shown that $(\mathcal{H}_1 - \mathbb{Y}_n)(\eta) \geq 0$ for all $\eta \in [a_1, T]$ as desired.

A similar argument proves the inequality $\min_{h \in \mathcal{K}_T(a_1)} \varphi(h) \leq \min_{h \in \mathcal{K}_T(a_2)} \varphi(h)$ for $a_2 > a_1 > a_0$. □

Proof of Lemma 3.3. For any two functions h and g in $\mathcal{K}(a)$ we calculate

$$\psi(h) - \psi(g) \geq \int_0^\infty \left\{ H(t) - G(t) + \left(1 - \frac{h(t)}{g(t)}\right) \mathbb{I}_{t \neq X_{(n)}} \right\} d\mathbb{F}_n(t)$$

since $-\log x \geq 1 - x$. Now, setting $g = \hat{h}_{n,a}$, and using the characterization of elements of $\mathcal{K}(a)$, (1.4), on h, we obtain that the right hand side is equal to

$$\begin{aligned} a\left\{ \int_0^\infty \left(t - \frac{1}{\hat{h}_{n,a}(t)} \mathbb{I}_{t \neq X_{(n)}} \right) d\mathbb{F}_n(t) \right\} + \left\{ 1 - \frac{1}{n} - \int_0^\infty \hat{H}_{n,a}(t) d\mathbb{F}_n(t) \right\} \\ + \int_0^\infty \left\{ \int_0^x \int_0^t \mathbb{S}_n(s) ds dt - \int_0^x \frac{x-t}{\hat{h}_{n,a}(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \right\} d\nu(x) \\ + \int_0^\infty \left\{ \int_x^\infty \int_t^\infty \mathbb{S}_n(s) ds dt - \int_0^x \frac{t-x}{\hat{h}_{n,a}(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \right\} d\mu(x). \end{aligned}$$

This is nonnegative since $\hat{h}_{n,a}$ is a function which satisfies conditions (3.5)-(3.8). It follows that these conditions are sufficient to describe a minimizer of ψ.

We next show that the conditions are necessary. To do this, recall the directional derivative

$$\nabla\psi(h)[\gamma] \equiv \lim_{\epsilon\to 0} \frac{\psi(h+\epsilon\gamma)-\psi(h)}{\epsilon} = \int_0^\infty \left\{\Gamma(t) - \frac{\gamma(t)}{h(t)}\mathbb{I}_{t\neq X_{(n)}}\right\} d\mathbb{F}_n(t). \qquad \text{(A-5)}$$

If $\hat{h}_{n,a}$ minimizes ψ, then for any γ such that $\hat{h}_{n,a}+\epsilon\gamma$ is in $\mathcal{K}(a)$ for sufficiently small ϵ it must be that $\nabla\psi(\hat{h}_{n,a})[\gamma] \geq 0$. If, however, $\hat{h}_{n,a} \pm \epsilon\gamma$ is in $\mathcal{K}$ for sufficiently small ϵ then, $\nabla\psi(\hat{h}_{n,a})[\gamma] = 0$. Choosing, respectively, $\gamma(t) \equiv 1, (t-y)_+$ for $y \in [0,a]$, and $(y-t)_+$ for $y \in [a, X_{(n)}]$ then $\hat{h}_{n,a}+\epsilon\gamma$ is in $\mathcal{K}(a)$, and we obtain the inequalities in conditions (3.5)-(3.7). Since $(1\pm\epsilon)\hat{h}_{n,a}$ is also in $\mathcal{K}(a)$, for sufficiently small ϵ, we obtain (3.8). Choosing, $\gamma = (\tau_i - t)_+, (t-\eta_j)_+$, yields the equalities in (3.5) and (3.6), since each of these functions $\hat{h}_{n,a} \pm \epsilon\gamma$ is in $\mathcal{K}(a)$. Thus the conditions are necessary. □

Proof of Proposition 3.4. The first two properties are immediate. By Lemma 3.1, we may restrict our search to all functions with total support measure $||\hat{\mu}|| \leq S < \infty$ and bounded below by some positive value $b > 0$. By (3.5), we may also assume that $\int_{[0,X_{(n)})} \frac{1}{h} d\mathbb{F}_n(t) < B < \infty$.

$$\begin{aligned}
\psi(h+\varepsilon g)-\psi(h) &= \int_0^\infty -\left\{\log\left(1+\varepsilon\frac{g}{h}\right) - \varepsilon\frac{g}{h}\right\}\mathbb{I}_{t\neq X_{(n)}} d\mathbb{F}_n \\
&\qquad + \varepsilon\int_0^\infty \left\{G - \frac{g}{h}\mathbb{I}_{t\neq X_{(n)}}\right\} d\mathbb{F}_n \\
&\leq \frac{\varepsilon^2}{2}\int_0^\infty \frac{g^2}{h^2}\mathbb{I}_{t\neq X_{(n)}} d\mathbb{F}_n + \varepsilon\nabla\psi(h)[g].
\end{aligned}$$

Therefore,

$$\psi(h+\varepsilon(e-h))-\psi(h) \leq \frac{\varepsilon^2}{2}\, 2(S^2+1)\, \max\{X_{(n)},1\}^2 \frac{B}{b} + \varepsilon\nabla\psi(h)[g].$$

Thus, if $\nabla\psi(h)[g] \leq -\delta$, then the last display is bounded above by $-\varepsilon\delta/2$ for all $\varepsilon \leq \bar{\varepsilon}$, as long as $\bar{\varepsilon}$ is chosen sufficiently small. The full result follows by applying Theorem 1 of [GJW1]. □

Proof of Proposition 3.5. As for the LSE, we need only prove the inequality on the right. A simple calculation shows that for any h, g we have

$$\psi(h)-\psi(g) \geq \int_0^\infty \left\{H(t) - G(t) + \left(1 - \frac{h(t)}{g(t)}\right)\mathbb{I}_{t\neq X_{(n)}}\right\} d\mathbb{F}_n(t).$$

Let $h_i = \operatorname{argmin}_{h\in\mathcal{K}_+(a_i)}\psi(h)$. Then we have that

$$\begin{aligned}
0 &\geq \psi(h_0)-\psi(h_1)\\
&\geq \int_0^\infty \left\{H_0(t)-H_1(t)+\left(1-\frac{h_0(t)}{h_1(t)}\right)\mathbb{I}_{t\neq X_{(n)}}\right\}d\mathbb{F}_n(t)\\
&= \alpha_0\left\{\int_0^\infty\left(t-\frac{1}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}\right)d\mathbb{F}_n(t)\right\}+\left\{1-\frac{1}{n}-\int_0^\infty H_1(t)d\mathbb{F}_n(t)\right\}\\
&\qquad+\int_0^{a_0}\left\{\frac{x^2}{2}-\int_0^x\frac{(x-t)^2}{2}d\mathbb{F}_n(t)-\int_0^x\frac{x-t}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}d\mathbb{F}_n(t)\right\}d\nu_0(x)\\
&\qquad\qquad+\int_{a_0}^\infty\left\{\int_x^\infty\frac{(t-x)^2}{2}d\mathbb{F}_n(t)-\int_x^\infty\frac{t-x}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}d\mathbb{F}_n(t)\right\}d\mu_0(x)\\
&\geq \int_{a_0}^{a_1}\left\{\int_x^\infty\frac{(t-x)^2}{2}d\mathbb{F}_n(t)-\int_x^\infty\frac{t-x}{h_1(t)}\mathbb{I}_{t\neq X_{(n)}}d\mathbb{F}_n(t)\right\}d\mu_0(x),
\end{aligned}$$

where the last inequality follows from Lemma 3.3. Since $h_0\neq h_1$ (otherwise there is nothing to prove), 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 (3.6) applied to h_1 on $[a_0,a_1]$, we obtain that

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

which implies that

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

Consider the function

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

By (3.5) this is an increasing function in y. Therefore, it follows that $f(y) \geq 0$ holds also for all $y \geq a_1$. Then, for all $x \geq a_1$,

$$\begin{aligned}
& \int_x^\infty \left\{ \frac{(t-x)^2}{2} - \frac{t-x}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} \right\} d\mathbb{F}_n(t) \geq 0 \\
\Rightarrow \quad & \int_0^\infty \left\{ \frac{x^2}{2} - \frac{(t-x)^2}{2} + \frac{t-x}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} \right\} d\mathbb{F}_n(t) + \int_x^\infty \left\{ \frac{(t-x)^2}{2} - \frac{t-x}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} \right\} d\mathbb{F}_n(t) \geq 0 \\
\Rightarrow \quad & \frac{x^2}{2} - \int_0^x \frac{(t-x)^2}{2} d\mathbb{F}_n(t) + \int_0^x \frac{t-x}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t) \geq 0 \\
\Rightarrow \quad & \frac{x^2}{2} - \int_0^x \frac{(t-x)^2}{2} d\mathbb{F}_n(t) \geq \int_0^x \frac{x-t}{h_1(t)} \mathbb{I}_{t \neq X_{(n)}} d\mathbb{F}_n(t). \qquad \text{(A-6)}
\end{aligned}$$

Next, calculating as above, we obtain that

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

by Lemma 3.3 and (A-6).

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 $a_2 > a_1 > a_0$. □

Department of Mathematics and Statistics, York University
e-mail: hkj@mathstat.yorku.ca

Department of Statistics, University of Washington
e-mail: jaw@stat.washington.edu