Local-EM and the EMS Algorithm

Chun-Po Steve Fan,* Jamie Stafford† and Patrick E. Brown‡

Abstract

The use of local likelihood methods (Hastie and Tibshirani 1987, Loader 1999) in the presence of data that is either interval or area censored leads naturally to the consideration of EM-type strategies, or rather local-EM algorithms. In this paper we consider a class of local-EM algorithms suitable for density or intensity estimation in the temporal or spatial context. We demonstrate that using a piecewise constant density function at the E-step results in the algorithm collapsing explicitly into an EMS algorithm of the type considered by Silverman et al. (1990).

This discovery has two advantages. Identifying a relationship between local likelihood and the EMS algorithm means the former provides a natural context for the latter, which is often referred to as *ad hoc* in the literature. In addition, the latter provides a set of tools to guide the use, and implementation, of local-EM algorithms. For example, we expose a previously unknown connection between local-EM algorithms and penalized likelihood that is analogous to the more familiar pairing of EM and likelihood. Examples include exploring the spatial structure of the disease Lupus in the City of Toronto.

Keywords: density estimation; intensity estimation; interval and area censoring; local likelihood; panel counts; lupus; penalized likelihood; self-consistency

1 Introduction

In this paper we consider the use of local likelihoods for density and intensity estimation when data are only partially observed. Here data may be interval censored, they may be

*Analysis Group
†University of Toronto
‡University of Toronto and Cancer Care Ontario

temporal and come in the form of panel counts, or they may be spatial and area censored. Whatever form the censoring takes, the use of local likelihood techniques naturally leads to the consideration of local-EM algorithms.

The use of local-EM algorithms has precedents in the literature. Examples include Tanner and Wong (1987), Betensky et al. (1999), Braun et al. (2005), Tolusso and Cook (2008) and so on. Here the method of implementation can vary where, for instance, implementation of the E-step might involve multiple imputation or Monte Carlo integration or numerical integration and so on. Despite this, and the interesting developments given in these papers, one common challenge that remains is demonstrating convergence of the local-EM algorithm to a fixed point. Furthermore, it is not clear whether this fixed point maximized any particular criterion. For example, while we know EM is a tool useful for maximizing likelihoods or posterior distributions, what is the analogy for local-EM?

In this paper we consider implementing the E-step of a local-EM algorithm by approximating conditional expectations using a piecewise constant density function. This results in the local-EM algorithm collapsing explicitly into an EMS algorithm of the type proposed in Silverman et al. (1990). There an EMS algorithm is constructed by simply adding a smoothing step to the expectation and maximization steps of the usual EM algorithm. Silverman et al. (1990) refer to this method as being *ad hoc*.

Identifying a relationship between local-EM and the EMS algorithm has two advantages. First, it embeds the EMS algorithm in the local likelihood context where it is seen to arise naturally as an implementation of a local-EM algorithm; thus it is not *ad hoc* [1]. Secondly, the EMS algorithm has been extensively studied. Much is known about its convergence (Latham, 1995) and its relationship to penalized likelihood (Nychka, 1990). This can be used to inform the use of local-EM. For example, the latter suggests a previously unknown connection between local-EM and penalized likelihood that is analogous to the more familiar pairing of EM and likelihood.

This paper has the following structure. In §2, we summarize a common notation used for three contexts: failure time processes, temporal and spatial inhomogeneous Poisson processes. In §3 we expose the relationship between local-EM and the EMS algorithm in a way that captures each of these contexts. We describe local-EM in §3.1 and in §3.2 demonstrate

[1] Nychka (1990) demonstrates that a modified EMS algorithm is related to penalized likelihood. As a result he also suggests that the EMS algorithm is not *ad hoc*.

that EMS arises naturally as an implementation of the local-EM algorithm. In §3.3 we prove the resulting EMS iterate converges uniformly to its local-EM counterpart.

In §4 we exploit the relationship between local-EM and the EMS algorithm to gain insight into convergence issues and to expose the role of local-EM. In §4.1 results of Latham (1985) are used to demonstrate the uniqueness of a fixed point. In addition, we provide an upper bound for local convergence and give conditions for this upper bound to shrink towards zero. In §4.2 we demonstrate that the use of an equivalent kernel in a local-EM algorithm leads to the modification necessary to maximize a penalized likelihood (Nychka 1990). This result suggests that, at least for the contexts considered in this paper, local-EM and penalized likelihood may be paired in a manner analogous to the pairing of EM and likelihood. The first example of §5 is useful in exemplifying this result.

We return to each context of §2 in the examples given in §5. Here we consider multivariate density estimation for failure time data, intensity estimation for panel count data, and finally, estimation of the spatial distribution of the disease lupus in the Greater Toronto Area. This last example may be viewed as extending the image reconstruction techniques of Silverman et al. (1990) to an epidemiological setting where, in the past, it was unclear how to proceed. One common theme throughout is that the local-EM algorithm is seen to explicitly extend the self-consistency algorithms of Turnbull (1976), Hu et al. (2008) and Vardi et al. (1985). Another is that, in simple cases, the local-EM algorithm reduces to the methods of Jones (1989) for smoothing histograms and Brillinger (1990, 1991, 1994) for smoothing spatially aggregated data. Simple cross validation is used to select an approriate bandwidth in the third example and the second example presents a simple simulation study that favour local-EM and suggest is has reasonable asymptotic properties. Rigorous treatment of these issues is difficult and beyond the scope of this paper. Concluding remarks may be found in §6

2 Some initial details

In the context of this paper we assume a study consists of n independent subjects where the time or location of events follows either a failure time process or an inhomogeneous Poisson process. Events are partially observed where they are only known to have fallen into a particular interval of time or region in space. To permit general developments in §3 we introduce a common notation where its meaning is context dependent. For example, in

the spacial context $\mathcal{M}$ will represent a geographic region, while in the temporal context it represents $\Re$ or $\Re^+$.

2.1 Processes in time

For processes in time each subject is observed at a set of points $\mathcal{T}_i = \{\tau_{ij}, j = 1 \ldots J_i\}$ that are either prearranged or determined by a visit process that is assumed to be independent of the event process.

When the event process for each subject is a common failure time process we denote the failure time by X_i and we assume it eithers fall between two adjacent elements of $\mathcal{T}_i$ or that it is right censored. In either case X_i is interval censored where we denote the relevant interval as $S_i = [L_i, R_i]$ where $L_i, R_i \in \mathcal{T}_i \cup \{\infty\}$ and of course $X_i \in S_i$. Note that if the event time for the ith subject is right censored we set $R_i = \infty$. The observed data is then a sequence of independent intervals $S_1 \ldots S_n$ some of which may overlap. We let $\mathcal{Q} = \{Q_j; j = 1 \ldots, J\}$ denote the partition of the data defined by the collection of endpoints $\{L_i, R_i; i = 1, \ldots, n\}$. For example, if $n = 2$ and $S_1 = [0, 3],\ S_2 = [1, 2]$ then we would have $\mathcal{Q} = \{[0, 1], [1, 2], [2, 3]\}$. For this setting the density, $\lambda(x)$, of the failure time process is the object of central interest.

When the event process for each subject is a common inhomogeneous Poisson process we denote the collection of events for the ith subject as $X_i = \{X_{ik}; k = 1 \ldots N_i\}$ where N_i denotes the number of events observed for the ith subject. Here event times are still interval censored but there may be multiple events in each interval. In this setting $S_{ij} = [\tau_{ij}, \tau_{ij+1}]$ is referred to as the jth panel for the ith individual and we denote the number of events in the interval S_{ij} by $N_{ij} = \#\{X_{ik} \in S_{ij}\}$. Following the setup of Hu et al. (2008) we let

$$\mathcal{T} = \cup_i^n \mathcal{T}_i = \{\tau_j; j = 0 \ldots J\}$$

and again $\mathcal{Q}$ denotes a partition of the data where now $Q_j = [\tau_{j-1}, \tau_j]$. For this setting the intensity, $\lambda(x)$, of the Poisson process is the object of central interest.

2.2 Processes in space

Rather than time, we may consider a series of inhomogeneous Poisson processes in space. Here the $X_i = \{X_{ik}; k = 1 \ldots N_i\}$ now represent sets of event locations for overlapping regions, or maps, $\mathcal{M}_i$ whose union we denote as $\mathcal{M} = \cup_i \mathcal{M}_i$. To accommodate nuisance

sources of spatial variation we assume the intensity surface of X_i takes the form

$$\rho_i(s) = \mathcal{O}_i(s)\lambda(s).$$

Here $\mathcal{O}_i(s)$ is an offset surface known *a priori* and $\lambda(s)$ is a relative risk surface that is assumed to be smooth.

The challenge in this context is that the maps have differing tessellations. We denote a subregion of map $\mathcal{M}_i$ by S_{ij} and the partition $\mathcal{Q}$ now results from overlaying the n individual maps. Here the X_{ik} are now assumed to be area censored and hence they are only known to fall in some subregion S_{ij}. Finally, for reasons given in §5.3 it is reasonable to assume that the offset surface O_i is constant over all regions S_{ij} such that

$$\mathcal{O}_i(u) = \mathcal{O}_{ij}, \;\; u \in S_{ij}.$$

Finally, for both the temporal and spatial context we have suggested the partition $\mathcal{Q}$ may be determined in a natural and obvious way from the data. However, there is no particular argument against choosing $\mathcal{Q}$ in any convenient fashion, say, as a grid on an axis or as a set of pixels for a map. In fact, for the theoretical developments in §4 and the appendix, it becomes convenient to consider an arbitrary partition whose elements shrink in volume. In any case, while the examples of §5 concern data driven partitions the results given in §3 & 4 hold for any particular partition.

3 Local-EM and the EMS algorithm

In this section we consider the use of local likelihoods for the flexible estimation of the function λ for the settings considered in §2. The treatment is general with the local likelihood having the form

$$\mathcal{L}_x(\lambda) = \sum_{ik} K_h(X_{ik} - x) \log\{\lambda(X_{ik})\} - \sum_i \int_{\mathcal{M}} \mathcal{O}_i(u) K_h(u - x)\lambda(u)\, \mathrm{d}u. \tag{1}$$

Here $K_h(z) = K(z/h)/h$ is a positive kernel function with $\int K(z)\,\mathrm{d}z = 1$. Following Loader (1999) we consider the polynomial approximation

$$\log\{\lambda(u)\} \approx \mathcal{P}(u - x) = \sum_{j=0}^{p} a_j (u - x)^j.$$

with coefficients $\mathbf{a} = \{a_0, a_1, \ldots, a_p\}$. Substituting the above into expression (1) and maximizing with respect to $\mathbf{a}$ yields the estimate $\hat{\lambda}(x) = \exp(\hat{a}_0)$.

3.1 Local-EM

When the data are interval or area censored with $X_{ik} \in S_{ij}$, one may consider replacing (1) with

$$\mathcal{L}_x(\mathbf{a}) = \sum_{ik} E_\lambda[K_h(X_{ik} - x)\mathcal{P}(X_{ik} - x; \mathbf{a})|X_{ik} \in S_{ij}] \\ - \sum_i \int_{\mathcal{M}} \mathcal{O}_i(u) K_h(u - x) \exp\{\mathcal{P}(u - x; \mathbf{a})\} \, \mathrm{d}u. \quad (2)$$

This leads to a local-EM algorithm that cycles through two steps at each iteration:

E-step: compute the relevant expectations using the current estimate of $\lambda(\cdot)$

M-step: maximize $\mathcal{L}_x(\mathbf{a})$ to get updated estimates of $\mathbf{a}$ and $\lambda(\cdot)$.

The algorithm differs from a typical EM algorithm because, at the E-step, expectation is computed with respect to an estimate of the infinite dimensional parameter λ while, at the M-step, we only estimate this parameter locally at x. As such the typical arguments concerning convergence of the EM algorithm cannot be brought to bear. Furthermore, if the local-EM algorithm converges to a fixed point $\hat{\lambda}$, it is not clear what criterion this fixed point optimizes.

Since the X_i are assumed to be realizations of a Poisson process the first term of (2) may be written as $\sum_{ij} N_{ij} E_\lambda[K_h(X - x)\mathcal{P}(X - x; \mathbf{a})|X \in S_{ij}]$, and in general the local-EM algorithm can be written as

$$\hat{\lambda}^{r+1}(x) = \sum_{ij} N_{ij} \mathrm{E}_{\hat{\lambda}^r} \left[K_h \left(X - x \right) | \, X \in S_{ij} \right] / \Psi_h(\hat{\mathbf{a}}^{r+1}) \quad (3)$$

where

$$\Psi_h(x; \mathbf{a}) \;\; = \;\; \sum_i \int_{\mathcal{M}} \mathcal{O}_i(u) K_h(u - x) \exp\left\{ \mathcal{P}(u - x) - a_0 \right\} \, \mathrm{d}u$$

and $\hat{\mathbf{a}}^{r+1}$ solves the local likelihood equations based on $\mathcal{L}_x(\mathbf{a})$ with λ replaced by $\hat{\lambda}^r$. Note that, given the offset surface is assumed to be constant over each region S_{ij}, expectation is computed with respect to the conditional density

$$\frac{\hat{\rho}^r(u)}{\int_{S_{ij}} \hat{\rho}^r(x) \, \mathrm{d}x} = \frac{\hat{\lambda}^r(u)}{\int_{S_{ij}} \hat{\lambda}^r(x) \, \mathrm{d}x} \quad (4)$$

at the rth iteration.

3.2 Implementation as an EMS algorithm

The above algorithm may be implemented using a discretization of λ. This effectively reduces an infinite dimensional estimation problem to one that has finite dimension. To this end we define a piecewise constant function over the partition $\mathcal{Q}$ as follows

$$g_\phi(x;\mathcal{Q}) = \|Q_j\|^{-1}\int_{Q_j}\phi(u)\,\mathrm{d}u \quad \text{for } x\in Q_j,$$

where ϕ is any integrable function over all Q_j's. The function g_ϕ may be formally referred to as the $\mathcal{Q}$-approximant of ϕ (see Royden (1988) for details). Now let $\Lambda_j = \int_{Q_j}\lambda(u)\,\mathrm{d}u$ and denote the collection of all Λ_j's by the vector $\mathbf{\Lambda}$. Finally, denote the indicator function of the set $S_{ij}\cap Q_l$ by $\mathcal{I}_{ijl}$.

Rather than using $\hat{\lambda}_r$ to compute conditional expectations, consider simplifying the iteration by using the piecewise constant function

$$\bar{\lambda}^r = g_{\hat{\lambda}^r}.$$

Here the conditional density (4) is replaced by

$$\bar{f}^r_{S_{ij}}(x) = \frac{\hat{\Lambda}^r_\ell\mathcal{I}_{ij\ell}}{\|Q_\ell\|\sum_m\hat{\Lambda}^r_m\mathcal{I}_{ijm}} \text{ for } x\in Q_\ell,$$

and $\mathrm{E}_{\hat{\lambda}^r}\left[K_h\left(X-x\right)|\,X\in S_{ij}\right]$ becomes

$$\begin{aligned}\mathrm{E}_{\bar{\lambda}^r}\left[K_h(X-x)|\,X\in S_{ij}\right] &= \int_{S_{ij}} K_h(u-x)\bar{f}^r_{S_{ij}}(u)\,\mathrm{d}u\\ &= \sum_\ell\int_{Q_\ell}K_h(u-x)\,\frac{\hat{\Lambda}^r_\ell\mathcal{I}_{ij\ell}}{\|Q_\ell\|\sum_m\hat{\Lambda}^r_m\mathcal{I}_{ijm}}\,\mathrm{d}u\\ &= \sum_\ell\frac{\hat{\Lambda}^r_\ell\mathcal{I}_{ij\ell}\int_{Q_\ell}K_h(u-x)\,\mathrm{d}u}{\|Q_\ell\|\sum_m\hat{\Lambda}^r_m\mathcal{I}_{ijm}}. \qquad (5)\end{aligned}$$

Substitution of the above into expression (3) leads to the computation of $\hat{\Lambda}^{r+1}_s$ at the next iteration. This in turns leads to the simple iteration

$$\begin{aligned}\hat{\Lambda}^{r+1}_s &= \int_{Q_s}\hat{\lambda}^{r+1}(x)\,\mathrm{d}x\\ &= \int_{Q_s}\left\{\sum_{ij\ell}N_{ij}\frac{\hat{\Lambda}^r_\ell\mathcal{I}_{ij\ell}}{\|Q_\ell\|\sum_m\hat{\Lambda}^r_m\mathcal{I}_{ijm}}\frac{\int_{Q_\ell}K_h(u-x)\,\mathrm{d}u}{\Psi_h(x;\hat{\mathbf{a}}^{r+1})}\right\}\mathrm{d}x\\ &= \sum_{ij\ell}N_{ij}\frac{\hat{\Lambda}^r_\ell\mathcal{I}_{ij\ell}}{\|Q_\ell\|\sum_m\hat{\Lambda}^r_m\mathcal{I}_{ijm}}\int_{Q_s}\frac{\int_{Q_\ell}K_h(u-x)\,\mathrm{d}u}{\Psi_h(x;\hat{\mathbf{a}}^{r+1})}\,\mathrm{d}x.\end{aligned}$$

The iteration may be conveniently expressed in terms of matrices as

$$\hat{\mathbf{\Lambda}}^{r+1} = \mathfrak{M}(\hat{\mathbf{\Lambda}}^r)\mathcal{K}_h(\hat{\mathbf{\Lambda}}^r), \tag{6}$$

Here $\mathcal{K}_h$ is a J-by-J smoothing matrix with entries

$$[\mathcal{K}_h]_{\ell s} = \frac{\tilde{\mathcal{O}}_\ell}{\|Q_\ell\|} \int_{Q_s} \frac{\int_{Q_\ell} K_h(u-x)\,\mathrm{d}u}{\Psi_h(x;\hat{\mathbf{a}}^{r+1})}\,\mathrm{d}x \tag{7}$$

and $\mathfrak{M}(\hat{\mathbf{\Lambda}}^r)$ is a J dimensional row vector whose ℓth entry is

$$[\mathfrak{M}(\hat{\mathbf{\Lambda}}^r)]_\ell = \sum_{ij} \frac{N_{ij}}{\tilde{\mathcal{O}}_\ell} \frac{\hat{\Lambda}_{r\ell}\mathcal{I}_{ij\ell}}{\sum_m \hat{\Lambda}_{rm}\mathcal{I}_{ijm}}, \tag{8}$$

where $\tilde{\mathcal{O}}_\ell = \sum_{ij} \mathcal{I}_{ij\ell}\mathcal{O}_{ij}$. The latter is recognized as a step in an EM algorithm (§5) and hence the iteration (6) is seen to explicitly involve an expectation, maximization *and* smoothing step. That is, by discretizing λ our implementation of the local-EM algorithm has resulted explicitly in an EMS algorithm. EMS algorithms were first proposed by Silverman et al. (1990) as an *ad hoc* method for improving the behaviour of the EM algorithm by including a smoothing step. Here they are seen to arise formally from local likelihood considerations when data are interval or area censored. Further comparisons with Silverman are given in §6.

Remark: The discretization of λ over the partition $\mathcal{Q}$ resulted in a local-EM algorithm collapsing explicitly into an EMS algorithm. We make a distinction between the local-EM iterate (3) and its EMS counterpart (9). At the $(r+1)$st iteration our estimate of λ will be given by

$$\left\{ \sum_{ij\ell} N_{ij} \frac{\hat{\Lambda}^r_\ell \mathcal{I}_{ij\ell}}{\sum_m \hat{\Lambda}^r_m \mathcal{I}_{ijm}} \frac{\int_{Q_l} K_h\left(u-x\right)\,\mathrm{d}u}{\|Q_\ell\|\Psi_h(x;\hat{\mathbf{a}}^{r+1})} \right\}. \tag{9}$$

This differs from (3), and the relation between the two is discussed in the next section.

3.3 Uniform Convergence of EMS to Local-EM

We consider the discretization (9) for an arbitrary partition where we let $J \to \infty$ and $\max_j \|Q_j\| \downarrow 0$. We demonstrate that the EMS iterate converges to its local-EM counterpart in the $\mathcal{L}^1$ norm as well as uniformly. This result suggests local-EM and EMS techniques may be thought of synonymously.

Consider a partition $\mathcal{Q}$ based on a set of J equally spaced grid points over a finite region $\mathcal{M}$. Without the loss of generality, we consider the partition where elements are squares centred at these grid points. We demonstrate that the EMS iterate will converge in $\mathcal{L}^1$ to its local-EM counterpart as $J \to \infty$ and $\max_j \|Q_j\| \downarrow 0$. For the sake of clarity, we restrict the attention to the locally constant case.

Without the loss of generality, assume $\mathcal{M}_i = \mathcal{M}$ for all i. Denote the EMS and local-EM iterates as $\hat{\lambda}_J^r$ and $\hat{\lambda}_\infty^r$, respectively. In addition, assume $S_{ij} \subseteq \mathcal{M}$ for all i, j where $\|\mathcal{M}\| < \infty$. Furthermore, let $K(z)$ be a symmetric positive kernel with compact support where $\int K(z)\,\mathrm{d}z = 1$. Finally, define a norm on $\mathcal{M}$ to be $\|\lambda\|_1 = \int_{\mathcal{M}} |\lambda(u)|\,\mathrm{d}u$ and interpret the convergence of the function f to the function g to mean that $\|f - g\|_1 \to 0$ as $J \to \infty$. This we denote as $f \xrightarrow{\mathcal{L}^1} g$. These details permit the statement of the following theorem:

Theorem 3.1. *Define $\mathcal{F}_1 = \{\lambda \in \mathcal{L}^1 \mid \lambda$ is non-negative with $\lambda(x) > 0$ for all $x \in \mathcal{M}\}$. For a common initial value $\hat{\lambda}_0 \in \mathcal{F}_1$, we have, for all $r = 1, 2, \ldots$,*

A. $\hat{\lambda}_J^r \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^r$, and

B. $\hat{\lambda}_J^r$, $\hat{\lambda}_\infty^r \in \mathcal{F}_1$.

Theorem 3.1 can be proved by induction using results from operator theory. Define $\mathcal{H}_x$ to be $\mathcal{H}_x : \mathcal{L}^1 \mapsto \mathcal{L}^1$ such that

$$\mathcal{H}_x(\lambda) = \int \frac{K_h(u - x)}{\int_{\mathcal{M}} \mathcal{O}(u) K_h(u - x)\,\mathrm{d}u} \lambda(u)\,\mathrm{d}u,$$

where $K_h(\cdot) = K(\cdot/h)/h$ for some $h > 0$ and $\mathcal{O} = \sum_i \mathcal{O}_i$. The proof relies on $\mathcal{H}_x(\lambda)$ being a bounded linear functional as well as some other basic results in operator theory stated as lemmas below. These lemma that may be found in Royden (1988).

Lemma 3.1. *Let $\lambda \in \mathcal{L}^1$. Then the $\mathcal{Q}_J$-approximant of λ converges in $\mathcal{L}^1$ to λ on $\mathcal{M}$ as $J \to \infty$; that is, $\bar{\lambda} \xrightarrow{\mathcal{L}^1} \lambda$.*

Lemma 3.2. *If $\int_{\mathcal{M}} \mathcal{O}(u) K_h(u - x)\,\mathrm{d}u \geq c > 0$, then $\mathcal{H}_x$ is a linear bounded functional for all $f \in \mathcal{L}^1$. That is, for all $x, a, b \in \Re$, $\mathcal{H}_t(af + b) = a\mathcal{H}_x(f) + b$, and there exists a real number M_h such that $\mathcal{H}_x(f) \leq M_h \|f\|_1$.*

Proof of Theorem 3.1: Consider a fixed h and n throughout the entire proof. Let $r = 1$. Assume $\int_{\mathcal{M}} \mathcal{O}(u) K_h(u-x)\,\mathrm{d}u \geq c > 0$. Note that $\int_{S_{ij}} \bar{\lambda}_0(u)\,\mathrm{d}u = \int_{S_{ij}} \hat{\lambda}_0(u)\,\mathrm{d}u$ by the definition of $\bar{\lambda}_0$. Repeated use of the triangle inequality gives

$$\begin{aligned}
\left\|\hat{\lambda}_J^1 - \hat{\lambda}_\infty^1\right\|_1 &\leq \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}^0(u)\,\mathrm{d}u\right)^{-1} \int_{\mathcal{M}} \int_{S_{ij}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u) K_h(u-x)\,\mathrm{d}u} \left|\bar{\lambda}^0(u) - \hat{\lambda}^0(u)\right| \mathrm{d}u\,\mathrm{d}x \\
&\leq \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}^0(u)\,\mathrm{d}u\right)^{-1} \int_{\mathcal{M}} \int_{\mathcal{M}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u) K_h(u-x)\,\mathrm{d}u} \left|\bar{\lambda}^0(u) - \hat{\lambda}^0(u)\right| \mathrm{d}u\,\mathrm{d}x \\
&\leq \|\mathcal{M}\| \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}^0(u)\,\mathrm{d}u\right)^{-1} M_h \left\|\bar{\lambda}^0 - \hat{\lambda}^0\right\|_1 .
\end{aligned}$$

Here the last inequality is due to the finiteness of $\mathcal{S} \subset \mathcal{M}$ and Lemma 3.2. By Lemma 3.1, $\bar{\lambda}^0 \xrightarrow{\mathcal{L}^1} \hat{\lambda}^0$, thus $\hat{\lambda}_J^1 \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^1$. Moreover, Lemma 3.2 also ensures that $\hat{\lambda}_J^1$ and $\hat{\lambda}_\infty^1$ both belong to the class $\mathcal{F}_1$.

Induction Step: Assume that $\hat{\lambda}_J^r \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^r$ and $\hat{\lambda}_J^r$, $\hat{\lambda}_\infty^r \in \mathcal{F}_1$. Let $b_{ij}^r = \dfrac{\int_{S_{ij}} \hat{\lambda}_\infty^r(v)\,\mathrm{d}v}{\int_{S_{ij}} \bar{\lambda}_J^r(v)\,\mathrm{d}v}$. With the repeated use of the triangle inequality, we have

$$\begin{aligned}
&\|\hat{\lambda}_J^{r+1} - \hat{\lambda}_\infty^{r+1}\| \\
&\quad\leq \sum_{ij} N_{ij} \int_{\mathcal{M}} \int_{S_{ij}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u) K_h(u-x)\,\mathrm{d}u} \left|\left(\frac{\bar{\lambda}_J^r(u)}{\int_{S_{ij}} \bar{\lambda}_J^r(v)\,\mathrm{d}v} - \frac{\hat{\lambda}_\infty^r(u)}{\int_{S_{ij}} \hat{\lambda}_\infty^r(v)\,\mathrm{d}v}\right)\right| \mathrm{d}u\,\mathrm{d}x \\
&\quad\leq \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}_\infty^r(v)\,\mathrm{d}v\right)^{-1} \int_{\mathcal{M}} \int_{\mathcal{M}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u) K_h(u-x)\,\mathrm{d}u} \left|b_{ij}^r \bar{\lambda}_J^r(u) - \hat{\lambda}_\infty^r(u)\right| \mathrm{d}u\,\mathrm{d}x \\
&\quad\leq \|\mathcal{M}\| \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}_\infty^r(v)\,\mathrm{d}v\right)^{-1} M_h \left\|b_{ij}^r \bar{\lambda}_J^r - \hat{\lambda}_\infty^r\right\|_1 .
\end{aligned}$$

The last inequality is again due to Lemma 3.2. Now since the induction assumption implies $\bar{\lambda}_J^r \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^r$ and $b_{ij}^r \to 1$, we have, for all i, j,

$$\left\|b_{ij}^r \bar{\lambda}_J^r - \hat{\lambda}_\infty^r\right\|_1 \leq \left\|b_{ij}^r\, \bar{\lambda}_J^r - \bar{\lambda}_J^r\right\|_1 + \left\|\bar{\lambda}_J^r - \hat{\lambda}_\infty^r\right\|_1 \to 0.$$

In addition, it is evident that $\hat{\lambda}_k^{r+1}$ and $\hat{\lambda}_\infty^{r+1}$ belong to $\mathcal{F}_1$ provided that $\hat{\lambda}_k^r$, $\hat{\lambda}_\infty^r \in \mathcal{F}_1$. Hence, we have (**A**) $\hat{\lambda}_k^{r+1} \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^{r+1}$ on $\mathcal{S}$, and (**B**) $\hat{\lambda}_k^{r+1}$, $\hat{\lambda}_\infty^{r+1} \in \mathcal{F}_1$ by induction.

The mode of convergence can be further strengthened to uniform convergence should the kernel function be continuous and the initial estimate $\|\hat{\lambda}_0\|_1$ be bounded on $\mathcal{M}$. To see this, note that the mode of the convergence in (**A**) can be *pointwise*, i.e. $\hat{\lambda}_J^r(x) \longrightarrow \hat{\lambda}_\infty^r(x)$ for all $x \in \mathcal{M}$. Furthermore, the functional class defined by $\mathcal{H}_x(\lambda)$, where $\lambda \in \mathcal{F}_1$, forms a class of *equicontinuous* functions. In this case, pointwise convergence implies uniform convergence by Ascoli-Arzelá Theorem (Royden, 1988, page 169). We will state this property as a corollary without proof since the proof is very much similar to the one above.

Corollary 3.1. *Assume the conditions in Theorem 3.1. Furthermore, if K is continuous and $\|\hat{\lambda}_0\|_1$ is bounded on $\mathcal{M}$, then $\hat{\lambda}_J^r \longrightarrow \hat{\lambda}_\infty^r$ uniformly.*

Remark: The above results may be immediately extended beyond the locally constant case to local linear, quadratic and so on, provided one can give conditions for $\mathcal{H}_x(\lambda)$ to be a bounded linear functional.

4 Convergence and the role of Local-EM

Thus far we have exposed an interesting relationship between classes of algorithms that demonstrates the EMS algorithm arises naturally from local likelihood considerations. This occurs because of the way we have chosen to implement the local-EM algorithm. However, we could have instead chosen to implement this algorithm through multiple imputation, or by using MCEM, or through some other favorite techniques. So why EMS?

In this section, we exploit the relationship between local-EM and the EMS algorithm to gain insight into convergence issues and to expose the role of local-EM as being paired with penalized likelihood in a manner analogous to the pairing of EM and likelihood. We primarily focus on the local constant case as details are easier to follow.

4.1 Convergence of the Local-EM Algorithm

Our objective is to allow the algorithm (6) to iterate until it converges to a fixed point $\hat{\mathbf{\Lambda}}$ that solves

$$\hat{\mathbf{\Lambda}} = \mathfrak{M}(\hat{\mathbf{\Lambda}})\mathcal{K}_h(\hat{\mathbf{\Lambda}}).$$

We rely on results given in Latham (1995) to demonstrate the uniqueness of $\hat{\mathbf{\Lambda}}$ and we study the rate of algorithmic convergence in the neighbourhood of $\hat{\mathbf{\Lambda}}$. Specifically, we derive an upper bound for the spectral radius of the EMS mapping and demonstrate that this upper bound shrinks toward zero as the value of h goes to infinity.

`Uniqueness:` In the situation in which the smoothing matrix is independent of $\mathbf{\Lambda}$, that is $\mathcal{K}_h(\mathbf{\Lambda}) = \mathcal{K}_h$ as in the local constant case, Latham (1995) shows the fixed-point solution is unique in the parameter space where $\Lambda_k > 0$ for all k. This implies that, if the EMS implementation of the local-EM algorithm is convergent at all, its iterations will converge to the unique fixed-point solution $\hat{\mathbf{\Lambda}}$. This result is consistent with the observation made in §2.4 of Braun et al. (2005).

`Convergence:` To demonstrate the convergence of (6) *locally*, it is sufficient to show that the EMS mapping has a spectral radius less than one at $\hat{\mathbf{\Lambda}}$ (see, e.g. Ortega (1990)). Let $\gamma(\mathbf{\Upsilon})$ be the spectral radius of the EMS mapping at $\mathbf{\Lambda}$. By the Perroni-Frobenius theorem,

$$\gamma(\mathbf{\Lambda}) \leq \max_s \sum_t \left[\partial\mathcal{M}\,\mathcal{K}_h\right]_{ts},$$

where $\partial\mathcal{M}$ is a $K \times K$ matrix with

$$[\partial\mathcal{M}]_{tk} \equiv \frac{\partial\mathcal{M}_k}{\partial\Lambda_t} = \begin{cases} \sum_{ij} \dfrac{N_{ij}}{\tilde{\mathcal{O}}_k} \dfrac{\sum_{\ell\neq k} \mathcal{I}_{ijk}\mathcal{I}_{ij\ell}\Lambda_\ell}{\left(\sum_\ell \mathcal{I}_{ij\ell}\Lambda_\ell\right)^2} & \text{when } k = t \\ \sum_{ij} \dfrac{N_{ij}}{\tilde{\mathcal{O}}_k} \dfrac{-\mathcal{I}_{ijt}\mathcal{I}_{ijk}\Lambda_k}{\left(\sum_\ell \mathcal{I}_{ij\ell}\Lambda_\ell\right)^2} & \text{otherwise} \end{cases}.$$

It follows that

$$\begin{aligned} \gamma(\mathbf{\Lambda}) &\leq \max_s \sum_t \left[\partial\mathcal{M}\,\mathcal{K}_h\right]_{ts} \\ &\leq \sum_k \left| \sum_{ij} \frac{N_{ij}}{\|Q_k\|} \frac{\sum_{t\neq k} \mathcal{I}_{ijk}\mathcal{I}_{ijt}(\Lambda_t - \Lambda_k)}{\left(\sum_\ell \mathcal{I}_{ij\ell}\Lambda_\ell\right)^2} \right| \max_s \int_{Q_s} \frac{\int_{Q_k} K_h(u-x)\,\mathrm{d}u}{\sum_\ell \tilde{\mathcal{O}}_\ell \int_{Q_\ell} K_h(u-x)\,\mathrm{d}u}\,\mathrm{d}x \end{aligned} \tag{10}$$

We set the upper bound for $\gamma(\mathbf{\Lambda})$ to be the last inequality of (10).

`Density estimation:` Consider the local-EM method for density estimation. In this context, our aim is twofold. First, we show that, for a fixed $\mathbf{\Lambda}$, the upper bound can be made as small as possible by increasing h. Second, given a bounded parameter space, the local-EM algorithm is globally convergent when h is sufficiently large.

By expressing (10) in terms of $\mathbf{p} = \{\int_{Q_j} f(u)\,\mathrm{d}u\}$, we have

$$\sum_k \left| \sum_i \frac{1}{\|Q_k\|} \frac{\sum_{t\neq k} \mathcal{I}_{ik}\mathcal{I}_{it}(p_t - p_k)}{\left(\sum_\ell \mathcal{I}_{i\ell} p_\ell\right)^2} \right| \max_s \int_{Q_s}\int_{Q_k} K_h(u-x)\,\mathrm{d}u\,\mathrm{d}x. \tag{11}$$

Since the kernel is symmetrical,

$$\begin{aligned}\max_s \int_{Q_s}\int_{Q_k} K_h(u-x)\,\mathrm{d}u\,\mathrm{d}x &= \int_{Q_k}\int_{Q_k} K_h(u-x)\,\mathrm{d}u\,\mathrm{d}x \\ &= \|Q_k\| \int_{(t_{k-1}-x^*)/h}^{(t_k-x^*)/h} K(z)\,\mathrm{d}z \quad \text{with some } x^* \in Q_k \end{aligned} \tag{12}$$

We obtain the following by taking the limit:

$$\begin{aligned}\lim_{h\to\infty} \|Q_k\| \int_{(t_{k-1}-x^*)/h}^{(t_k-x^*)/h} K(z)\,\mathrm{d}z &= \lim_{h\to\infty} \|Q_k\| \Big(\int_{-\infty}^{(t_k-x^*)/h} K(z)\,\mathrm{d}z - \int_{-\infty}^{(t_{k-1}-x^*)/h} K(z)\,\mathrm{d}z \Big) \\ &= \|Q_k\| \Big(\int_{-\infty}^{0} K(z)\,\mathrm{d}z - \int_{-\infty}^{0} K(z)\,\mathrm{d}z \Big) = 0\end{aligned}$$

Thereby, the upper bound shrinks toward 0 as h increases. The shrinking spectral radius, in turn, accelerates the convergence of the local-EM algorithm. Moreover, the upper bound (11) is a decreasing function of h. To see this, suppose $0 < h_1 < h_2$. Then

$$\begin{aligned}&\int_{J_s}\int_{J_k} K_{h_1}(u-x)\,\mathrm{d}u\,\mathrm{d}x - \int_{J_s}\int_{J_k} K_{h_2}(u-x)\,\mathrm{d}u\,\mathrm{d}x \\ &\quad= \int_{J_s} \left(\int_{(t_{k-1}-x)/h_1}^{(t_k-x)/h_1} K(z)\,\mathrm{d}z - \int_{(t_{k-1}-x)/h_2}^{(t_k-x)/h_2} K(z)\,\mathrm{d}z \right) \mathrm{d}x \\ &\quad= \int_{J_s} \left(\int_{(t_k-x)/h_2}^{(t_k-x)/h_1} K(z)\,\mathrm{d}z + \int_{(t_{k-1}-x)/h_1}^{(t_{k-1}-x)/h_2} K(z)\,\mathrm{d}z \right) \mathrm{d}x \geq 0\end{aligned}$$

This implies that, provided that the parameter space is bounded, there exists an $H > 0$ such that, for all $h > H$, the upper bound (11) is less than one for all $\mathbf{p}$. In this case, the EMS mapping is a contraction mapping, and the local-EM algorithm is globally convergent in the parameter space, as conjectured in Braun, Duchesne and Stafford (2005).

`Intensity estimation:` In the context of intensity estimation , the convergence of the local-EM algorithm is more complicated, and we can only demonstrate that the local-EM algorithm is locally convergent based on the result given by Green (1990). Let γ denote the spectral radius of the EMS mapping at the fixed point $\hat{\mathbf{\Lambda}}$. In this case, Green (1990) shows that $\gamma < 1$ for all $h > 0$. This implies that local-EM iterations converge at least in the

neighbourhood of $\hat{\mathbf{\Lambda}}$. Our experience suggests that, when h is sufficiently large, the iteration of the local-EM algorithm never fails to converge. In addition, the local-EM requires fewer steps than the EM algorithm to meet convergence criteria.

4.2 Local-EM and the Modified EMS algorithm

Nychka (1990) identified a relationship between EMS and penalized likelihood by demonstrating that a modified EMS algorithm maximizes

$$\mathcal{L}(\boldsymbol{\theta}) + \mathbf{Pen}(\boldsymbol{\theta}, \mathcal{K}). \tag{13}$$

Here $\mathcal{L}(\boldsymbol{\theta})$ is the appropriate nonparametric likelihood and is context dependent (Turnbull, 1976; Wellner and Zhang, 2000; Vardi et al., 1985). The parameter $\boldsymbol{\theta}$ is a vector with components θ_{ij} such that $\theta_{ij}^2 = \Lambda_{ij}$ for all i, j, and $\mathbf{Pen}(\boldsymbol{\theta}, \mathcal{K})$ is a penalty function that depends on both $\boldsymbol{\theta}$ and some smoothing matrix $\mathcal{K}$. Below we demonstrate that, with the appropriate choice of kernel, namely an equivalence kernel, the local-EM algorithm may be used to maximize a penalized likelihood function. This occurs because the equivalent kernel leads, under the discretization of λ, to Nychka's modification of the EMS algorithm.

We begin by first considering the following penalty:

$$\mathbf{Pen}(\boldsymbol{\theta}, \mathcal{K}) = \boldsymbol{\theta}^T \mathbf{R} \boldsymbol{\theta}.$$

$\mathbf{R} = \mathcal{K}^{-1} - \tilde{\boldsymbol{\mathcal{O}}}$, where $\mathcal{K}$ is the smoothing matrix in (7) and $\tilde{\boldsymbol{\mathcal{O}}}$ is a diagonal matrix with entries $\tilde{\mathcal{O}}_\ell$. Next, we explore the relationship between this penalized likelihood and the local-EM algorithm by considering the following function

$$(1/\lambda(u))^{1/2} K_h(u - x),$$

where λ is the smooth component of the true density or intensity, and $K_h(u - x)$ is any symmetric positive kernel with compact support. Let $c(x)$ be the normalizing constant respect to u such that,

$$K_h^*(u - x) = c^{-1}(x) K_h(u - x)/\lambda^{1/2}(u) \quad \text{and} \quad \int K_h^*(u - x)\, \mathrm{d}u = 1. \tag{14}$$

We refer to K_h^* as an equivalent kernel. Next, consider the use of the equivalent kernel with the $\mathcal{Q}$-approximant $\bar{\lambda}^r$ in our local-EM algorithm while assuming $\|Q_j\| = \|Q\|$ for all j. Using the first-order approximation of (14), i.e.

$$K_h^*(u - x) = (\lambda(x)/\lambda(u))^{1/2} K_h(u - x) + o(h),$$

results in the conditional expectation $\mathrm{E}_{\hat{\lambda}^r}\left[K_h^*(X-x) \mid X \in S_{ij}\right]$ being approximated as follows:

$$\mathrm{E}_{\bar{\lambda}^r}\left[K_h^*(X-x) \mid S_{ij}\right] \;=\; \sum_k \left(\frac{\hat{\Lambda}_\ell^r}{\hat{\Lambda}_k^r}\right)^{1/2} \|Q\|^{-1} \int_{Q_k} K_h(u-x)\,\mathrm{d}u \frac{\mathcal{I}_{ijk}\hat{\Lambda}_k^r}{\sum_m \mathcal{I}_{ijm}\hat{\Lambda}_m^r} + o(h)$$

for $x \in J_\ell$. This in turn gives the following iteration for $\mathbf{\Lambda}$:

$$\begin{aligned}\hat{\Lambda}_\ell^{r+1} &= n^{-1}\sum_{ijk}\left(\frac{\hat{\Lambda}_\ell^r}{\hat{\Lambda}_k^r}\right)^{1/2} \|Q\|^{-1} \int_{Q_\ell} \frac{\int_{Q_k} K_h(u-x)\,\mathrm{d}u}{\int_{\sum_i \mathcal{M}} \mathcal{O}_i(u) K_h^*(u-x)\,\mathrm{d}u}\,\mathrm{d}x \frac{\mathcal{I}_{ijk}\hat{\Lambda}_k^r}{\sum_m \mathcal{I}_{ijm}\hat{\Lambda}_m^r} \\ &= n^{-1}\sum_{ijk}\left(\frac{\hat{\Lambda}_\ell^r}{\hat{\Lambda}_k^r}\right)^{1/2} \|Q\|^{-1} \int_{Q_\ell}\int_{Q_k} K_h(u-x)\,\mathrm{d}u\,\mathrm{d}x \frac{1}{\tilde{\mathcal{O}}_\ell} \frac{\mathcal{I}_{ijk}\hat{\Lambda}_k^r}{\sum_m \mathcal{I}_{ijm}\hat{\Lambda}_m^r} + o(h) \qquad (15)\end{aligned}$$

The expression (15) can be re-expressed in the following matrix form:

$$\hat{\mathbf{\Lambda}}^{r+1} = \mathfrak{M}(\hat{\mathbf{\Lambda}}^r)\mathcal{K}_h^*\left(\hat{\mathbf{\Lambda}}^r\right). \qquad (16)$$

$\mathcal{K}_h^*\left(\hat{\mathbf{\Lambda}}^r\right) = \left(\hat{\mathbf{\Theta}}^r\right)^{-1} \tilde{\boldsymbol{\mathcal{O}}}\,\mathcal{K}_h\,\tilde{\boldsymbol{\mathcal{O}}}^{-1}\,\hat{\mathbf{\Theta}}^r$, where $\hat{\mathbf{\Theta}}^r = \mathrm{diag}(\hat{\theta}_k^r)$. Note that $\mathcal{O}_k = 1$ in the context considered by Silverman et al. (1990) and Nychka (1990). In addition, provided that $\mathcal{O}_k = 1$ for all k, the iteration (16) is recognized as Nychka's modified EMS algorithm with the smoothing matrix equal to $\mathcal{K} = \mathcal{K}_h$. In other words, this iteration can be regarded as a generalization of Nychka's modified EMS algorithm since it allows for offsets.

Remark: The theoretic results in $\mathcal{L}^1$ and uniform convergence stated in §3.3 can also be extended with the equivalent kernel in the local-EM algorithm. We will re-state Theorem (3.1) with the equivalent kernel (14) as the following proposition and include the proof in the appendix.

Proposition 4.1. *When the equivalence kernel (14) is used, we instead define*

$$\mathcal{F}_2 = \left\{\lambda \in \mathcal{L}^1 \middle| \lambda \text{ is non-negative with } \lambda(x) > 0 \text{ for all } x \in \mathcal{M} \text{ and } \int_{\mathcal{M}} \lambda^{1/2} < \infty\right\}.$$

For a common initial value $\hat{\lambda}_0 \in \mathcal{F}_2$, we have, for all $r = 1, 2, \ldots$,

A. $\hat{\lambda}_J^r \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^r$, *and*

B. $\hat{\lambda}_J^r,\ \hat{\lambda}_\infty^r \in \mathcal{F}_2$.

5 Examples

5.1 Density Estimation for Failure Time Data

Consider a bivariate failure time process, such as time to HIV infection and AIDS onset, or in flu epidemics an infected individual's time to onset of symptoms and the first time they transmit the infection to another individual. As infections are usually only revealed through repeated testing the event times are interval-censored and only known to fall between two consecutive clinic visits, one where the patient tests negative for the presence of the virus and a follow-up visit where they test positive. Multivariate density estimation in this context may be facilitated by (2), which now simplifies to

$$\mathcal{L}_x(\mathbf{a}) = \sum_{i=1}^{n} \mathrm{E}_\lambda[K_h(X_i - x)\mathcal{P}(X_i - x)|I_i] - n \int_{\mathcal{M}} K_h(u - x) \exp\{\mathcal{P}(u - x)\} \, \mathrm{d}u$$

where $\mathcal{O}_i(u) = 1$ for all $u \in \mathcal{M}$.

5.1.1 One dimension

For univariate failure time data, $\mathcal{M} = \Re$ and X_i is the event time for the ith individual which is only known to fall in the interval I_i. For the iteration (6) we recognize $\mathfrak{M}(\hat{\mathbf{\Lambda}}_r)$ as a step in the EM algorithm of Turnbull (1976). Furthermore, in the case of a histogram, $\{I_i\} \equiv \mathcal{Q}$, the local-EM algorithm (3) will only iterate once and reduces to methods of Jones (1989) for smoothing histograms. Finally, Braun et al. (2005) proposes a local-EM algorithm based on $\mathcal{L}_x(\mathbf{a})$ and, without being aware of it, develop an EMS implementation.

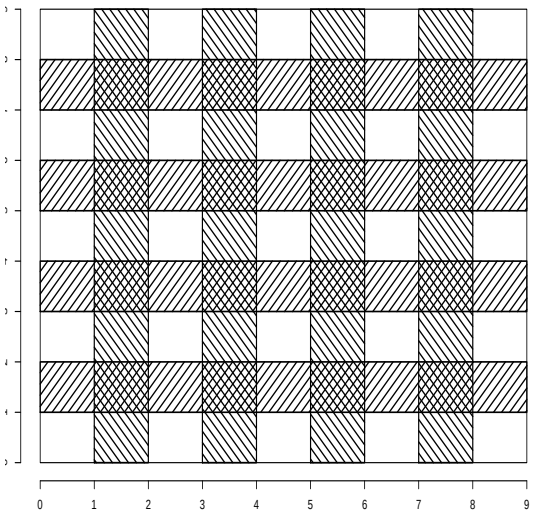
5.1.2 Two dimensions

When two events are being modelled, X_i is bivariate and $\mathcal{M} = \Re^2$. Typically, when data are doubly interval-censored, the estimation of NPMLE for the joint distribution is complex and a treatment of some of the issues can be found in Maathuis (2005). In the rest of this example, we consider a simple case that suggests local-EM in this context may ease these complexities and thus deserves further exploration. For example, the use of a local-EM algorithm does not require identifying maximal sets (the analog of inner most intervals in the univariate case) and thus does require use of the height map algorithm of Maathuis (2005).

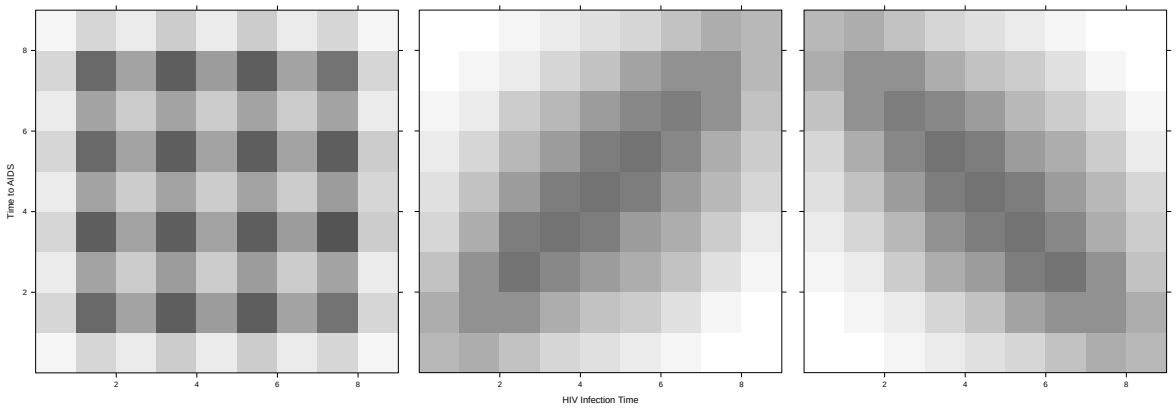
Also the solution is unique while the NPMLE is not. A Bayesian interpretation of local-EM provides insight into the latter.

A hypothetical example of bivariate interval-censored data is shown in Figure 1a. These data consists of eights bivariate intervals, represented by four horizontal and four vertical rectangles. Overlaying these observations forms a partition of 81 unit squares, and the intersections of these rectangles form the maximal sets. A NPMLE places all probability on these sets however it is not unique. For example, a uniform weight of 1/16 on all 16 sets, a weight of 1/4 on the positive diagonal sets, and a weight of 1/4 on the negative diagonal sets all maximize the nonparametric likelihood. Consequently, the EM iteration will converge to one of the solutions depending on the initial value. However, empirical evidence suggests otherwise for the local-EM algorithm.

Figure 1b shows the estimated intensity surface for a circular kernel. In this case the EMS iteration will always converge to a solution that favours uniform weighting regardless of the starting value. Figure 1c gives the intensity estimate using a kernel with an elliptical contour such that bandwidths are 1.5 and .15 in the x- and y-direction and rotated by 45 degrees. In this case, the local-EM algorithm converges to a solution that favours the positive diagonal. Similarly, when this kernel is instead rotated by -45 degrees, Figure 1d shows the local-EM algorithm converges to a solution that favours the negative diagonal. This behaviour can be interpreted in terms of the penalized likelihood of §4.2. Here the local-EM algorithm aims to maximize the penalized likelihood (13) where the penalty depends on the choice of kernel. The kernel leads us to favour one NPMLE over another *a priori*. Explicitly when the kernel is radially symmetrical, any deviations from the maximal eigenfunction are equally penalized. However, as the kernel becomes more elliptical, deviations in the direction of the major axis of the elliptical contour are penalized less than those in the direction of the minor axis.



(a) Regions corresponding to eight bivariate interval-censored data points



(b) EMS intensity estimate from circular kernel

(c) Elliptical kernel, positive-sloped major axis

(d) Elliptical kernel, negative-sloped major axis

Figure 1: An artificial example of bivariate interval-censored data (a), with EMS estimates of the intensity using three different kernels (b-d).

5.2 Intensity Estimation for Panel Count Data

In this example we consider the situation described in §2.1.2 where individuals are monitored in time and events follow an inhomogeneous Poisson process. Monitoring involves periodic assessments, so events are interval censored. Here, individuals may drop out of the study at different time points and we use $Y_i(x)$ to indicate if the ith individual is in the study at time x. Since an individual's event process is observable only before he/she drops out, the intensity of the observable process equals $Y_i(x)\lambda(x)$ (see Andersen (1993)). Finally, the number of at-risk individuals at any time x is denoted by $Y(x) = \sum_i Y_i(x)$.

We assume that the event, assessment and drop-out processes are independent of one another, and that the drop-out process is monotone. Under this setting, the local-EM algorithm still derives from (2) which now becomes

$$\sum_{ij} N_{ij} \mathrm{E}_\lambda \left[Y_i(X) K_h(X-x) \mathcal{P}(X-x) \mid X \in S_{ij} \right] - \int_{\mathcal{M}} Y(u) K_h(u-x) \exp\{\mathcal{P}(u-x)\} \, \mathrm{d}u$$

and the corresponding EMS implementation (6) has

$$[\mathfrak{M}(\hat{\mathbf{\Lambda}}^r)]_\ell = \sum_{ij} N_{ij} \frac{Y_i(\tau_\ell) \hat{\Lambda}^r_\ell \mathcal{I}_{ij\ell}}{Y(\tau_\ell) \sum_m \hat{\Lambda}^r_m \mathcal{I}_{ijm}}$$

and

$$[\mathcal{K}_h(\hat{\mathbf{\Lambda}}^r)]_{\ell s} = \frac{Y(\tau_\ell)}{\|Q_\ell\|} \int_{J_s} \frac{\int_{J_\ell} K_h(u-x) \, \mathrm{d}u}{\Psi_h(x; \mathbf{a}^{r+1})} \, \mathrm{d}x.$$

Note $\mathfrak{M}(\hat{\mathbf{\Lambda}}_r)$ is a step in the self-consistent algorithm of Hu et al. (2008).

5.2.1 A simulation study

A simulation study was carried out to examine the mean integrated squared error (MISE) of the local-EM estimator as well as several alternatives. Event times follow a Poisson process with intensity $\lambda(x)$ equal to a re-scaled gamma density function (shape = 9 and rate=3/4). Each subject is assumed to have a sequence of predetermined observation times $\tau_1, \tau_2, \ldots, \tau_J$ where $\tau_j = j$ and $J = 20$. However, subjects miss a visit with increasing probability, specifically, the probability of missing a visit equals $(\tau_j/20)^{1/4} - 0.05$. Finally, a subject's panel counts are obtained by aggregating events times among consecutive observed visits. Note that each subject is assumed to have no event at time 0.

For each of S samples, and for a fixed window size h, we compute several estimators of the intensity using a Gaussian kernel. For each estimator, $\hat{\lambda}$, we approximate its MISE as

$S^{-1}\sum_k \int(\hat{\lambda}_k(u) - \lambda(u))^2\,\mathrm{d}u$. This was performed for 40 different values of h between 0.05 and 3.95 with $S = 300$. The resulting MISE's for each estimator are plotted in Figure 2. The first estimator assumes no interval censoring has taken place and uses the exact event times themselves, rather than the panel counts. This is the gold standard. For the panel counts we use the partition $\mathcal{Q}$ and compute the local-EM estimator in both the constant and linear cases, that is, where the polynomial is truncated at the first or second term. In addition, as an alternative to, and competitor of, the local-EM estimators we consider simply smoothing the self-consistent estimator of Hu et al. (2008) after their EM algorithm has converged.

The results favour the local-EM estimator considerably. While the gold standard achieves the smallest MISE, the local-EM estimators track it quite closely and attain the next smallest MISE for a similar window size. Smoothing after the EM algorithm converges has the worst performance achieving a minimum MISE that is larger for a larger window size. This result is perhaps not all that surprising given that the λ is quite non-linear. In cases where λ is linear the improvements in MISE for the local-EM estimator are not as dramatic. Another simulation was performed in the spatial context with similar results.

5.3 The Spatial Structure of Lupus in the City of Toronto

In this example, we consider the setting described in §2.2 and investigate the spatial structure of lupus in the Greater Toronto Area (GTA). The lupus clinic at the Toronto Western Hospital records the census tracts where individuals with lupus reside, and has data from 1965 to 2007. If lupus is affected by a spatially varying environmental or social risk factor, it should result in a spatially smooth relative risk surface $\lambda(s)$.

Disease incidence is assumed to arise from an inhomogeneous Poisson process in space and time, in which the intensity is given as $\rho_k(x,t) = \lambda(x)\mathcal{O}_k(x,t)$ with the offset surface $\mathcal{O}_k$ given as

$$\mathcal{O}_k(x,t) = \beta(t)\theta_k P_k(x,t).$$

Here the subscript denotes the kth age-sex group, θ_k is the incidence rate for this group, $P_k(x,t)$ is the population intensity (in persons per km square), and $\beta(t)$ is the time trend. Using regionally aggregated case counts to estimate relative risk surface $\lambda(s)$ is the objective. The main complication is that boundaries of census regions used to aggregate the data change repeatedly over the study period. In the past it has been unclear how to proceed formally

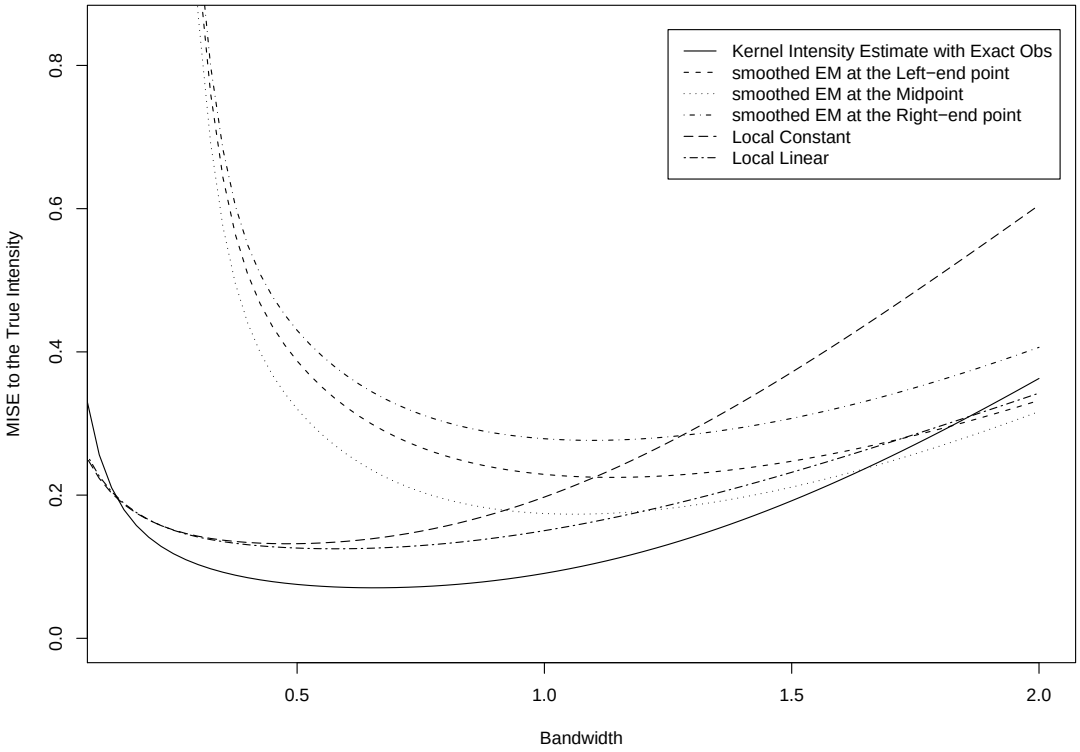


Figure 2: Mean integrated squared error as a function of bandwidth for various estimation methods. The proposed local-EM intensity estimate achieves the lowest overall MISE with a small bandwidth of 0.195, comparing to the smoothed EM estimate by placing expected increments at the centres of pixels.

with such time dependent boundaries and techniques in use are *ad hoc*. Local-EM resolves these issues by providing an automatic way to proceed.

Census periods are defined as beginning and ending at the mid-points between census years before and after a given census. Period i covers the years t_{i-1} to t_i and $i = 1 \ldots T$ where T is the total number of census periods during the study. The jth census region for the ith census period is denoted as S_{ij} and these regions have boundaries that vary between census periods. For simplicity, we assume $\beta(t)$ and the population $P_k(s,t)$ are constant within a census period so that $\beta(t) = \beta_i$ when t is in period i and

$$P_k(x,t) = P_{ik}(x) = P_{ijk}/\|S_{ij}\| \quad \text{for } x \in S_{ij}.$$

where P_{ijk} is the population count for group k in region S_{ij}. As a result of these simplifications, the offset is constant over the region S_{ij} within a census period.

$$\mathcal{O}_k(x,t) = \mathcal{O}_{ik}(x) = \beta_i \theta_k P_{ik}(x).$$

Finally, the available data are case counts of the form N_{ijk} for individuals in group k who were diagnosed with lupus during census period i while living in region S_{ij}.

The model is fitted in two stages. At the first stage, the spatial variation in $\lambda(x)$ is ignored so that case counts N_{ijk} may be assumed to be distributed as

$$N_{ijk} \sim \text{Poisson}(\theta_k \beta_i (t_i - t_{i-1}) P_{ijk}).$$

This allows β_i and θ_k to be estimated from a generalized linear model. At the second stage, β_i and θ_k are set to the values estimated at the first stage and treated as if they were known. They are then used to construct the offsets

$$\mathcal{O}_i(x) = \sum_k \mathcal{O}_{ik}(x).$$

These offsets are in turn used in the iteration (6) to estimate $\lambda(x)$.

Figure 3 shows the estimated intensity surface using the local-EM algorithm with locally constant risks within square grid cells and a bandwidth of 1.35km. The offsets were computed by calculating empirical rates by age and sex groups and applying these to the population data from the census for census data from 1971, 1981, 1991, and 2001. The ideal window size h was chosen by cross validation. The risk surface is fairly flat and near unity throughout most of the region, with an area of elevated risk near the centre of the downtown area of

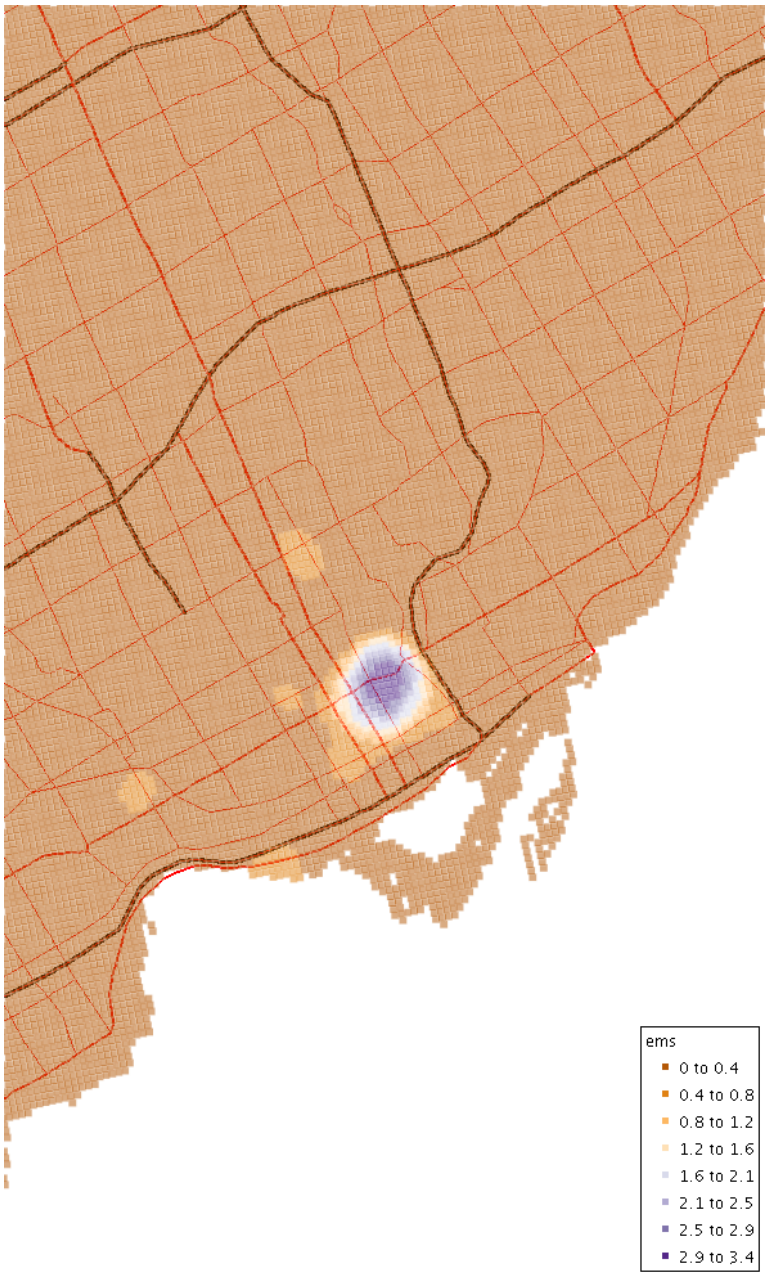


Figure 3: Estimated risk surface for lupus using EMS with a bandwith of 1350 m.

the city. This could be due to a risk factor not accounted for, such as ethnicity, or reporting bias due to the proximity to the clinic.

We conclude this example by making a comparison between the local-EM algorithm in this context to methods in the literature. Note that if we only have a single map ($j = 1$) then Q_ℓ and $S_{j\ell}$ coincide so that $I_{ij\ell} = 0$ for all $j \neq \ell$. As a result the kernel weight given in (5) simplifies to $\int_{Q_\ell} K_h(u-t)\,\mathrm{d}u/\|Q_\ell\|$, the algorithm (3) iterates once and the local-EM estimator simply becomes the Nadaraya-Watson estimator advocated by Brillinger (1990, 1991, 1994) in a series of papers concerning spatial smoothing where data are aggregated to regions within a map.

6 Discussion

Local likelihood can be seen as a semi-parametric method, providing a compromise between the power and theoretical rigour of parametric methods and the flexibility of kernel smoothing algorithms. local-EM provides a method for applying local likelihood in situations where interval or area censoring with irregular observed regions. By demonstrating that local-EM and the EMS algorithm are related, it is hoped that the computational advantages offered by EMS will lead to greater adoption of local-EM methods. Formulating EMS problems in the context of local likelihood allows for a natural and rigorous method of incorporating offsets.

A final comparison of local-EM to Silverman et al. (1990) permits further insights beyond what has already been discussed in the paper. In Silverman et al. (1990) quantities analogous to S_{ij} and Q_ℓ are referred to as observation and reconstruction bins respectively and the context concerns image reconstruction involving a single image rather than multiple maps say. As a result, example §5.3 could well be thought of as an extension of the image reconstruction techniques of Silverman et al. (1990) to an epidemiological setting. Furthermore, noting there are no offsets in Silverman et al. (1990), the expression (2.2) given there and $\mathfrak{M}(\hat{\mathbf{\Lambda}}^r)$ are related. For example, their weights p_{st} simplify in our setting to the indicator variables $\mathcal{I}_{ij\ell}$ because we have assumed the locations of events have been measured without error. This observation provides an avenue for extending the local-EM toolbox to settings where data are mismeasured, but this is beyond the scope of this paper. Finally, we note that in our context $\mathfrak{M}(\hat{\mathbf{\Lambda}}^r)$ is an extension of Vardi et al. (1985) to multiple maps where data are not mismeasured.

Acknowledgements

We are grateful to Dr. Paul Fortin for the provision of the lupus data. We would also like to acknowledge the Natural Sciences and Engineering Research Council of Canada for supporting this research through individual operating grants.

References

Andersen, P. (1993). *Statistical Models Based on Counting Processes*. Springer.

Betensky, R., Lindsey, J., Ryan, L., and Wand, M. (1999). Local EM estimation of the hazard function for interval-censored data. *Biometrics*, 55(1):238–245.

Braun, J., Duchesne, T., and Stafford, J. (2005). Local likelihood density estimation for interval censored data. *The Canadian Journal of Statistics*, 33(1):39–60.

Brillinger, D. (1990). Spatial-temporal modeling of spatially aggregate birth data. *Survey Methodology*, 16:255–269.

Brillinger, D. (1991). Locally weighted analysis of spatially aggregate birth data: Uncertainty estimation and display. In *Spatial Issues in Statistics*, volume 91 of *Proceedings of Statistics Canada Symposium*, Ottawa. Statistics Canada.

Brillinger, D. (1994). Examples of scientific problems and data analyses in demography, neurophysiology, and seismology. *Journal of Computational and Graphical Statistics*, pages 1–22.

Green, P. J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, (3):443–452.

Hu, X., Lagakos, S. W., and Lockhart, R. A. (2008). Generalized least squares estimation of the mean function of a counting process based on panel counts. *Statistica Sinica*, page Preprint.

Latham, G. (1995). Existence of EMS solutions and a-priori estimates. *SIAM Journal on Matrix Analysis and Applications*, 16:943–953.

Loader, C. (1999). *Local Regression and Likelihood*. Springer.

Maathuis, M. (2005). Reduction algorithm for the NPMLE for the distribution function of bivariate interval-censored data. *Journal of Computational & Graphical Statistics*, 14(2):352–362.

Nychka, D. (1990). Some properties of adding a smoothing step to the EM algorithm. *Statistics and Probability Letters*, 9(2):187–193.

Ortega, J. (1990). *Numerical Analysis: a Second Course*. Society for Industrial Mathematics.

Royden, H. (1988). *Real analysis*. Prentice-Hall.

Silverman, B., Jones, M., Wilson, J., and Nychka, D. (1990). A smoothed EM approach to indirect estimation problems, with particular, reference to stereology and emission tomography. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(2):271–324.

Tanner, M. A. and Wong, W. H. (1987). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, 29:23–32.

Tolusso, D. and Cook, R. (2008). Local likelihood for right and interval-censored recurrent event data. *University of Waterloo Technical Report*.

Turnbull, B. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38:290–295.

Vardi, Y., Shepp, L. A., and Kaufman, L. (1985). A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20.

Wellner, J. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Annals of Statistics*, 28(3):779–814.

A Appendix

A.1 Proof of Proposition 4.1

The proof is very much similar to the one for Theorem 3.1. However, it requires another useful result in integral theories, which will be stated as a lemma as follows:

Lemma A.1. *Let $\gamma_h(u,x) = g(x)K_h(u-x)f(u)$, where $f,\ g \in \mathcal{L}^1$ and $K_h \in \mathcal{L}^\infty$. Then $\gamma(u,t)$ is an $\mathcal{L}^1$ function on $\Re^2$ with*

$$\iint |\gamma_h(u,x)| \,\mathrm{d}u\,\mathrm{d}x \le M_h \cdot \|g\|_1 \cdot \|f\|_1 .$$

Proof of Proposition 4.1: Let $r = 1$. Assume $\int_{\mathcal{M}} \mathcal{O}(u)K_h(u-x)\,\mathrm{d}u \ge c > 0$. By the triangle inequality and Lemma A.1, we have

$$\left\|\hat{\lambda}_J^1 - \hat{\lambda}_\infty^1\right\|_1 \le \sum_{ij} N_{ij} \int_{\mathcal{M}} \int_{S_{ij}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u)K_h(u-x)\,\mathrm{d}u} \left| \frac{\bar{\lambda}_0^{1/2}(u)\,\bar{\lambda}_0^{1/2}(x) - \hat{\lambda}_0^{1/2}(u)\,\hat{\lambda}_0^{1/2}(x)}{\int_{S_{ij}} \hat{\lambda}_0(v)\,\mathrm{d}v} \right| \mathrm{d}u\,\mathrm{d}x$$

$$\le \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}_0(v)\,\mathrm{d}v\right)^{-1} \left\{ \int_{\mathcal{M}} \int_{\mathcal{M}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u)K_h(u-x)\,\mathrm{d}u} \bar{\lambda}_0^{1/2}(x) \left|\bar{\lambda}_0^{1/2}(u) - \hat{\lambda}_0^{1/2}(u)\right| \mathrm{d}u\,\mathrm{d}x \right.$$
$$\left. + \int_{\mathcal{M}} \int_{\mathcal{M}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u)K_h(u-x)\,\mathrm{d}u} \hat{\lambda}_0^{1/2}(u) \left|\bar{\lambda}_0^{1/2}(x) - \hat{\lambda}_0^{1/2}(x)\right| \mathrm{d}u\,\mathrm{d}x \right\}$$

$$\le \|\mathcal{M}\| \sum_{ij} N_{ij} \left(\int_{R_{ij}} \hat{\lambda}^0(v)\,\mathrm{d}v\right)^{-1} M_h \left(\left\|(\bar{\lambda}^0)^{1/2}\right\|_1 + \left\|(\hat{\lambda}^0)^{1/2}\right\|_1\right) \left\|(\bar{\lambda}^0)^{1/2} - (\hat{\lambda}^0)^{1/2}\right\|_1$$

By Lemma 3.1 and the continuous mapping theorem (CMT), $\left\|\bar{\lambda}_0^{1/2} - \hat{\lambda}_0^{1/2}\right\|_1 \to 0$. Therefore, $\hat{\lambda}_J^1 \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^1$. Furthermore, if we choose $\hat{\lambda}^0$ to be bounded above, then $\hat{\lambda}_J^1$ and $\hat{\lambda}_\infty^1$ will be also bounded. This, in turn, ensures that $\hat{\lambda}_J^1,\ \hat{\lambda}_\infty^1 \in \mathcal{F}_2$.

Induction Step: Assume that $\hat{\lambda}_J^r \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^r$, $\hat{\lambda}_J^r,\ \hat{\lambda}_\infty^r \in \mathcal{F}_2$, and bounded. Then the induction assumption immediately implies that

$$c_{ij}^r = \frac{\int_{S_{ij}} \hat{\lambda}_\infty^r(v)\,\mathrm{d}v}{\int_{S_{ij}} \hat{\lambda}_J^r(v)\,\mathrm{d}v} \to 1 \text{ for all } i,j \text{ and } (\bar{\lambda}_J^r)^{1/2} \xrightarrow{\mathcal{L}^1} (\hat{\lambda}_\infty^r)^{1/2} \text{ on } \mathcal{M}.$$

Similar to part (I), Lemma 3.1 and A.1 imply that

$$
\begin{aligned}
&\left\|\hat{\lambda}_J^{r+1} - \hat{\lambda}_\infty^{r+1}\right\|_1 \\
&\quad \leq \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}_\infty^r(x)\,\mathrm{d}x\right)^{-1} \int_{\mathcal{M}} \int_{S_{ij}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u) K_h(u-x)\,\mathrm{d}u} \Big| c_{ij}^r \left[\bar{\lambda}_J^r(u)\,\bar{\lambda}_J^r(x)\right]^{1/2} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad - \left[\hat{\lambda}_\infty^r(u)\,\hat{\lambda}_\infty^r(x)\right]^{1/2} \Big| \,\mathrm{d}u\,\mathrm{d}x \\
&\quad \leq \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}_\infty^r(x)\,\mathrm{d}x\right)^{-1} \int_{\mathcal{M}} \int_{\mathcal{M}} \frac{K_h(u-x)}{\int_{\mathcal{M}} \mathcal{O}(u) K_h(u-x)\,\mathrm{d}u} \left\{ c_{ij}^r \left(\bar{\lambda}_J^r(x) \left|\bar{\lambda}_J^r(u) - \hat{\lambda}_\infty^r(u)\right|\right)^{1/2} \right. \\
&\qquad\qquad \left. + \left|c_{ij}^r - 1\right| \left[\hat{\lambda}_\infty^r(u)\bar{\lambda}_J^r(x)\right]^{1/2} + \left(\hat{\lambda}_\infty^r(u) \left|\bar{\lambda}_J^r(x) - \hat{\lambda}_\infty^r(x)\right|\right)^{1/2} \right\} \mathrm{d}u\,\mathrm{d}x \\
&\quad \leq \|\mathcal{M}\| \sum_{ij} N_{ij} \left(\int_{S_{ij}} \hat{\lambda}_\infty^r(x)\,\mathrm{d}x\right)^{-1} M_h \left[|c_{ij}^r| \left\|(\bar{\lambda}_J^r)^{1/2}\right\|_1 \cdot \left\|(\bar{\lambda}_J^r)^{1/2} - (\hat{\lambda}_\infty^r)^{1/2}\right\|_1 \right. \\
&\qquad\qquad \left. + |c_{ij}^r - 1| \left\|(\bar{\lambda}_J^r)^{1/2}\right\|_1 \cdot \left\|(\hat{\lambda}_\infty^r)^{1/2}\right\|_1 + \left\|(\hat{\lambda}_\infty^r)^{1/2}\right\|_1 \cdot \left\|(\bar{\lambda}_J^r)^{1/2} - (\hat{\lambda}_\infty^r)^{1/2}\right\|_1 \right] \to 0.
\end{aligned}
$$

Provided that $\hat{\lambda}_J^r$ and $\hat{\lambda}_\infty^r$ are bounded, $\hat{\lambda}_J^{r+1}$ and $\hat{\lambda}_\infty^{r+1}$ are bounded, implying that $\int_{\mathcal{M}} (\hat{\lambda}_J^r)^{1/2} < \infty$ and $\int_{\mathcal{M}} (\hat{\lambda}_\infty^{r+1})^{1/2} < \infty$. It follows that (**A**) $\hat{\lambda}_J^r(x) \xrightarrow{\mathcal{L}^1} \hat{\lambda}_\infty^r(x)$, and (**B**) $\hat{\lambda}_J^r$, $\hat{\lambda}_\infty^r \in \mathcal{F}_2$ by induction.