

---

# A Robust Functional EM Algorithm for Incomplete Panel Count Data

---

**Alexander Moreno**  
Georgia Institute of Technology  
alexander.f.moreno@gatech.edu

**Zhenke Wu**  
University of Michigan  
zhenkewu@umich.edu

**Jamie Yap**  
University of Michigan  
jamiyap@umich.edu

**Cho Lam**  
University of Utah  
cho.lam@hci.utah.edu

**David W. Wetter**  
University of Utah  
david.wetter@hci.utah.edu

**Inbal Nahum-Shani**  
University of Michigan  
inbal@umich.edu

**Walter Dempsey**  
University of Michigan  
wdem@umich.edu

**James M. Rehg**  
Georgia Institute of Technology  
rehg@gatech.edu

## Abstract

Panel count data describes aggregated counts of recurrent events observed at discrete time points. To understand dynamics of health behaviors and predict future negative events, the field of quantitative behavioral research has evolved to increasingly rely upon panel count data collected via multiple self reports, for example, about frequencies of smoking using in-the-moment surveys on mobile devices. However, missing reports are common and present a major barrier to downstream statistical learning. As a first step, under a missing completely at random assumption (MCAR), we propose a simple yet widely applicable functional EM algorithm to estimate the counting process mean function, which is of central interest to behavioral scientists. The proposed approach wraps several popular panel count inference methods, seamlessly deals with incomplete counts and is robust to misspecification of the Poisson process assumption. Theoretical analysis of the proposed algorithm provides finite-sample guarantees by expanding parametric EM theory [3, 34] to the general non-parametric setting. We illustrate the utility of the proposed algorithm through numerical experiments and an analysis of smoking cessation data. We also discuss useful extensions to address deviations from the MCAR assumption and covariate effects.

## 1 Introduction

A major goal in behavioral medicine is identifying temporal patterns of risk factors preventing an individual from successfully modifying a health-related behavior. In smoking cessation, one would like to know times of day, locations, and other contextual factors such as smoking opportunity [17] that may precipitate lapse to inform interventions to prevent lapse [24]. A basic task is to describe when smoking occurs through modeling the *counting process* of repeated negative events. One goal is to estimate the mean function of the counting process to characterize the dynamics of health-related behaviors at the population level.

The counting process *mean function* can describe population-level temporal patterns of smoking. It converts the discrete patterns of smoking counts from a population of participants (see Fig. 1) into a temporally-continuous summary of smoking behavior (see Fig. 2b). However, the *missingness* inherent in EMA data makes consistent estimation of the mean function difficult. There are several

conditions causing missingness. An EMA may be ignored by the user or opened and then abandoned. Second, the mobile app itself may postpone the triggering of an EMA for any of several reasons (e.g. battery low). While EMAs are triggered randomly, the random process is constrained to have a minimum temporal spacing between EMAs, in order to keep participant burden at an acceptable level. If EMAs are postponed or ignored too many times, it will not be possible to trigger the full set of EMAs for the day, resulting in missing EMAs. Missing data is a common issue in studies that involve EMAs, with [16] noting that over 126 studies, the average missingness rate is 25%.

For mean function estimation, missing EMAs cause problems when they lead to inaccurate counts of the total cigarettes smoked between EMAs. Due to recall bias, longer intervals between EMAs are less reliable.<sup>1</sup> Behavioral scientists have developed heuristic imputation schemes to adjust for missing counts [13], which may not consistently estimate the mean function. We present the first self-contained and systematic treatment of missing data for panel count data, by providing a simple Expectation-Maximization (EM) algorithm [9, 3] to estimate the mean function with finite-sample theoretical guarantees. In doing so, we provide a framework for finite-sample theoretical guarantees for non-parametric EM more generally, potentially allowing finite sample theory for infinite dimensional missing data and latent variable models.

Our primary methodological contribution is a functional EM algorithm to wrap standard non-parametric mean function estimators to handle missing data under a missing completely at random (MCAR) assumption [18]. The E-step uses estimates from a fitting method to impute missing data, and the M-step calls that fitting method to estimate a mean function. This extends several classic non-parametric methods [31, 19] and the baseline-only version of a semi-parametric mean function estimation method [29] to the setting of missing data.

We analyze our EM algorithm using the frameworks described in [3, 34] and obtain finite sample guarantees. This requires care as we extend their work from the parametric to the non-parametric setting. This paper addresses three major theoretical challenges in the context of functional EM by: (i) noting that an infinite dimensional derivative in our setting is analogous to the inner products used in [34, 3], (ii) using a more general technique for showing local uniform strong concavity of our population E-step than previous work, which relied on quadratic dependence on parameters, and (iii) a high probability finite sample bound on the convergence of the M-step. The lack of ground truth for missing data in real-world settings makes evaluation difficult and theoretical guarantees important. Finally, our proposed algorithm can consistently estimate the mean function even when the Poisson process assumption is violated. This is achieved by recovering the population MLE, and noting that under certain integrability conditions the population MLE of the Poisson process log-likelihood is the true mean function of the counting process, even when the counting process is not Poisson.

We analyze two datasets empirically. We first look at a bladder tumor dataset where there are no missing intervals, and simulate missingness under missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) for various initializations and missingness probabilities. We find that recovery under MCAR tends to be very close to if we had complete data, as long as initialization is good. MAR also shows strong performance and MNAR shows relatively strong performance. We further analyze a recent smoking cessation dataset, where the system targets 3-4 prompts per day asking about smoking counts since the last assessment, but sometimes there are fewer responses. We treat self-reports over 24 hours as unreliable or missing data. We find that there is a difference of 21.7% in the final estimates of smoking counts of a 14 day study when using our EM algorithm vs believing their responses over these long intervals.

## 2 Related Work

**Ecological Momentary Assessments** are frequently used to record counts for behavior, including smoking counts [26, 28, 12], alcohol counts [8] and promiscuous behaviors [33]. However, none of these papers use panel count data methods to estimate the mean function.

**Panel count data** analysis comes from nonparametric statistics, but our missing data problem has not been addressed. The closest works to ours are [29, 31, 19]. These papers estimate the mean function

---

<sup>1</sup>In principle a participant who completes very few EMAs but reports their counts with 100% accuracy is not creating missing data, because standard mean function estimators [31] would still be asymptotically unbiased. In practice, however, behavioral scientists frequently treat EMAs as if they are missing if the interval to the last EMA exceeds a cut-off, such as 24 hours, due to recall bias.

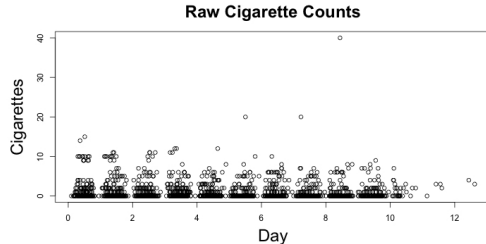


Figure 1: Raw cigarette counts smoked between observations. We want to convert this to a mean function of cumulative cigarettes smoked (Figure 2b), but some counts are missing or unreliable.

from panel count data, but cannot account for missing EMAs as in our motivating application. [29] develop an EM approach under a more limited missingness model assuming that the total counts for each participant are known prior to data analysis. This is unsuitable for analyzing data with missing EMAs. However, two strengths are that they can incorporate baseline covariate information, and they relax the standard assumption of independence between observation times and the counting process.

Neither of [31, 19] handle missing counts, and cannot be applied directly to our setting. We propose an EM algorithm for missing data to wrap any of [29, 31, 19], building on existing methods in the M-step. [31] provides the first strong theoretical guarantees for a mean function estimator in the panel count setting. [19] improve their rate of convergence with spline mean function estimators.

Many other papers in the literature focus on either extending to the semi-parametric setting (which includes baseline covariates), relaxing assumptions in either the non-parametric or semi-parametric setting, or improving estimation under specific distributional assumptions. They do not address missing data. [32] extended [31] to the semi-parametric setting, [15] analyzed panel count with informative observation times and subject-specific frailty, and [14] proposed using a smooth semi-parametric estimator to handle over-dispersion in panel count data. [10] proposed using a squared Gaussian process intensity function.

**Identification in Missing Data Models** a number of recent works [22, 4, 21, 23] analyze when the full law is identified in missing data models, including when recovery is possible under a missing not at random (MNAR) mechanism. However, only [21] discusses estimation, which is our focus. For the parametric non-temporal setting, they give a flexible method for estimation even under MNAR, as long as certain dependency conditions are satisfied by random variables. However, our setting is non-parametric and continuous time, and thus we cannot apply their method.

**EM Theory.** Also relevant is recent work on convergence of the EM algorithm to the true parameter [3, 34], which we extend to the non-parametric setting, proving convergence to the true *mean function*. In [3], they show that under good initialization and certain regularity conditions, population EM converges to the true parameter. With finite sample uniform convergence of the M-step, one can also show a finite-sample bound for EM. [34] show that by assuming the ability to optimize over a ball around the true parameter, one can weaken regularity conditions and replace uniform convergence of the M-step with three concentration inequalities, often easier to prove. We base our population guarantees on [34]. We base our finite-sample guarantees on [3], because [34] assumes that empirical and population norms are the same which does not hold in our non-parametric setting.

### 3 Model

#### 3.1 Complete Data

Let  $N = \{N(u) : u \geq 0\}$  be a univariate counting process. The goal is to estimate the mean function  $\Lambda^*(u) = \mathbb{E}[N(u)]$  over a study window  $[0, T_{\text{end}}]$ . Let  $K$  be the random number of observations for a participant. Let  $T = (T_1, \dots, T_K) \in \mathbb{R}_+^K$  be a random vector of observation times, with  $T_0 = 0$ . Let  $\Delta N = (\Delta N_1, \dots, \Delta N_K)$  be the count increments:  $\Delta N_j = N(T_j) - N(T_{j-1})$ . Let  $\Delta \Lambda(T_j) = \Lambda(T_j) - \Lambda(T_{j-1})$  be mean function increments. For each participant  $i = 1, \dots, n$ , we observe  $Y = (\Delta N, T, K) \in \mathcal{N} \times \mathcal{T} \times \mathcal{K}$  where  $\mathcal{N}, \mathcal{T}, \mathcal{K}$  are the corresponding sample spaces. Let  $P_n, P$  denote the empirical and true measures on  $\mathcal{N} \times \mathcal{T} \times \mathcal{K}$ , respectively. For a measurable function

$f$ , let  $P_n f = \frac{1}{n} \sum_{i=1}^n f(Y_i)$  and  $Pf = \int f dP$ . Under a non-homogeneous Poisson process, the complete data sample log-likelihood is

$$l_n(Y|\Lambda) = nP_n \left\{ \sum_{j=1}^K \Delta N_j \log[\Delta \Lambda(T_j)] - \Lambda(T_K) \right\},$$

and the population log-likelihood is similar but with  $P$  instead of  $nP_n$ . The goal is *consistent mean function estimation even when the Poisson process assumption is violated*.

### 3.2 Missing Data

Unlike previous work, we assume certain observations  $\Delta N_j$  are missing. For  $\Delta N \in \mathbb{R}^K$ , each observation  $\Delta N_j$  is missing completely at random with probability (w.p.)  $\epsilon \in [0, 1)$ . Let  $\tau \in \{0, 1\}^K$  be the missingness pattern, where  $0 = 0$  but the notation represents missingness. Let  $s = 1 - \tau$ , thus  $s_j = 1$  if  $\Delta N_j$  is missing. Let  $\Delta N^{(\tau)} = \Delta N \odot \tau$  and  $\Delta N^{(s)} = \Delta N \odot s$ , where  $\odot$  is elementwise product. Then  $\Delta N = \Delta N^{(s)} + \Delta N^{(\tau)}$ . We observe  $\Delta N^{(\tau)}$  but not  $\Delta N^{(s)}$ . Let  $Z \equiv \Delta N^{(s)}$  represent our missing data vector, and let  $\mathcal{Z}$  be the space of values for our missing data. Then  $Y = (\Delta N^{(\tau)}, T, K)$ , and  $(Y, Z) \in \mathcal{N} \times \mathcal{T} \times \mathcal{K} \times \mathcal{Z}$  gives us the observed and missing parts of the data, respectively. In this case  $P_n$  and  $P$  are now the empirical and true measures for  $\mathcal{N} \times \mathcal{T} \times \mathcal{K} \times \mathcal{Z}$ .

## 4 EM Algorithm

We first describe the general setting for EM. Let  $f_\Lambda(y, z)$ ,  $p_\Lambda(y)$ , and  $k_\Lambda(y|z)$  be joint, marginal, and conditional densities respectively. Let  $\Theta$  be a convex set of functions and  $\{\Theta_k\}$  a sieve: a nested sequence  $\Theta_1 \subset \Theta_2 \subset \dots$  such that  $\cup_{k=1}^\infty \Theta_k \subseteq \Theta$  is dense in  $\Theta$ . We define the following.

**Definition 1.** (Population  $Q$ -functional)  $Q(\Lambda'|\Lambda) \equiv \int_{\mathcal{Y}} \left( \int_{\mathcal{Z}(y)} \log(f_{\Lambda'}(y, z)) k_\Lambda(z|y) dz \right) p_{\Lambda^*}(y) dy$ .

**Definition 2.** (Sample  $Q$ -functional)  $Q_n(\Lambda'|\Lambda; \{Y_i\}_{i=1}^n) \equiv \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\Lambda[\log f_{\Lambda'}(y, z)|Y_i]$ .

**Definition 3.**  $B_r(\Lambda^*) \equiv \{\Lambda : \|\Lambda - \Lambda^*\|_\infty \leq r\}$  where  $\|\cdot\|_\infty$  is the essential supremum.

**Definition 4.** (Population  $M$ -step)  $M(\Lambda^{(t)}) = \arg \max_{\Lambda' \in \Theta \cap B_r(\Lambda^*)} Q(\Lambda'|\Lambda^{(t)})$ .

**Definition 5.** (Sample  $M$ -step)  $M_n(\Lambda^{(t)}) = \arg \max_{\Lambda' \in \Theta_k \cap B_r(\Lambda^*)} Q_n(\Lambda'|\Lambda^{(t)})$ .

A  $Q$ -functional is an  $E$ -step of the EM algorithm. Population EM repeatedly takes  $\Lambda^{(t+1)} = M(\Lambda^{(t)})$ , where  $\Lambda^{(t)}$  denotes iteration  $t$ 's estimate. Sample EM repeatedly takes  $\Lambda^{(t+1)} = M_n(\Lambda^{(t)})$ . Next we describe EM computation. Algorithm 1 describes sample EM; population EM is similar, but uses population  $Q$ -functionals and  $M$ -steps.

---

#### Algorithm 1 Sample-based EM Algorithm for Panel Count Data with Missing Counts.

---

- 1: Initialize  $\Lambda_0 \in \Theta_k \cap B_r(\Lambda^*)$ , let  $t \leftarrow 0$
  - 2: **while** not converged **do**
  - 3:     (E-step): Compute  $Q_n(\Lambda'|\Lambda^{(t)}; \{Y_i\}_{i=1}^n)$  using current mean function estimate
  - 4:     (M-step):  $\Lambda^{(t+1)} \leftarrow \arg \max_{\Lambda' \in \Theta_k \cap B_r(\Lambda^*)} Q_n(\Lambda'|\Lambda^{(t)}; \{Y_i\}_{i=1}^n)$  using existing method
  - 5:      $t \leftarrow t + 1$
  - 6: **end while**
- 

Algorithm 1 assumes  $B_r(\Lambda^*)$  is known following [34]: in practice it is unknown, and we suggest trying multiple initializations and choosing the one that leads to the highest final log-likelihood. In our numerical experiments and data analysis, for illustration we assume  $\Theta_k = \{\text{monotone step functions with at most } k \text{ steps}\}$ . We let  $k$  be the total number of intervals across all subjects. Next we describe the E and M-steps.

## 4.1 E-Step

The population and sample  $Q$ -functionals replace missing counts with mean function estimates. The sample  $Q$ -functional (the population version is similar but uses  $P$  instead of  $P_n$ ) is

$$Q_n(\Lambda'|\Lambda; \{Y_i\}_{i=1}^n) = P_n \left\{ \sum_{j=1}^K \Delta N_j^{T_j} \Delta \Lambda^{s_j}(T_j) \log[\Delta \Lambda'(T_j)] - \Lambda'(T_K) \right\}$$

## 4.2 M-Step

The population M-step maximizes  $Q(\Lambda'|\Lambda)$  using

$$\Theta = \{ \Lambda : [0, T_{\text{end}}] \rightarrow [0, \infty) \mid \Lambda \text{ is nondecreasing, } \Lambda(0) = 0, \Lambda(T_{\text{end}}) \leq U_{\text{all}} \},$$

where  $U_{\text{all}}$  is a uniform upper bound for functions in this set. The sample  $M$ -step uses a convex set  $\Theta_k \subset \Theta$  (sieve estimator), where  $k \rightarrow \infty$  as  $n \rightarrow \infty$ .  $\Theta_k$  is potentially monotonic step functions [31, 29] or monotonic splines [19], subject to the constraint of having the upper bound of  $U_{\text{all}}$  over the study time. Maximization proceeds via existing methods [31, 19, 29]. Like  $B_r(\Lambda^*)$ ,  $U_{\text{all}}$  is unknown.

## 5 Theory

Section 5.1 defines assumptions, 5.2 defines distances and a quasi-inner product, and 5.3 gives two regularity condition lemmas based on [34]. Proposition 1 shows that with good initialization and sufficiently small missingness probability, a population EM step gives a contraction, moving our estimate closer to the true mean function after every iteration. Theorem 1 then shows that population EM converges linearly to the true mean function.

We then show finite-sample theory. Proposition 2 states that with high probability, the sample M-step converges uniformly to the population M-step. Theorem 2 states that with high probability the distance between the current estimate and the true mean function is bounded by two terms: one describes applying population EM, and the other involves the uniform convergence of the  $M$ -step.

### 5.1 Assumptions

We make the following assumptions, similar assumptions were made in [31, 19, 29]:

1. The counting process  $N$  is independent of the number of observations  $K$  and observation times  $T$ , respectively;
2. The observation times are random variables taking values in the bounded set  $[\tau_0, T_{\text{end}}]$  where  $0 < \tau_0 < T_{\text{end}}$  and  $T_{\text{end}} \in (0, \infty)$ ;
3. The number of observations is bounded, i.e., there exists  $k_0 > 0$  such that  $P(K \leq k_0) = 1$ ;
4. The true mean function is uniformly bounded over the study, satisfying  $\Lambda^*(u) \leq U \leq U_{\text{all}}$  for some  $U \in (0, \infty)$  and all  $u \in [\tau_0, T_{\text{end}}]$ . Recall  $U_{\text{all}}$  is a uniform upper bound on functions in  $\Theta$ ;
5. The first derivative of  $\Lambda^*(u)$  has a positive lower bound in  $[\tau_0, T_{\text{end}}]$ .<sup>2</sup>
6. The observation times are  $\alpha$ -separated. That is,  $P(T_j - T_{j-1} \geq \alpha) = 1$  for some  $\alpha > 0$  and all  $j = 1, \dots, K$ ;
7.  $\mathbb{E}[\exp(aN(t))]$  is uniformly bounded for  $t \in [0, T_{\text{end}}]$  some constant  $a$ ;
8. The count increments  $\Delta N_j$  are missing completely at random (MCAR) w.p.  $\epsilon > 0$ . That is,  $s_{j_i}, j_i = 1, \dots, K_i, i = 1, \dots, n$  are iid Bernoulli( $\epsilon$ ) random variables.

In the context of an observational smoking study, these assumptions can be scientifically expressed: 1) EMA delivery times are independent of smoking times 2) EMAs are delivered at random times

<sup>2</sup>This is not strong, and doesn't imply at least a constant first derivative/superlinear function in general. Consider  $\Lambda^*(t) = t^{1/2}$  over  $[0, T_{\text{end}}]$ . The derivative  $\frac{1}{2\sqrt{t}} \geq \frac{1}{2\sqrt{T_{\text{end}}}}$ , but this function is sub-linear.

throughout the study 3) there is a maximum number of EMAs delivered over the study 5) there is a minimum smoking risk throughout the study 6) there is a minimum time between EMAs 7) a sufficient condition is that the number of cigarettes smoked is bounded over the study 8) the probability that an EMA is missed is  $\epsilon > 0$ . The only new assumption for panel count is assumption 8.

## 5.2 Measures, Distance Metrics, and Quasi-Inner Product

We next define measures for constructing distance metrics between mean functions. We then define a quasi-inner product that is used to show regularity conditions for population EM theory. Let  $B, B_1, B_2$  be the intersection of Borel sets in  $\mathbb{R}$  with  $[0, T_{\text{end}}]$ .

**Definition 6** (Measures for sets containing observation times).  $\mu(B) \equiv \mathbb{E} \left[ \sum_{j=1}^K 1_B(T_j) \right]$  and  $\mu_2(B_1 \times B_2) \equiv \mathbb{E} \left[ \sum_{j=1}^K 1_{B_1}(T_{j-1}) 1_{B_2}(T_j) \right]$ .

Then  $\mu(B) = \mathbb{E}|\{\text{observations in } B\}|$  and  $\mu_2(B_1 \times B_2) = \mathbb{E}|\{\text{one observation in } B_1, \text{ next in } B_2\}|$ .

**Definition 7** ( $d_2$  metric for mean functions).  $\|\Lambda_1 - \Lambda_2\| \equiv \left[ \int_0^{T_{\text{end}}} \int_0^{T_{\text{end}}} |(\Lambda_1(v) - \Lambda_1(u)) - (\Lambda_2(v) - \Lambda_2(u))|^2 d\mu_2(u, v) \right]^{1/2}$ .

This is the  $d_2$  metric of [31]. Under assumption 3, convergence in  $\|\cdot\|$  implies convergence in  $L^2(\mu)$ . We base our theory on convergence in this norm. Now define

$$\begin{aligned} \langle \nabla Q(\Lambda^{(l)}|\Lambda), \Lambda' \rangle &\equiv \lim_{\zeta \downarrow 0} \frac{Q(\Lambda^{(l)} + \zeta \Lambda'|\Lambda) - Q(\Lambda^{(l)}|\Lambda)}{\zeta} \\ &= P \left\{ \sum_{j=1}^K \left( \frac{\Delta N(T_j)^{\tau_j} \Delta \Lambda(T_j)^{s_j}}{\Delta \Lambda_j^l} - 1 \right) (\Delta \Lambda_j') \right\}. \end{aligned} \quad (1)$$

We do not claim that (1) is a valid inner product, but it is closely related to inner products in [34, 3], the  $Q$ -function's directional derivative in the direction of a parameter: (1) is the  $Q$ -functional's right Gateaux derivative in the direction of a mean function, analogous in function space. The connection to inner products from [3, 34] is key to extending EM theory to the non-parametric setting, allowing us to prove our Lemmas 1 and 2. We prove Equation (1) in A.1 in the supplementary material.

## 5.3 Population Theory

Before stating our main population EM convergence theorem, we define important constants and state two lemmas for the population  $Q$ -functional. Lemmas 1 and 2 mirror gradient stability and local uniform strong concavity from [34], respectively, but they studied  $Q$ -functions with quadratic dependence on parameters. Our objective function does not have quadratic dependence on the mean function, which introduces new challenges. However, we can decompose our  $Q$ -functional into a sum of two terms: one locally uniformly strongly concave, and one strictly concave. The sum is then locally uniformly strongly concave. Lemma 2 shows this lower bound holds, and is a key step that allows us to apply [34] to our setting. We then prove Proposition 1 which states that population EM steps contract, bringing estimates closer to the true mean function after each iteration.

**Definition 8.** Let  $c \equiv \inf \Delta \Lambda^* > 0$  be a uniform lower bound on increments of the true mean function. This exists by assumptions 5 and 6.

**Definition 9.** Let  $b \equiv \sup \{ \Delta \Lambda : \Lambda \in B_r(\Lambda^*) \} > 0$  be a uniform upper bound on increments of mean functions in  $B_r(\Lambda^*)$ . Note  $b \leq U_{\text{all}}$ , where  $U_{\text{all}}$  is a uniform upper bound on functions in  $\Theta$ .

**Definition 10.** Let  $\gamma = \frac{\epsilon}{c}$  and  $\nu = \frac{1-\epsilon}{3b}$

**Lemma 1.** (Gradient Stability) Assume assumptions 1, 2, 3, 4, 5, 6, and 8 hold. Then for  $\Lambda, \Lambda' \in \Theta$ ,

$$\langle \nabla Q(\Lambda^*|\Lambda) - \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle \leq \gamma \|\Lambda - \Lambda^*\| \|\Lambda' - \Lambda^*\|.$$

*Proof.* See A.2 of the supplementary material.  $\square$

**Lemma 2.** (Local Uniform Strong Concavity) Assume all assumptions hold. Then if  $r \leq \frac{\epsilon}{4}$  and  $\Lambda', \Lambda \in B_r(\Lambda^*)$ ,

$$Q(\Lambda^*|\Lambda) - Q(\Lambda'|\Lambda) + \langle \nabla Q(\Lambda^*|\Lambda), \Lambda' - \Lambda^* \rangle \geq \nu \|\Lambda' - \Lambda^*\|^2.$$

*Proof.* See A.3 of the supplementary material.  $\square$

One can think of this similarly to a second order Taylor expansion, where we take the functional of  $\Lambda'$  expanded at  $\Lambda^*$  conditional on  $\Lambda$ , and use that to derive an inequality.

**Proposition 1.** (*Population EM Contraction*) Assume all assumptions hold. If  $r \leq \frac{c}{4}$ ,  $\Lambda', \Lambda \in B_r(\Lambda^*)$  and  $Q(\Lambda'|\Lambda) \geq Q(\Lambda^*|\Lambda)$ , then

$$\|\Lambda' - \Lambda^*\| \leq \frac{\gamma}{\nu} \|\Lambda - \Lambda^*\|.$$

See section A.5 in the supplementary material for a detailed proof. This is similar to Proposition 3.2 in [34]. In order for this to give a contraction, we need  $\gamma < \nu$ , which holds if  $\epsilon < \frac{c}{3b+c}$ . Thus if the uniform lower bound on increments of the true mean function goes up or the uniform upper bound on increments in the ball  $B_r(\Lambda^*)$  goes down, we can tolerate a higher probability of missing data. For the smoking setting, these are if the uniform lower bound on mean number of cigarettes smoked between EMAs goes up and uniform upper bound on possible number of cigarettes smoked in the ball around the true mean function goes down.

**Theorem 1.** (*Population EM Convergence to True Mean Function*) Suppose the assumptions of the above hold and  $0 < \gamma < \nu$ . Take the EM sequence  $\Lambda^{(t+1)} = \arg \max_{\Lambda' \in \Theta \cap B_r(\Lambda^*)} Q(\Lambda'|\Lambda^{(t)})$  and  $\Lambda^0 \in B_r(\Lambda^*) \cap \Theta$ . Then

$$\|\Lambda^{(t)} - \Lambda^*\| \leq \left(\frac{\gamma}{\nu}\right)^t \|\Lambda^{(0)} - \Lambda^*\|.$$

*Proof.* By induction. It holds for  $t = 0$ . Assume it holds for  $t \geq 0$ . Then  $\Lambda^{(t+1)} \in B_r(\Lambda^*)$  and by assumption  $Q(\Lambda^{(t+1)}|\Lambda^{(t)}) \geq Q(\Lambda^*|\Lambda^{(t)})$ . Applying population contractivity and the induction assumption,

$$\begin{aligned} \|\Lambda^{t+1} - \Lambda^*\| &\leq \frac{\gamma}{\nu} \|\Lambda^t - \Lambda^*\| \\ &\leq \left(\frac{\gamma}{\nu}\right)^{t+1} \|\Lambda^0 - \Lambda^*\| \end{aligned}$$

$\square$

At each EM step, we move towards the true function by a multiplicative factor of  $\left(\frac{\gamma}{\nu}\right)$ . This is similar to Theorem 3.1 in [34].

## 5.4 Sample Theory

We next discuss convergence of the sample EM algorithm. We again require the initial estimate  $\Lambda^0$  and subsequent estimates  $\Lambda^{(t)}$  to be close, i.e., in the set  $B_r(\Lambda^*) \cap \Theta_k$ . The key additional assumption is that the sample size is large enough to satisfy certain conditions. For all  $n$  greater than this minimum sample size, we can guarantee approximate contraction of the sample EM algorithm, where the term that does not contract goes to 0 in large samples.

**Proposition 2.** Suppose all assumptions hold. Assume  $B_r(\Lambda^*)$  has radius  $r \leq \frac{c}{4}$ . For any  $L > 0$ , there exists  $u_L \xrightarrow{L \rightarrow \infty} 0$  such that w.p.  $1 - u_L$

$$\|M_n(\Lambda) - M(\Lambda)\| \leq 2^L n^{-1/3}.$$

*Proof.* See B.1 of the supplementary material. The term  $u_L$  is also defined there.  $\square$

This is the rate of convergence of an M-estimator, and generalizes the rate of convergence of [1] from maximum likelihood estimates to M-steps. We now state our main result.

**Theorem 2.** Suppose  $0 < r \leq \frac{c}{4}$  and  $0 < \gamma < \nu$  and all assumptions hold such that the population contractivity holds. Let  $\kappa = \frac{\gamma}{\nu}$ . Take the EM sequence  $\Lambda^{(t+1)} = \arg \max_{\Lambda' \in B_r(\Lambda^*) \cap \Theta_k} Q_n(\Lambda'|\Lambda^{(t)})$  and  $\Lambda^0 \in B_r(\Lambda^*) \cap \Theta_k$ . Then if the sample size is large enough that  $2^L n^{-1/3} \leq (1 - \kappa)r$ , then w.p. at least  $1 - u_L$

$$\|\Lambda^{(t)} - \Lambda^*\| \leq \kappa^t \|\Lambda^0 - \Lambda^*\| + \frac{1}{1 - \kappa} 2^L n^{-1/3}.$$

*Proof.* Note that if  $\forall \Lambda \in B_r(\Lambda^*) \cap \Theta_k$ ,  $\|M_n(\Lambda) - M(\Lambda)\| \leq 2^L n^{-1/3}$ , then we have  $\sup_{\Lambda \in B_r(\Lambda^*) \cap \Theta_k} \|M_n(\Lambda) - M(\Lambda)\| \leq 2^L n^{-1/3}$ . We claim for any  $t > 0$ ,

$$\|\Lambda^{(t+1)} - \Lambda^*\| \leq \kappa \|\Lambda^{(t)} - \Lambda^*\| + 2^L n^{-1/3}$$

We prove by induction. First, this holds for  $t = 1$ .

$$\begin{aligned} \|\Lambda^{(1)} - \Lambda^*\| &\leq \|M(\Lambda^{(0)}) - \Lambda^*\| + \|M_n(\Lambda^{(0)}) - M(\Lambda^{(0)})\| \\ &\leq \kappa \|\Lambda^{(0)} - \Lambda^*\| + 2^L n^{-1/3} \end{aligned}$$

Now assume holds true for  $t > 0$ . Then for  $t + 1$ ,

$$\begin{aligned} \|\Lambda^{(t+1)} - \Lambda^*\| &\leq \|M(\Lambda^{(t)}) - \Lambda^*\| + \|M_n(\Lambda^{(t)}) - M(\Lambda^{(t)})\| \\ &\leq \kappa \|\Lambda^{(t)} - \Lambda^*\| + 2^L n^{-1/3} \end{aligned}$$

Now iterating we have

$$\|\Lambda^{(t)} - \Lambda^*\| \leq \kappa^t \|\Lambda^{(0)} - \Lambda^*\| + \frac{1}{1 - \kappa} 2^L n^{-1/3}$$

□

The proof follows that of Theorem 5 of [3] closely. This immediately implies that our algorithm can recover the true population MLE in large samples. Further by [31], the population MLE is the true mean function under assumptions 1,4 and 7 even under Poisson process violations.

## 6 Experiments

Here we show numerical and real data results to illustrate the utility of the proposed functional EM algorithm. In the M-step, we can choose any reasonable mean function estimator. In our experiments, we use a general likelihood-based augmented estimating equation (AEE) method [29], which uses monotone step functions when obtaining the mean function estimate with complete data. [30, 7] provide software implementations of AEE and a wide variety of other panel count methods. First, we perform synthetic experiments that demonstrate accurate recovery of the true mean function (see Section C of the supplementary material). In particular, we also demonstrate good recovery of the true mean function using simulated data based on mixed Poisson processes (which violate the Poisson process assumption) hence confirming theoretical robustness of functional EM to Poisson process assumption. We also show a simulation where simple baselines (complete case analysis, median imputation, and last value carried forward) perform poorly. In the following, we focus on two experiments involving real data.

### 6.1 Real Data with Synthetic Missingness

We analyze blaTum (bladder tumor dataset) [5] with 85 patients and counts of tumors taken at appointment times. We artificially delete intervals completely at random with probability 0.2. We then initialize  $\Lambda^{(0)}$  by replacing the missing data with Poisson(1) random variables and fitting AEE. We bootstrap 1,000 times, and plot the sample mean of our learned mean functions under complete data. We also set counts to zero in intervals with missing counts, which biases the mean function estimates. Figure 2a compares inference from complete vs partially missing data using our wrapper vs zeroing out missing data. Our wrapper learns a model much closer to the complete data than its initialization or the zeroing model. Section D of the supplementary material has additional experiments: we investigate sensitivity to initialization and different missingness probabilities, replace AEE with other M-step methods, investigate missing at random (MAR) and missing not at random (MNAR), and show results for the baselines discussed above.

### 6.2 Smoking Cessation Study

We analyze data from an ongoing smoking cessation study in which the following EMA question was delivered randomly 3-4 times per day: "Since the last assessment, how many cigarettes did you smoke?" We have 125 participants tracked over 14 days, with 3-4 random EMAs targeted. The first



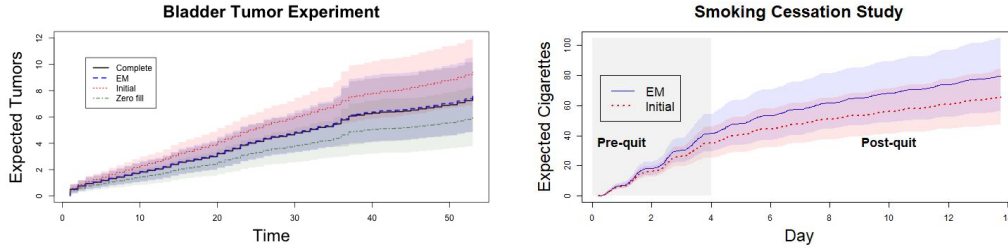


Figure 2: The estimated mean function and the 95% bootstrap confidence band from **a)** bladder tumor dataset: synthetic 20% missingness probability. Initialization has missing data set to Poisson(1) values. **b)** smoking cessation study. Initialization treats self-reported counts in intervals over 24 hours (about 5.6% of observations) as valid.

four days are their normal smoking behavior, while the remaining days involve them attempting to quit. In both curves the smoking rates are much higher pre-quit than post-quit, suggesting that they do in fact attempt to quit. After discussion with psychologists, we treat counts in intervals of longer than 24 hours as missing because they are considered to be unreliable. We use the counts to initialize our model. Unlike the previous experiments, we lack a ground truth. Note that this study is currently ongoing with more participants being added and we cannot share the dataset. However, a different subset of questions (as well as wearable sensors) was also used in [6], where they provide further description of the study.

Figure 2b shows the results based on 1,000 non-parametric bootstrap samples and two models: one treating long intervals as valid (“initial”), and the other treating them as missing where we use EM. Long intervals make up 5.6% of observations. The EM algorithm estimates that on average smokers attempting to quit smoked 80.4 cigarettes by the end of the study; in contrast, AEE underestimated (66.11), a difference of 21.7%. This is consistent with our collaborating behavioral scientists’ hypothesis that participants may under-report a cigarette count when the gap between completed EMAs is large. The proposed functional EM is able to borrow count information across multiple shorter overlapping intervals from other participants to alleviate this under-reporting. We discuss other analyses of the dataset further in section E of the supplementary material.

## 7 Discussion

This paper proposed a functional EM algorithm to estimate the mean function for incomplete panel count data. Extending existing EM algorithm analysis to general non-parametric settings, we provided finite sample convergence guarantees to the truth. We conducted extensive experiments to illustrate the effectiveness of the proposed algorithm in recovering the true mean function. We applied the proposed algorithm to a smoking cessation study and found that participants may underestimate their cigarettes smoked over intervals longer than 24 hours. To the best of our knowledge, we are the first to apply panel count data methods to EMAs, despite their central role in behavioral science research.

The main theoretical limitation is the MCAR assumption. Deviations from MCAR can happen if non-reporting depends on a subject’s emotional state, which may influence smoking counts. Extensions to other missingness mechanisms such as missing at random (MAR) warrant future research; our analysis provides a general framework and first step, which can eventually be extended to also incorporate covariates under a more complex missing data model. Like existing theoretical analysis of EM algorithms [34], another limitation is the need to optimize over a ball around the true mean function, which is unknown. Relaxing this condition is likely very challenging but also important future work. Finally, we would like test statistics for our estimator. This is challenging as we are not aware of any results showing that the difference between the true mean function and its estimator converges to a Gaussian process. However, there are test statistics based on weighting [2] that could be extended to this setting.

## Broader Impact

Understanding the dynamics for individuals who attempt to change and maintain behaviors to improve health has important societal value, for example, a comprehensive understanding of how smokers attempt to quit smoking may guide behavioral scientists to design better intervention strategies that tailor to the highest risk windows of relapse. Our theory and method provide an approach to understanding a particular aspect of the smoking behavior (mean function). The resulting algorithm is robust to Poisson process violations, readily adaptable and simple to implement, highlighting the potential for its wider adoption. The negative use case could be lack of sensitivity analysis around the assumptions such as missing data mechanism which may lead to misleading conclusions. Our current recommendation is to consult scientists about the plausibility of the assumption about missing data.

## 8 Funding and Other Acknowledgments

We acknowledge funding from the national institute of health (NIH) under awards U01CA229437, R01CA224537, R01MD010362, R01CA190329, R01 MH101459, R01 DA039901 as well as U54EB020404 (by NBIB) through funds provided by the Big Data to Knowledge (BD2K) initiative. We also acknowledge an investigator grant from Precision Health Initiative at the University of Michigan, Ann Arbor.

We thank Soujanya Chatterjee, Shahin Samiei, Timothy Hnat and Santosh Kumar at the University of Memphis for helpful discussions.

## References

- [1] BALAKRISHNAN, N., AND ZHAO, X. A class of multi-sample nonparametric tests for panel count data. *Annals of the Institute of Statistical Mathematics* 63, 1 (2011), 135–156.
- [2] BALAKRISHNAN, N., ZHAO, X., ET AL. New multi-sample nonparametric tests for panel count data. *The Annals of Statistics* 37, 3 (2009), 1112–1149.
- [3] BALAKRISHNAN, S., WAINWRIGHT, M. J., YU, B., ET AL. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics* 45, 1 (2017), 77–120.
- [4] BHATTACHARYA, R., NABI, R., SHPITSER, I., AND ROBINS, J. M. Identification in missing data models represented by directed acyclic graphs. In *Uncertainty in Artificial Intelligence* (2020), PMLR, pp. 1149–1158.
- [5] BYAR, D. The veterans administration study of chemoprophylaxis for recurrent stage i bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa. In *Bladder tumors and other topics in urological oncology*. Springer, 1980, pp. 363–370.
- [6] CHATTERJEE, S., MORENO, A., LIZOTTE, S. L., AKTHER, S., ERTIN, E., FAGUNDES, C. P., LAM, C., REHG, J. M., WAN, N., WETTER, D. W., ET AL. Smokingopp: Detecting the smoking’opportunity’context using mobile sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–26.
- [7] CHIOU, S. H., HUANG, C.-Y., XU, G., AND YAN, J. Semiparametric regression analysis of panel count data: A practical review. *International Statistical Review* 87, 1 (2019), 24–43.
- [8] COLLINS, R. L., KASHDAN, T. B., AND GOLLNISCH, G. The feasibility of using cellular phones to collect ecological momentary assessment data: Application to alcohol consumption. *Experimental and clinical psychopharmacology* 11, 1 (2003), 73.
- [9] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1 (1977), 1–22.
- [10] DING, H., LEE, Y., SATO, I., AND SUGIYAMA, M. Variational inference for gaussian process with panel count data. *Uncertainty in Artificial Intelligence* (2018).
- [11] GERW ([HTTPS://MATH.STACKEXCHANGE.COM/USERS/58577/GERW](https://math.stackexchange.com/users/58577/gerw)). Taylor expansion for gteaux derivative. Mathematics Stack Exchange. URL:<https://math.stackexchange.com/q/3076726> (version: 2019-01-17).

- [12] GRIFFITH, S. D., SHIFFMAN, S., AND HEITJAN, D. F. A method comparison study of timeline followback and ecological momentary assessment of daily cigarette consumption. *Nicotine & tobacco research* 11, 11 (2009), 1368–1373.
- [13] HALLGREN, K. A., AND WITKIEWITZ, K. Missing data in alcohol clinical trials: a comparison of methods. *Alcoholism: Clinical and Experimental Research* 37, 12 (2013), 2152–2160.
- [14] HUA, L., ZHANG, Y., AND TU, W. A spline-based semiparametric sieve likelihood method for over-dispersed panel count data. *Canadian Journal of Statistics* 42, 2 (2014), 217–245.
- [15] HUANG, C.-Y., WANG, M.-C., AND ZHANG, Y. Analysing panel count data with informative observation times. *Biometrika* 93, 4 (2006), 763–775.
- [16] JONES, A., REMMERSWAAL, D., VERVEER, I., ROBINSON, E., FRANKEN, I. H., WEN, C. K. F., AND FIELD, M. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. *Addiction* 114, 4 (2019), 609–619.
- [17] KIRCHNER, T. R., CANTRELL, J., ANESETTI-ROTHERMEL, A., GANZ, O., VALLONE, D. M., AND ABRAMS, D. B. Geospatial exposure to point-of-sale tobacco: real-time craving and smoking-cessation outcomes. *American journal of preventive medicine* 45, 4 (2013), 379–385.
- [18] LITTLE, R. J., AND RUBIN, D. B. *Statistical analysis with missing data*, vol. 793. John Wiley & Sons, 2019.
- [19] LU, M., ZHANG, Y., AND HUANG, J. Estimation of the mean function with panel count data using monotone polynomial splines. *Biometrika* 94, 3 (2007), 705–718.
- [20] MA, S., AND KOSOROK, M. R. Robust semiparametric m-estimation and the weighted bootstrap. *Journal of Multivariate Analysis* 96, 1 (2005), 190–217.
- [21] MALINSKY, D., SHPITSER, I., AND TCHETGEN, E. J. T. Semiparametric inference for non-monotone missing-not-at-random data: the no self-censoring model. *arXiv preprint arXiv:1909.01848* (2019).
- [22] MOHAN, K., PEARL, J., AND TIAN, J. Graphical models for inference with missing data. In *Advances in neural information processing systems* (2013), pp. 1277–1285.
- [23] NABI, R., BHATTACHARYA, R., AND SHPITSER, I. Full law identification in graphical models of missing data: Completeness results. *arXiv preprint arXiv:2004.04872* (2020).
- [24] REHG, J. M., MURPHY, S. A., AND KUMAR, S. *Mobile Health*. Springer, 2017.
- [25] SEN, B. A gentle introduction to empirical process theory and applications.
- [26] SHIFFMAN, S. How many cigarettes did you smoke? assessing cigarette consumption by global report, time-line follow-back, and ecological momentary assessment. *Health Psychology* 28, 5 (2009), 519.
- [27] VAN DER VAART, A. W., AND WELLNER, J. A. Weak convergence. In *Weak convergence and empirical processes*. Springer, 1996, pp. 16–28.
- [28] WANG, H., SHIFFMAN, S., GRIFFITH, S. D., AND HEITJAN, D. F. Truth and memory: Linking instantaneous and retrospective self-reported cigarette consumption. *The annals of applied statistics* 6, 4 (2012), 1689.
- [29] WANG, X., MA, S., AND YAN, J. Augmented estimating equations for semiparametric panel count regression with informative observation times and censoring time. *Statistica Sinica* (2013), 359–381.
- [30] WANG, X., AND YAN, J. Fitting semiparametric regressions for panel count survival data with an r package `spcf`. *Computer methods and programs in biomedicine* 104, 2 (2011), 278–285.
- [31] WELLNER, J. A., ZHANG, Y., ET AL. Two estimators of the mean of a counting process with panel count data. *The Annals of statistics* 28, 3 (2000), 779–814.
- [32] WELLNER, J. A., ZHANG, Y., ET AL. Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics* 35, 5 (2007), 2106–2142.
- [33] WRAY, T. B., KAHLER, C. W., AND MONTI, P. M. Using ecological momentary assessment (ema) to study sex events among very high-risk men who have sex with men (msm). 2231–2242.
- [34] WU, C., YANG, C., ZHAO, H., AND ZHU, J. On the convergence of the em algorithm: A data-adaptive analysis. *arXiv preprint arXiv:1611.00519* (2016).

## A Proofs Related to Theorem 1

### A.1 Proof of Equality in Equation 1

We start with the numerator

$$\begin{aligned}
Q(\Lambda^{(l)} + \zeta \Lambda' | \Lambda) - Q(\Lambda^{(l)} | \Lambda) &= P\left(\sum_{j=1}^K \Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log\left(\Delta \Lambda_j^{(l)} + \zeta \Delta \Lambda'_j\right) - (\Lambda'_K + \zeta \Lambda'_K)\right) \\
&\quad - P\left(\sum_{j=1}^K \Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log\left(\Delta \Lambda_j^{(l)}\right) - (\Lambda'_K)\right) \\
&= P\left(\sum_{j=1}^K \Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log \frac{\Delta \Lambda_j^{(l)} + \zeta \Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}} - \zeta \Lambda'_K\right)
\end{aligned}$$

Now consider

$$\begin{aligned}
\lim_{\zeta \downarrow 0} \frac{\log \frac{\Delta \Lambda_j^{(l)} + \zeta \Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}}}{\zeta} &= \lim_{\zeta \downarrow 0} \log \left( \frac{\Delta \Lambda_j^{(l)} + \zeta \Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}} \right)^{1/\zeta} \\
&= \lim_{\zeta \downarrow 0} \log \left( 1 + \zeta \frac{\Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}} \right)^{1/\zeta} \\
&= \log \exp \left( \frac{\Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}} \right) \\
&= \frac{\Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}}
\end{aligned}$$

We next need to show that we can pull the limit as  $\zeta \downarrow 0$  under integrals. There are two relevant terms:  $\frac{P(\sum_{j=1}^K \zeta \Lambda'_K)}{\zeta}$ , which we can trivially handle by pulling  $\zeta$  outside the integral, and

$$\frac{P(\sum_{j=1}^K \Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log \frac{\Delta \Lambda_j^{(l)} + \zeta \Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}})}{\zeta} = P\left(\sum_{j=1}^K \Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \frac{1}{\zeta} \log \frac{\Delta \Lambda_j^{(l)} + \zeta \Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}}\right)$$

Noting that  $\frac{1}{\zeta} \log \frac{\Delta \Lambda_j^{(l)} + \zeta \Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}}$  is monotone increasing to  $\frac{\Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}}$  as  $\zeta \downarrow 0$ , we can apply the monotone convergence theorem to pull the limit under the integral. Then

$$\begin{aligned}
\lim_{\zeta \downarrow 0} \frac{Q(\Lambda^{(l)} + \zeta \Lambda' | \Lambda) - Q(\Lambda^{(l)} | \Lambda)}{\zeta} &= P\left(\sum_{j=1}^K \Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \frac{\Delta \Lambda'_j}{\Delta \Lambda_j^{(l)}} - \Lambda'_K\right) \\
&= P\left(\sum_{j=1}^K \left(\frac{\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j}}{\Delta \Lambda_j^{(l)}} - 1\right) (\Delta \Lambda'_j)\right) \tag{2}
\end{aligned}$$

### A.2 Proof of Lemma 1

$$\begin{aligned}
\left| \sum_{j=1}^K s_j \left( \frac{\Delta \Lambda_j - \Delta \Lambda_j^*}{\Delta \Lambda_j^*} \right) (\Delta \Lambda'_j - \Delta \Lambda_j^*) \right| &\leq k_0 \frac{\sup |(\Delta \Lambda - \Delta \Lambda^*)(\Delta \Lambda' - \Delta \Lambda^*)|}{\inf \Delta \Lambda^*} \\
&= k_0 \frac{b^2}{c} \\
&< \infty
\end{aligned}$$

then any integral of this over  $\mathcal{T} \times \mathcal{K} \times \mathcal{Z}$  will also be finite, and we can apply Fubini's theorem to such integrals. Then recalling that  $\epsilon > 0$  is the MCAR probability (assumption 8),

$$\begin{aligned} \langle \nabla Q(\Lambda^*|\Lambda) - \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle &= P \left[ \sum_{j=1}^K s_j \left( \frac{\Delta\Lambda_j - \Delta\Lambda_j^*}{\Delta\Lambda_j^*} \right) (\Delta\Lambda'_j - \Delta\Lambda_j^*) \right] \\ &\leq \epsilon \left[ P \left( \sum_{j=1}^K \left( \frac{\Delta\Lambda_j - \Delta\Lambda_j^*}{\Delta\Lambda_j^*} \right)^2 \right) \right]^{1/2} \\ &\quad \times \left[ P \left( \sum_{j=1}^K (\Delta\Lambda'_j - \Delta\Lambda_j^*)^2 \right) \right]^{1/2} \end{aligned}$$

Here we used Fubini's theorem to pull out  $\epsilon$ . Applying the CS inequality for inner products in  $l^2$ , and finally applying the CS inequality for expectations gives us the result. The first term on rhs is then:

$$\left[ P \left( \sum_{j=1}^K \left( \frac{\Delta\Lambda_j - \Delta\Lambda_j^*}{\Delta\Lambda_j^*} \right)^2 \right) \right]^{1/2} \leq \frac{1}{c} \left[ P \left( \sum_{j=1}^K (\Delta\Lambda_j - \Delta\Lambda_j^*)^2 \right) \right]^{1/2}$$

This gives us

$$P \left[ \sum_{j=1}^K s_j \left( \frac{\Delta\Lambda_j - \Delta\Lambda_j^*}{\Delta\Lambda_j^*} \right) (\Delta\Lambda'_j - \Delta\Lambda_j^*) \right] \leq \frac{\epsilon}{c} \|\Lambda - \Lambda^*\| \|\Lambda' - \Lambda^*\|$$

and thus

$$\langle \nabla Q(\Lambda^*|\Lambda) - \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle \leq \frac{\epsilon}{c} \|\Lambda - \Lambda^*\| \|\Lambda' - \Lambda^*\|$$

### A.3 Proof of Lemma 2

Note that

$$\begin{aligned} Q(\Lambda'|\Lambda) &= P \left( \sum_{j=1}^K \Delta N_j^{\tau_j} \Delta\Lambda_j^{s_j} \log[\Delta\Lambda'_j] - \Lambda'(T_K) \right) \\ &= P \left( \sum_{j=1}^K \tau_j [\Delta N_j \log[\Delta\Lambda'_j] - \Lambda'(T_K)] \right) + P \left( \sum_{j=1}^K s_j [\Delta\Lambda_j \log[\Delta\Lambda'_j] - \Lambda'(T_K)] \right) \end{aligned}$$

Define

$$\begin{aligned} Q_1(\Lambda'|\Lambda) &= P \left( \sum_{j=1}^K \Delta N_j \log[\Delta\Lambda'_j] - \Lambda'(T_K) \right) \\ Q_2(\Lambda'|\Lambda) &= P \left( \sum_{j=1}^K \Delta\Lambda_j \log[\Delta\Lambda'_j] - \Lambda'(T_K) \right) \end{aligned}$$

and note that if  $\sum_{j=1}^K \tau_j [\Delta N_j \log[\Delta\Lambda'_j] - \Lambda'(T_j)]$  and  $\sum_{j=1}^K s_j [\Delta\Lambda_j \log[\Delta\Lambda'_j] - \Lambda'(T_j)]$  are integrable over  $\mathcal{N} \times \mathcal{T} \times \mathcal{K} \times \mathcal{Z}$  with respect to the measure  $P$ , we can apply Fubini's theorem to obtain

$$Q(\Lambda'|\Lambda) = (1 - \epsilon)Q_1(\Lambda'|\Lambda) + \epsilon Q_2(\Lambda'|\Lambda)$$

This proof takes place in four parts. We first show an inequality that allows us to characterize  $r$  for  $B_r(\Lambda^*)$ . Next we show that in this ball the integrability conditions above hold. We then show that  $Q_1$  and  $Q_2$  are both strictly concave. Finally we show that  $Q_1$  is strongly concave and thus  $Q$ , a positive linear combination of a strongly concave and a strictly concave function, is strongly concave.

### A.3.1 Characterizing the Radius of Contraction

**Claim 1.** Let  $h(x) = x(\log(x) - 1) + 1$ . For  $\|\Lambda^* - \Lambda'\|_\infty \leq \frac{c}{4}$

$$h\left(\frac{\Delta\Lambda_j^*}{\Delta\Lambda_j'}\right) \geq \frac{1}{3}\left(\frac{\Delta\Lambda_j^*}{\Delta\Lambda_j'} - 1\right)^2$$

*Proof.* Let  $\phi(y) = (1+y)(\log(1+y) - 1) + 1$ . Then we can Taylor expand

$$\phi(y) = \phi(0) + \phi'(0)y + \frac{\phi''(\xi)}{2}y^2$$

where  $\xi$  lies between 0 and  $y$ . Then noting that  $\phi(0) = \phi'(0) = 0$  and  $\phi''(y) = \frac{1}{1+y}$ ,

$$\begin{aligned} \phi(y) &= \frac{\phi''(\xi)}{2}y^2 \\ &= \frac{1}{2(1+\xi)}y^2 \\ &\geq \frac{1}{3}y^2 \end{aligned}$$

for  $y \in (-1, 1)$ . Letting  $h(x) = x(\log x - 1) + 1$  this implies that  $h(x) \geq \frac{1}{3}(x-1)^2$  for  $(x-1) \in (-1, 1)$  i.e.  $x \in (0, 2)$ . Let  $x = \frac{\Delta\Lambda_j^*}{\Delta\Lambda_j'}$ . Then we need  $\Delta\Lambda_j' \geq \frac{1}{2}\Delta\Lambda_j^*$  for  $h(x) \geq \frac{1}{3}(x-1)^2$  to hold. Note that if

$$|\Delta\Lambda_j^* - \Delta\Lambda_j'| \leq z\Delta\Lambda_j^* \quad (3)$$

Then

$$\begin{aligned} \Delta\Lambda_j^* - \Delta\Lambda_j' &\leq z\Delta\Lambda_j^* \\ (1-z)\Delta\Lambda_j^* &\leq \Delta\Lambda_j' \\ \Delta\Lambda_j^* &\leq \frac{1}{1-z}\Delta\Lambda_j' \end{aligned}$$

and if  $z = 0.5$  the desired result holds. For equation 3 to hold it suffices to have  $(\Delta\Lambda_j^* - \Delta\Lambda_j')^2 \leq z^2\Delta\Lambda_j^{*2} \leq z^2c^2$ . Now noting that  $(a-b)^2 \leq 2(a^2 + b^2)$ ,

$$\begin{aligned} (\Delta\Lambda_j^* - \Delta\Lambda_j')^2 &= ([\Lambda_j^* - \Lambda_j'] - [\Lambda_{j-1}^* - \Lambda_{j-1}'])^2 \\ &\leq 2([\Lambda_j^* - \Lambda_j']^2 + [\Lambda_{j-1}^* - \Lambda_{j-1}']^2) \\ &\leq 4\|\Lambda^* - \Lambda'\|_\infty^2 \end{aligned}$$

w.p. 1, and noting that we want  $z = 0.5$  a sufficient condition is

$$\begin{aligned} 4\|\Lambda^* - \Lambda'\|_\infty^2 &\leq z^2c^2 \\ \|\Lambda^* - \Lambda'\|_\infty &\leq \frac{c}{4} \end{aligned}$$

□

### A.4 Integrability Conditions

First note by assumption 7, all moments of  $N(T_{\text{end}})$  are uniformly bounded and thus  $\|N(T_{\text{end}})\|_\infty < \infty$ . Then for  $\Lambda, \Lambda' \in B_r(\Lambda^*)$  and by assumption 3,

$$P \left| \sum_{j=1}^K \tau_j [\Delta N_j \log[\Delta\Lambda_j'] - \Lambda'(T_j)] \right| \leq k_0 \|N(T_{\text{end}})\|_\infty \max(|\log \frac{c}{2}|, \log b) + k_0 b < \infty$$

$$P \left| \sum_{j=1}^K s_j [\Delta\Lambda_j \log[\Delta\Lambda_j'] - \Lambda'(T_j)] \right| \leq k_0 b \max(|\log \frac{c}{2}|, \log b) + k_0 b < \infty$$

#### A.4.1 Strict Concavity of $Q_1$ and $Q_2$

For  $Q_1$  we have

$$\begin{aligned}
Q_1\left(\frac{\Lambda_1 + \Lambda_2}{2} \mid \Lambda\right) &= P \left( \sum_{j=1}^K \Delta N_j \log \left[ \frac{\Delta \Lambda_{1,j} + \Delta \Lambda_{2,j}}{2} \right] - \frac{\Lambda_{1,K} + \Lambda_{2,K}}{2} \right) \\
&> P \left( \sum_{j=1}^K \Delta N_j \log [\sqrt{\Delta \Lambda_{1,j}} \sqrt{\Delta \Lambda_{2,j}}] - \frac{\Lambda_{1,K} + \Lambda_{2,K}}{2} \right) \text{ AM-GM inequality} \\
&= \frac{1}{2} Q_1(\Lambda_1 \mid \Lambda) + \frac{1}{2} Q_1(\Lambda_2 \mid \Lambda)
\end{aligned}$$

and for  $Q_2$  the same argument can be made.

#### A.4.2 Strong Concavity of $Q$

Now note that

$$\begin{aligned}
Q_1(\Lambda^* \mid \Lambda) - Q_1(\Lambda' \mid \Lambda) + \langle \nabla Q_1(\Lambda^* \mid \Lambda), \Lambda' - \Lambda^* \rangle \\
&= P \left( \sum_{j=1}^K \left[ \Delta N_j \log \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} - (\Delta \Lambda_j^* - \Delta \Lambda_j') \left( 1 - \left( \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} - 1 \right) \right) \right] \right) \\
&= P \left( \sum_{j=1}^K \left[ \Delta \Lambda_j^* \log \left[ \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} \right] - (\Delta \Lambda_j^* - \Delta \Lambda_j') \right] \right) \\
&= P \left( \sum_{j=1}^K \Delta \Lambda_j' \left( \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} \log \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} - \left( \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} - 1 \right) \right) \right) \\
&= P \left( \sum_{j=1}^K \Delta \Lambda_j' h \left( \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} \right) \right) \\
&\geq \frac{1}{3} P \left( \sum_{j=1}^K \Delta \Lambda_j' \left( \frac{\Delta \Lambda_j^*}{\Delta \Lambda_j'} - 1 \right)^2 \right) \\
&\geq \frac{1}{3b} \|\Lambda' - \Lambda^*\|^2 \\
&\geq \frac{1}{3b} \|\Lambda' - \Lambda^*\|^2
\end{aligned}$$

where the first inequality holds by claim 1. Now by strict concavity of  $Q_2$  as shown in A.4.1 we have

$$Q_2(\Lambda^* \mid \Lambda) - Q_2(\Lambda' \mid \Lambda) + \langle \nabla Q_2(\Lambda^* \mid \Lambda), \Lambda' - \Lambda^* \rangle \geq 0$$

Summing  $(1 - \epsilon)Q_1$  and  $\epsilon Q_2$  we obtain

$$Q(\Lambda^* \mid \Lambda) - Q(\Lambda' \mid \Lambda) + \langle \nabla Q(\Lambda^* \mid \Lambda), \Lambda' - \Lambda^* \rangle \geq \frac{(1 - \epsilon)}{3b} \|\Lambda' - \Lambda^*\|^2$$

#### A.5 Proof of Population Contractivity

We now state the main proof of population contractivity. Denote

$$V(\Lambda' \mid \Lambda) = Q(\Lambda' \mid \Lambda) - Q(\Lambda^* \mid \Lambda) - \langle \nabla Q(\Lambda^* \mid \Lambda), \Lambda' - \Lambda^* \rangle$$

then

$$\begin{aligned}
0 &\leq Q(\Lambda'|\Lambda) - Q(\Lambda^*|\Lambda) \\
&= V(\Lambda'|\Lambda) + \langle \nabla Q(\Lambda^*|\Lambda), \Lambda' - \Lambda^* \rangle \\
&= V(\Lambda'|\Lambda) + \langle \nabla Q(\Lambda^*|\Lambda) - \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle + \langle \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle \\
&\leq V(\Lambda'|\Lambda) + \langle \nabla Q(\Lambda^*|\Lambda) - \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle \text{ KKT conditions} \\
&\leq -\nu \|\Lambda' - \Lambda^*\|^2 + \gamma \|\Lambda - \Lambda^*\| \|\Lambda' - \Lambda^*\| \text{ technical Lemmas}
\end{aligned}$$

and rearranging terms and dividing both sides by  $\|\Lambda' - \Lambda^*\|$  gives the desired result. Note that we used  $\langle \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle \leq 0$ , which if  $\langle \cdot, \cdot \rangle$  were a valid inner product would be the KKT conditions. However since  $\langle \cdot, \cdot \rangle$  may not be a valid inner product, they must be checked specifically. [31] does it in the sample case for the true log-likelihood: it is easy to verify that it still holds in the population case for  $Q$ -functionals. Noting that  $\Lambda^*$  maximizes  $Q(\Lambda'|\Lambda^*)$ , we have that  $Q(\Lambda^* + \zeta(\Lambda' - \Lambda^*)|\Lambda^*) - Q(\Lambda^*|\Lambda^*) \leq 0$

$$\begin{aligned}
\langle \nabla Q(\Lambda^*|\Lambda^*), \Lambda' - \Lambda^* \rangle &= \lim_{\zeta \downarrow 0} \frac{Q(\Lambda^* + \zeta(\Lambda' - \Lambda^*)|\Lambda^*) - Q(\Lambda^*|\Lambda^*)}{\zeta} \\
&\leq 0
\end{aligned}$$

## B Proofs Related to Theorem 2

### B.1 Proof of Proposition 2

#### B.1.1 Definitions and Background from the Literature

Before proving the proposition, we restate several important results that we use from existing literature. We repeat two definitions and two theorems from [25], adjusted to our notation. Note that  $\eta$ -brackets are normally called  $\epsilon$ -brackets. However, since we have already used  $\epsilon$  to denote MCAR probabilities, we call them  $\eta$ -brackets.

**Definition 11.** ( *$\eta$ -bracket*) Let  $(\mathcal{F}, d)$  be a normed space of functions with distance metric  $d$  induced by some norm. Given two functions  $l(\cdot)$  and  $g(\cdot)$ , the bracket  $[l, g]$  is the set of all functions  $f \in \mathcal{F}$  with  $l(u) \leq f(u) \leq g(u) \forall u \in [0, T_{end}]$ . An  $\eta$ -bracket is a bracket  $[l, g]$  with  $d(l, g) < \eta$ .

**Definition 12.** (*Bracketing numbers*). The bracketing number  $N_{[]}(\eta, \mathcal{F}, L_2(P))$  is the minimum number of  $\eta$ -brackets needed to cover  $\mathcal{F}$  using  $L^2(P)$  distance.

**Definition 13.** (*Bracketing Integral*) The bracketing integral is defined as

$$J_{[]}(\delta, \mathcal{F}, L_2(P)) \equiv \int_0^\delta \sqrt{\log(N_{[]}(\eta, \mathcal{F} \cup \{0\}, L_2(P)))} d\eta$$

Note that since any non-empty set requires at least one bracket to cover it and  $\log(x+1) \leq 1 + \log(x)$  for  $x \geq 1$ ,

$$\int_0^\delta \sqrt{\log(N_{[]}(\eta, \mathcal{F} \cup \{0\}, L_2(P)))} \leq \int_0^\delta \sqrt{1 + \log(N_{[]}(\eta, \mathcal{F}, L_2(P)))}$$

the right hand side is sometimes used as the definition of the bracketing integral, but we use the left hand side, following [25]. We now restate a theorem from [25], with the notation heavily adapted to our setting for clarity. The theorem is otherwise the same.

**Theorem 3.** (*Theorem 5.1 in [25]*) Let  $(\Theta \cap B_r(\Lambda^*), \|\cdot\|)$  be a semi-metric space. Fix  $n \geq 1$ . Let  $\{Q_n(\Lambda'|\Lambda) : \Lambda' \in \Theta \cap B_r(\Lambda^*)\}$  be a stochastic process and  $\{Q(\Lambda'|\Lambda) : \Lambda' \in \Theta \cap B_r(\Lambda^*)\}$  be a deterministic process. Assume

$$Q(\Lambda'|\Lambda) - Q(M(\Lambda)|\Lambda) \leq -c_1 \|\Lambda' - M(\Lambda)\|^2 \quad (4)$$

for some  $c_1 > 0$ . We call this the separation condition. Further, let

$$U_n(\Lambda'|\Lambda) = Q_n(\Lambda'|\Lambda) - Q(\Lambda'|\Lambda)$$

and assume that there exists some function  $\phi_n(\cdot)$  satisfying the following three conditions



1. The following expected supremum condition holds

$$\mathbb{E} \left[ \sup_{\Lambda': \|\Lambda' - M(\Lambda)\| \leq \delta} \sqrt{n} |U_n(\Lambda'|\Lambda) - U_n(M(\Lambda)|\Lambda)| \right] \lesssim \phi_n(\delta) \quad (5)$$

2. there exists  $\alpha < 2$  so that

$$\phi_n(dx) \leq d^\alpha \phi_n(x) \forall d > 1, x > 0 \quad (6)$$

3. for the rate of convergence  $r_n$

$$\phi_n(r_n) \lesssim \sqrt{n} r_n^2 \quad (7)$$

as  $n$  varies.

Here  $\lesssim$  means  $\leq$  the right hand side times a constant. Then for every  $L > 0$ ,  $\|M_n(\Lambda) - M(\Lambda)\| \leq 2^L r_n$  with probability at least  $1 - u_L$ . Here  $u_L = \tilde{c} \sum_{j>M} 2^{j(\alpha-2)}$ , where  $\tilde{c}$  only depends on the constants in the separation condition and the expected supremum bound.

Note that this is essentially a special case of Theorem 3.2.5 of [27], but it uses expectations instead of outer expectations. It makes the stronger version of their assumptions and draws a stronger conclusion, giving a finite sample bound. The key in our setting will be to ensure that  $\tilde{c}$  does not vary across iterations, which requires that the constants for the separation condition and the expected supremum bound do not vary across iterations. Importantly, we cannot always apply this theorem in the general functional EM setting: it requires that the sample  $Q$ -functional at its maximizer over a Sieve is equal to the sample  $Q$ -functional at its maximizer over the full function space. This holds for the Poisson process log-likelihood for panel count data for a range of cases: for instance with step functions or splines where jumps/knots occur at observation times. It does not necessarily hold for arbitrary models. In the latter case we may be able to use Theorem 6.1 of [25], but this would require checking it carefully for each potential model. We also note

**Theorem 4.** (Theorem 4.12 of [25]) For any class  $\mathcal{F}$  of measurable functions  $f : \mathcal{X} \rightarrow \mathbb{R}$  such that  $Pf^2 < \delta^2$  and  $\|f\|_\infty \leq M$  for every  $f$ ,

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |\sqrt{n}(P_n - P)|] \leq \tilde{K} J_{\square}(\delta, \mathcal{F}, L_2(P)) \left( 1 + \frac{J_{\square}(\delta, \mathcal{F}, L_2(P))}{\delta^2 \sqrt{n}} M \right)$$

where  $\tilde{K} > 0$  is some constant.

Importantly,  $\tilde{K}$  is a *universal* constant and does not depend on  $\mathcal{F}$ . This was noted by [20]. A version of this theorem was originally Lemma 3.4.2 in [27].

### B.1.2 Outline

We follow [1], which proves the rate of convergence for the maximum pseudo-likelihood of a Poisson process objective function for panel count data: extending their proof to the expected complete data log-likelihood case is straightforward. However, we face the issue that we want a high probability uniform bound on the distance between sample and population M-steps *across* EM iterations, whereas they needed an asymptotic high probability rate of convergence for the pseudo MLE. This poses three challenges: 1) our objective function at each iteration is the expected log-likelihood rather than the log-likelihood. Thus we cannot prove that the separation condition holds using the same techniques. 2) the separation condition must hold always rather than only in a neighborhood of the optimum 3) we need to check that constants are the same across EM iterations. Our aim is to apply Theorem 3 and show that the constant  $\tilde{c}$  in  $u_L$  does not vary across EM iterations. Note that other than checking the separation condition, the majority of this proof simply repeats the proof strategy of [1] but fills in details of results they call to make sure that constants don't vary across iterations of EM.

This proof takes place in four parts. We first show that the separation condition holds. We then bound the expectation of the supremum of the magnitude of an empirical process. We next prove the two properties of the function involved in that bound to show the rate of convergence, and finally conclude by applying Theorem 3.

### B.1.3 Separation Condition

We first prove that the separation condition given by Equation 4 holds. This involves applying functional second order Taylor expansions to  $Q_1$  and  $Q_2$  and using the remainder terms to obtain quadratic lower bounds.

**Claim 2.** For any  $\Lambda, \Lambda' \in B_r(\Lambda^*)$ ,

$$Q(M(\Lambda)|\Lambda) - Q(\Lambda'|\Lambda) \geq \left( (1 - \epsilon) \frac{c}{b^2} + \epsilon \frac{c}{2b^2} \right) \|M(\Lambda) - \Lambda'\|^2$$

*Proof.* Consider the functional second order expansion of  $Q_1(\Lambda'|\Lambda)$  and  $Q_2(\Lambda'|\Lambda)$  at  $M(\Lambda)$ . Note that this is valid by considering  $\Lambda' = M(\Lambda) + \zeta(\Lambda' - M(\Lambda))$  for  $\zeta = 1$ . We can then consider  $g_1(\zeta) = Q(M(\Lambda) + \zeta(\Lambda' - M(\Lambda))|\Lambda)$  for fixed  $\Lambda, M(\Lambda), \Lambda'$  and similarly for  $g_2$  and do a second order Taylor expansion at  $\zeta = 0$  as (See [11] for further details)

$$g_1(1) = g_1(0) + g_1'(0) + \frac{1}{2}g_1''(\zeta)$$

for some  $\zeta_1 \in (0, 1)$ . We first assume that  $g_1$  is twice differentiable in the interval  $[0, 1)$ , and then show it rigorously. We then have

$$\begin{aligned} Q_1(M(\Lambda)|\Lambda) - Q_1(\Lambda'|\Lambda) &= -\langle \nabla Q_1(M(\Lambda)|\Lambda), \Lambda' - M(\Lambda) \rangle \\ &\quad - P \left( - \sum_{j=1}^K (\Delta M(\Lambda)_j - \Delta \Lambda'_j) \frac{\Delta N_j}{\Delta \xi_{1,j}^2} (\Delta M(\Lambda)_j - \Delta \Lambda'_j) \right) \\ &= -\langle \nabla Q_1(M(\Lambda)|\Lambda), \Lambda' - M(\Lambda) \rangle \\ &\quad + P \left( \sum_{j=1}^K (\Delta M(\Lambda)_j - \Delta \Lambda'_j) \frac{\Delta \Lambda_j^*}{\Delta \xi_{1,j}^2} (\Delta M(\Lambda)_j - \Delta \Lambda'_j) \right) \\ &\geq -\langle \nabla Q_1(M(\Lambda)|\Lambda), \Lambda' - M(\Lambda) \rangle + \frac{c}{b^2} \|M(\Lambda) - \Lambda'\|^2 \end{aligned}$$

here  $\xi_{1,j} = M(\Lambda) + \zeta_1(\Lambda' - M(\Lambda))$ . Similarly for  $g_2$  and  $Q_2$ ,

$$\begin{aligned} Q_2(M(\Lambda)|\Lambda) - Q_2(\Lambda'|\Lambda) &= -\langle \nabla Q_2(M(\Lambda)|\Lambda), \Lambda' - M(\Lambda) \rangle \\ &\quad - P \left( - \sum_{j=1}^K (\Delta M(\Lambda)_j - \Delta \Lambda'_j) \frac{\Delta \Lambda_j}{\Delta \xi_{2,j}^2} (\Delta M(\Lambda)_j - \Delta \Lambda'_j) \right) \\ &= -\langle \nabla Q_2(M(\Lambda)|\Lambda), \Lambda' - M(\Lambda) \rangle \\ &\quad + P \left( \sum_{j=1}^K (\Delta M(\Lambda)_j - \Delta \Lambda'_j) \frac{\Delta \Lambda_j}{\Delta \xi_{2,j}^2} (\Delta M(\Lambda) - \Delta \Lambda') \right) \\ &\geq -\langle \nabla Q_2(M(\Lambda)|\Lambda), \Lambda' - M(\Lambda) \rangle + \frac{c}{2b^2} \|M(\Lambda) - \Lambda'\|^2 \end{aligned}$$

where the last line follows since  $\Lambda \in B_r(\Lambda^*)$  so that  $\|\Lambda - \Lambda^*\|_\infty \leq \frac{c}{4}$  and thus  $\Delta \Lambda \geq \frac{1}{2} \Delta \Lambda^* \geq \frac{1}{2}c$  w.p. 1. Noting that  $\langle \nabla Q(M(\Lambda)|\Lambda), \Lambda' - M(\Lambda) \rangle \leq 0$  by the KKT conditions,

$$Q(M(\Lambda)|\Lambda) - Q(\Lambda'|\Lambda) \geq \left( (1 - \epsilon) \frac{c}{b^2} + \epsilon \frac{c}{2b^2} \right) \|M(\Lambda) - \Lambda'\|^2$$

Thus the separation condition holds since we optimized over  $\Theta \cap B_r(\Lambda^*)$ .

We now return to showing that  $g_1$  and  $g_2$  are twice differentiable on  $[0, 1)$ . Note that the derivatives of  $g_1$  and  $g_2$  at 0 are in fact the Gateaux derivatives of  $Q_1$  and  $Q_2$  in the direction of  $\Lambda' - M(\Lambda)$ . In A.1 of the supplementary material we previously showed the existence of right Gateaux derivatives. The change here is that we need to handle the fact that  $\zeta$  may approach 0 from the left now. We first show that  $g_1$  is differentiable at 0.

Recall

$$Q_1(M(\Lambda) + \zeta(\Lambda' - M(\Lambda))|\Lambda) = P\left(\sum_{j=1}^K \Delta N_j \log[\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)]\right) \\ - (M(\Lambda)(T_K) + \zeta(\Lambda'(T_K) - M(\Lambda)(T_K)))$$

so that

$$q'_1(0) = \lim_{\zeta \rightarrow 0} \frac{P(\sum_{j=1}^K \Delta N_j \log \frac{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}{\Delta M(\Lambda)_j}) - \zeta(\Lambda'(T_K) - M(\Lambda)(T_K))}{\zeta}$$

To show existence, we must show first that we can pull the derivative under the integral and second that we can differentiate the logarithm terms, which we can as long as the terms inside the logarithm are positive. For the first part, we show that we can bound the terms with log in the derivative, so that we can apply the bounded convergence theorem. The second part follows since in showing the first part we also show that the terms inside the logarithm are positive.

In order to apply the bounded convergence theorem we need to show that in a neighborhood of  $\zeta = 0$ , that

$$\frac{1}{\zeta} \log \frac{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}{\Delta M(\Lambda)_j}$$

exists and is bounded. Since  $1 - \frac{1}{x} \leq \log x \leq x - 1$  if  $x > 0$  (we will show that  $x > 0$  in this setting), we have

$$\log \frac{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}{\Delta M(\Lambda)_j} \leq \zeta \frac{\Delta \Lambda'_j - \Delta M(\Lambda)_j}{\Delta M(\Lambda)_j} \\ \log \frac{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}{\Delta M(\Lambda)_j} \geq \zeta \frac{\Delta \Lambda'_j - \Delta M(\Lambda)_j}{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}$$

so that

$$\left| \frac{1}{\zeta} \log \frac{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}{\Delta M(\Lambda)_j} \right| \leq \max \left( \left| \frac{\Delta \Lambda'_j - \Delta M(\Lambda)_j}{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)} \right|, \left| \frac{\Delta \Lambda'_j - \Delta M(\Lambda)_j}{\Delta M(\Lambda)_j} \right| \right)$$

Now  $\Delta M(\Lambda)_j \geq \frac{c}{2}$  and we previously showed that  $|\Delta \Lambda_j - \Delta \Lambda_j^*| \leq 2\|\Lambda - \Lambda^*\|_\infty \leq \frac{c}{2}$  (first inequality was proven in Claim 1, second by assumption), so that

$$|\Delta \Lambda'_j - \Delta M(\Lambda)_j| \leq |\Delta \Lambda'_j - \Delta \Lambda_j^*| + |\Delta M(\Lambda)_j - \Delta \Lambda_j^*| \\ \leq c$$

thus if we consider  $\zeta \in (-\frac{1}{4}, \frac{1}{4})$ ,

$$\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j) \geq \frac{c}{2} + \min(\zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j), -\zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)) \\ \geq \frac{c}{2} + \min(\zeta c, -\zeta c) \\ \geq \frac{c}{4}$$

giving us

$$\left| \frac{1}{\zeta} \log \frac{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}{\Delta M(\Lambda)_j} \right| \leq \max \left( \frac{c}{4}, \frac{c}{2} \right) \\ = 4$$

We also needed to show that  $\frac{\Delta M(\Lambda)_j + \zeta(\Delta \Lambda'_j - \Delta M(\Lambda)_j)}{\Delta M(\Lambda)_j} > 0$ . However, we did this by showing that for  $\zeta \in (-\frac{1}{4}, \frac{1}{4})$ , the numerator is positive. Thus  $g_1$  is differentiable at 0. Showing the existence of the second derivative is straightforward. For the derivative at points in  $(0, 1)$  all terms are non-negative and thus it is simple to show existence using similar methods. Essentially the same argument can be applied to  $g_2$  and  $Q_2$ .

□

### B.1.4 Bounding the Expectation of the Supremum of the Magnitude of the Empirical Process

Our aim in this section is to apply Theorem 4 and use the result to show that the expected supremum condition, Equation 5 in Theorem 3 holds. Let

$$\Theta_\delta \equiv \{\Lambda' : \|\Lambda' - M(\Lambda)\| \leq \delta, \Lambda' \in \Theta \cap B_r(\Lambda^*)\}$$

Let  $m_{\Lambda', \Lambda}(Y) \equiv \sum_{j=1}^K \Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log[\Delta \Lambda'_j] - \Lambda'(T_K)$  and

$$\mathcal{M}_\delta \equiv \{m_{\Lambda', \Lambda}(Y) - m_{M(\Lambda), \Lambda}(Y) : \Lambda' \in \Theta_\delta\}$$

This section proceeds as follows. We first show that for all  $f \in \mathcal{M}_\delta$ ,  $Pf^2 \leq c_2\delta^2$  for some constant  $c_2 > 0$  and  $\|f\|_\infty \leq c_3$  for some  $c_3 > 0$ . We next show a bound on the bracketing entropy in terms of the bracket size. We then use this to bound the bracketing integral using  $\delta^{1/2}$ . We combine all of this to bound the expectation of the supremum of interest. We must carefully note that relevant constants do not vary across iterations.

**Claim 3.** For  $\Lambda' \in \Theta_\delta$ ,  $P|m_{\Lambda', \Lambda}(Y) - m_{M(\Lambda), \Lambda}(Y)|^2 \leq c_2\delta^2$  for some  $c_2 > 0$  that does not depend on  $\Lambda'$  or  $\Lambda$ .

*Proof.*

$$\begin{aligned} & P(m_{\Lambda', \Lambda}(Y) - m_{M(\Lambda), \Lambda}(Y))^2 \\ &= P \left( \sum_{j=1}^K [\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log \frac{\Delta \Lambda'_j}{\Delta M(\Lambda)_j} - (\Delta \Lambda'_j - \Delta M(\Lambda)_j)] \right)^2 \\ &\leq k_0 P \left( \sum_{j=1}^K [\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log \frac{\Delta \Lambda'_j}{\Delta M(\Lambda)_j} - (\Delta \Lambda'_j - \Delta M(\Lambda)_j)]^2 \right) \\ &\quad \text{Cauchy Schwarz} \\ &\leq 2k_0 P \left( \sum_{j=1}^K [(\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j})^2 (\log \frac{\Delta \Lambda'_j}{\Delta M(\Lambda)_j})^2 + (\Delta \Lambda'_j - \Delta M(\Lambda)_j)^2] \right) \\ &\quad \text{since } (a - b)^2 \leq 2(a^2 + b^2) \\ &= 2k_0 \left[ P \left( \sum_{j=1}^K (\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j})^2 (\log \frac{\Delta \Lambda'_j}{\Delta M(\Lambda)_j})^2 \right) + \|\Lambda' - M(\Lambda)\|^2 \right] \quad (8) \end{aligned}$$

Then note

$$\begin{aligned} P \left( \sum_{j=1}^K (\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log \frac{\Delta \Lambda'_j}{\Delta M(\Lambda)_j})^2 \right) &\leq P \left( \sum_{j=1}^K (\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j})^2 \frac{(\Delta \Lambda'_j - \Delta M(\Lambda)_j)^2}{\min(\Delta \Lambda'_j, \Delta M(\Lambda)_j)^2} \right) \\ &\leq \frac{1}{4c^2} P \left( \sum_{j=1}^K (\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j})^2 (\Delta \Lambda'_j - \Delta M(\Lambda)_j)^2 \right) \\ &\leq \frac{\max(\|N(T_{\text{end}})\|_\infty^2, b^2)}{4c^2} P \left( \sum_{j=1}^K (\Delta \Lambda'_j - \Delta M(\Lambda)_j)^2 \right) \\ &\quad \text{assumption 7} \\ &= \frac{\max(\|N(T_{\text{end}})\|_\infty^2, b^2)}{4c^2} \|\Lambda' - M(\Lambda)\|^2 \end{aligned}$$

where the first line uses the inequality  $1 - \frac{1}{x} \leq \log(x) \leq x - 1$  for  $x > 0$  which implies  $\left(\log\left(\frac{x}{y}\right)\right)^2 \leq (x - y)^2 / \min(x, y)^2$ . Plugging this back into equation 8 we obtain

$$\begin{aligned} P(m_{\Lambda', \Lambda}(Y) - m_{M(\Lambda), \Lambda}(Y))^2 &\leq 2k_0 \left[ \frac{\max(\|N(T_{\text{end}})\|_\infty^2, b^2)}{4c^2} \|\Lambda' - M(\Lambda)\|^2 + \|\Lambda' - M(\Lambda)\|^2 \right] \\ &= \left( 2k_0 \frac{\max(\|N(T_{\text{end}})\|_\infty^2, b^2)}{4c^2} + 1 \right) \|\Lambda' - M(\Lambda)\|^2 \\ &\leq \left( 2k_0 \frac{\max(\|N(T_{\text{end}})\|_\infty^2, b^2)}{4c^2} + 1 \right) \delta^2 \end{aligned}$$

so that we have  $P|m_{\Lambda', \Lambda}(Y) - m_{M(\Lambda), \Lambda}(Y)|^2 \leq c_2 \delta^2$  for some constant  $c_2$ , and  $c_2$  does not depend on either  $\Lambda'$  or  $\Lambda$ .  $\square$

**Claim 4.** For  $\Lambda' \in \Theta_\delta$ ,  $\|m_{\Lambda', \Lambda}(Y) - m_{M(\Lambda), \Lambda}(Y)\|_\infty \leq c_3$ , where  $c_3$  does not depend on  $\Lambda$  or  $\Lambda'$ .

*Proof.* Again using  $\left|\log\left(\frac{x}{y}\right)\right| \leq |x - y| / \min(x, y)$ , we have

$$\begin{aligned} |m_{\Lambda', \Lambda}(Y) - m_{M(\Lambda), \Lambda}(Y)| &= \left| \sum_{j=1}^K [\Delta N_j^{\tau_j} \Delta \Lambda_j^{s_j} \log \frac{\Delta \Lambda'_j}{\Delta M(\Lambda)_j} - (\Delta \Lambda'_j - \Delta M(\Lambda)_j)] \right| \\ &\leq \left| \sum_{j=1}^K \max(\|N(T_{\text{end}})\|_\infty, b) \frac{\Delta \Lambda'_j - \Delta M(\Lambda)_j}{\min(\Delta \Lambda'_j, \Delta M(\Lambda)_j)} \right| \\ &\quad + \left| \sum_{j=1}^K (\Delta \Lambda'_j - \Delta M(\Lambda)_j) \right| \\ &\leq k_0 \frac{\max(\|N(T_{\text{end}})\|_\infty, b)}{2c} [2\|\Lambda' - M(\Lambda)\|_\infty] + 2k_0 \|\Lambda' - \Lambda^*\|_\infty \\ &\leq k_0 \frac{\max(\|N(T_{\text{end}})\|_\infty, b)}{2c} 2[\|\Lambda' - \Lambda^*\|_\infty \\ &\quad + \|M(\Lambda) - \Lambda^*\|_\infty] + 2k_0 \|\Lambda' - \Lambda^*\|_\infty \\ &\leq k_0 \frac{\max(\|N(T_{\text{end}})\|_\infty, b)}{2} + k_0 \frac{c}{2} \\ &= c_3 \end{aligned}$$

where we used that  $\Lambda', \Lambda^*, M(\Lambda) \in B_r(\Lambda^*)$  an  $L^\infty$  ball with  $r = \frac{c}{4}$   $\square$

By theorem 2.7.5 of [27], which bounds the bracketing number of monotone functions mapping to  $[0, 1]$ , and noting that  $\mathcal{M}_\delta$  has bracketing number less than or equal to that of  $\Theta \cap B_r(\Lambda^*)$ , which was shown in [2],

$$\log N_{[]}(\eta, \mathcal{M}_\delta, L_2(P)) \leq c_4 \eta^{-1}$$

where  $c_4$  only depends on  $U_{\text{all}}$ , the uniform upper bound in  $\Theta$ . Noting that  $\mathcal{M}_\delta \cup \{0\} = \mathcal{M}_\delta$ , we have

$$\int_0^\delta \sqrt{\log N_{[]}(\eta, \mathcal{M}_\delta \cup \{0\}, L_2(P))} d\eta \leq c_5 \delta^{1/2}$$

where again  $c_5$  only depends on  $U_{\text{all}}$ . Let

$$\|\sqrt{n}(P_n - P)\|_{M_\delta} = \sup_{f \in M_\delta} |\sqrt{n}(P_n - P)f|$$

and note that

$$\mathbb{E} \|\sqrt{n}(P_n - P)\|_{M_\delta} = \mathbb{E} \left[ \sup_{\Lambda': \|\Lambda' - M(\Lambda)\| \leq \delta} \sqrt{n} |U_n(\Lambda' | \Lambda) - U_n(M(\Lambda) | \Lambda)| \right]$$

**Claim 5.** Because  $P|m_{\Lambda',\Lambda}(Y) - m_{M(\Lambda),\Lambda}(Y)|^2 \leq c_2\delta^2$  and  $\|m_{\Lambda',\Lambda}(Y) - m_{M(\Lambda),\Lambda}(Y)\|_\infty \leq c_3$ , we have

$$\mathbb{E}\|\sqrt{n}(P_n - P)\|_{\mathcal{M}_\delta} \leq c_6\phi_n(\delta)$$

for  $\phi_n(\delta) = \delta^{1/2} + \delta^{-1}n^{-1/2}$ , where  $c_6$  does not depend on  $\Lambda'$  or  $\Lambda$ .

*Proof.* Here we use Theorem 4 in order to prove Equation 5 in Theorem 3. By Theorem 4 and again noting that  $\mathcal{M}_\delta \cup \{0\} = \mathcal{M}_\delta$ , we have

$$\begin{aligned} & \mathbb{E}\|\sqrt{n}(P_n - P)\|_{\mathcal{M}_\delta} \\ & \leq \tilde{K} \left( \frac{(\int_0^{\sqrt{c_2}\delta} \sqrt{\log N_{[]}(\eta, \mathcal{M}_\delta, L_2(P))} d\eta)^2 c_3}{c_2\delta^2\sqrt{n}} + \int_0^{\sqrt{c_2}\delta} \sqrt{\log N_{[]}(\eta, \mathcal{M}_\delta, L_2(P))} d\eta \right) \\ & \leq \tilde{K} \left( \frac{c_5^2(c_2^{1/2})c_3\delta}{c_2\delta^2\sqrt{n}} + c_5c_2^{1/4}\delta^{1/2} \right) \\ & = \tilde{K} \left( \frac{c_5^2c_3}{c_2^{1/2}\delta\sqrt{n}} + c_5c_2^{1/4}\delta^{1/2} \right) \\ & \leq \tilde{K} \max \left( \frac{c_5^2c_3}{c_2^{1/2}}, c_5c_2^{1/4} \right) \left( \frac{1}{\delta\sqrt{n}} + \delta^{1/2} \right) \end{aligned}$$

Set  $c_6 = \tilde{K} \max \left( \frac{c_5^2c_3}{c_2^{1/2}}, c_5c_2^{1/4} \right)$  and we are done.  $\square$

With this we have the bound on the expected supremum and have proven Equation 5 for Theorem 3.

### B.1.5 Characterizing the Function in the Bound

We first prove Equation 6 in Theorem 3.

$$\begin{aligned} \phi_n(d\delta) &= (d\delta)^{1/2} + \frac{1}{(d\delta)\sqrt{n}} \\ &= \sqrt{d} \left( \delta^{1/2} + \frac{1}{d^{3/2}\delta\sqrt{n}} \right) \\ &\leq \sqrt{d} \left( \delta^{1/2} + \frac{1}{\delta\sqrt{n}} \right) \text{ for } d \geq 1 \\ &= \sqrt{d}\phi_n(\delta) \end{aligned}$$

and thus  $\alpha = \frac{1}{2}$ .

Next we prove Equation 7. Let  $r_n = n^{-1/3}$ . Then

$$\phi_n(r_n) = 2n^{-1/6}$$

and

$$\begin{aligned} \sqrt{nr_n^2} &= \sqrt{n}(n^{-1/3})^2 \\ &= n^{-1/6} \end{aligned}$$

and thus

$$\phi_n(r_n) \leq 2\sqrt{nr_n^2}$$

### B.1.6 Putting it All Together

We have now proven that all of the assumptions of Theorem 3 hold, and that constants do not vary across iterations. Then for some constant  $\tilde{c}$  which does not vary across iterations, for any  $L > 0$  and using  $\alpha = \frac{1}{2}$ , w.p.  $1 - \tilde{c} \sum_{j>M} 2^{-3j/2}$ ,

$$\|M_n(\Lambda) - M(\Lambda)\| \leq 2^L n^{-1/3}$$

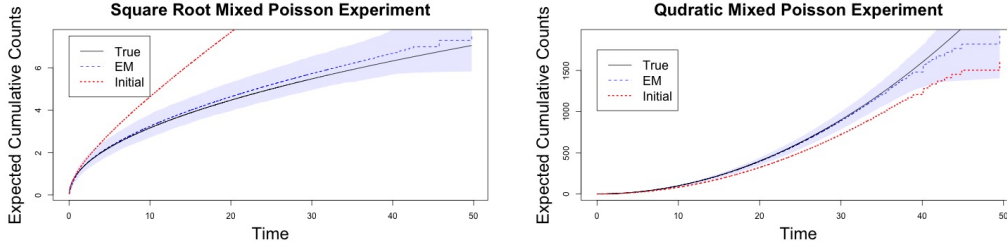


Figure 3: Mean+95% Monte Carlo CIs of 1000 runs of Mixed Poisson processes, which violate the Poisson process assumption. Synthetic datasets: 20% missingness using AEE wrapped with our method. Black/solid is the true mean function. Red/dotted is initialization with missing data set to Poisson(1) values. Blue/dashed is our EM algorithm initialized at red. (a)  $\Lambda^*(u) = \sqrt{u}$  a square root mean function (b)  $\Lambda^*(u) = u^2$  a quadratic mean function

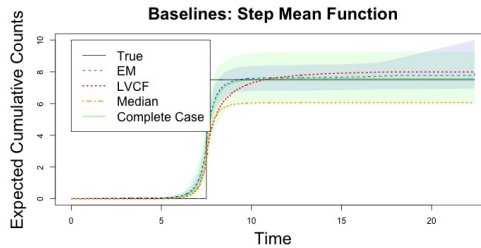


Figure 4: Comparison of our EM algorithm to last value carried forward (LVCF), median imputation, and complete case analysis, all under 20% missingness for estimating a step function. We see that last value carried forward performs poorly after a jump and median imputation is a poor method. We show confidence intervals for EM and complete case analysis: the latter has very high variance.

## C Synthetic Analysis

### C.1 Square Root Synthetic Experiment

We generate synthetic panel count data from mixed inhomogeneous Poisson processes with conditional mean functions  $\Lambda^*(u)|X = Xu^{1/2}$ ,  $\Lambda^*(u)|X = Xu^2$ , where  $X \sim \text{uniform}(0, 2)$ . The mean functions are then  $\Lambda^*(u) = u^{1/2}$  and  $\Lambda^*(u) = u^2$ , respectively. The counting process conditional on  $X$  is Poisson, but the marginal counting process is not. We use 100 trajectories, each with 30 observations and for each observation set it to missing with probability 0.2. We initialize the mean function  $\Lambda^{(0)}$  by replacing the missing data with Poisson(1) random variables and fitting a model. We generate 1000 Monte Carlo runs and create Monte Carlo marginal confidence intervals from those runs.

Fig 3 compares the true mean function against AEE wrapped with our method vs AEE directly on the corrupted data. Taking the corrupted data as given learns highly biased results, while wrapping AEE with our algorithm learns close to the true mean function for both experiments.

### C.2 A Step Function Where Two Baselines Perform Poorly

Here we show that three simple baselines perform poorly in estimating a scaled indicator step function. We use last value carried forward, median imputation, and complete case analysis (where we delete each participant with missing data), all under MCAR. Figure 4 shows the results. We see that complete case has very high variance, median is very biased, and LVCF performs poorly after a discontinuity.

## D Further Analysis of Bladder Tumor Dataset

### D.1 Comment on Mean and Confidence Intervals

In order to form the mean and marginal confidence intervals, we can note that the observation times are all discrete valued at times (months) 1 to 50. Thus we can simply take mean and marginal confidence intervals at those points. To do so, for each bootstrap replicate, we add a small amount of noise  $\pm 1e - 6$  to the observed time points for identifiability purposes in order to train, and then take mean and marginal confidence intervals at the described points. While last value carried forward is often a surprisingly competitive baseline, it performs poorly after discontinuities, as we see here. Median imputation tends to be highly biased, and complete case analysis tends to have high variance.

### D.2 Varying Missingness Probabilities

In this experiment we vary the missingness probability with  $\epsilon = 0.1, 0.2, 0.3, 0.4$ . We do so for each of the five following methods: the non-parametric step function maximum pseudo-likelihood (NPMPLE) of [31], the smoothed maximum pseudo-likelihood (MPLs) and maximum likelihood (MLs) estimators of [19], and the step function solution to the augmented estimating equation (AEE) and the informative censoring (AEEX) version methods of [29]. We show that bias is low but increasing as a function of the missingness probability for the five methods. Figure 5 shows results. The bias is very low in all cases. Further, in all cases it is much lower than the initialization with corrupted data shown in Figure 2a.

### D.3 Varying Initialization

In this experiment we vary the initialization. We note that the mean count across all observations is 0.44. We investigate Poisson(1) to Poisson(4) initializations under both  $\epsilon = 0.2$  and  $\epsilon = 0.4$ . We use AEE in all cases. Figure 6 shows results. We see that further initializations increase bias, but only slightly for  $\epsilon = 0.2$  and more so for  $\epsilon = 0.4$ . This is not surprising as our theory requires good initialization and sufficiently low missingness probability.

### D.4 Heterogeneity in Missingness Probabilities

Even if the MCAR assumption does hold, the missingness probability may vary between subjects. In this experiment, we let  $\epsilon|X = \epsilon_{\text{mean}}X$  where  $X \sim \text{uniform}(0, 2)$  and  $\epsilon_{\text{mean}} = 0.2, 0.4$ . This can capture between subject heterogeneity in missingness. Within each subject, we compare initialization of Poisson(1) to Poisson(4). Figure 7 shows results using the AEE method. They look very similar to results without heterogeneity in the missingness probabilities between subjects.

### D.5 Missing at Random

Here we investigate the missing at random (MAR) setting. We compare full data, last value carried forward, median imputation, complete case analysis, EM under MCAR, and EM under MAR. The baselines are under an MCAR assumption. We mimic MAR by artificially masking counts with probability  $\epsilon$  conditional on the previous count being zero ( $\epsilon = 0.3, 0.4$ ) or the previous count being non-zero ( $\epsilon = 0.2$ ). Figure 2a) has results, which are generally good and only have a slightly higher bias than the MCAR setting.

### D.6 Missing Not at Random

Here we investigate the missing not at random (MNAR) setting. This is similar to the previous but now we mask counts base on the *current* count being zero or non-zero. Figure 2b) has results. Again, the baselines are under an MCAR assumption. An increase in the zero missingness probability of 1% leads to an approximately 0.28% increase in the final estimate of expected tumors over MCAR: this is still useful for moderate differences in conditional missingness probabilities. For large deviations in the difference in missingness probability, further work is needed.



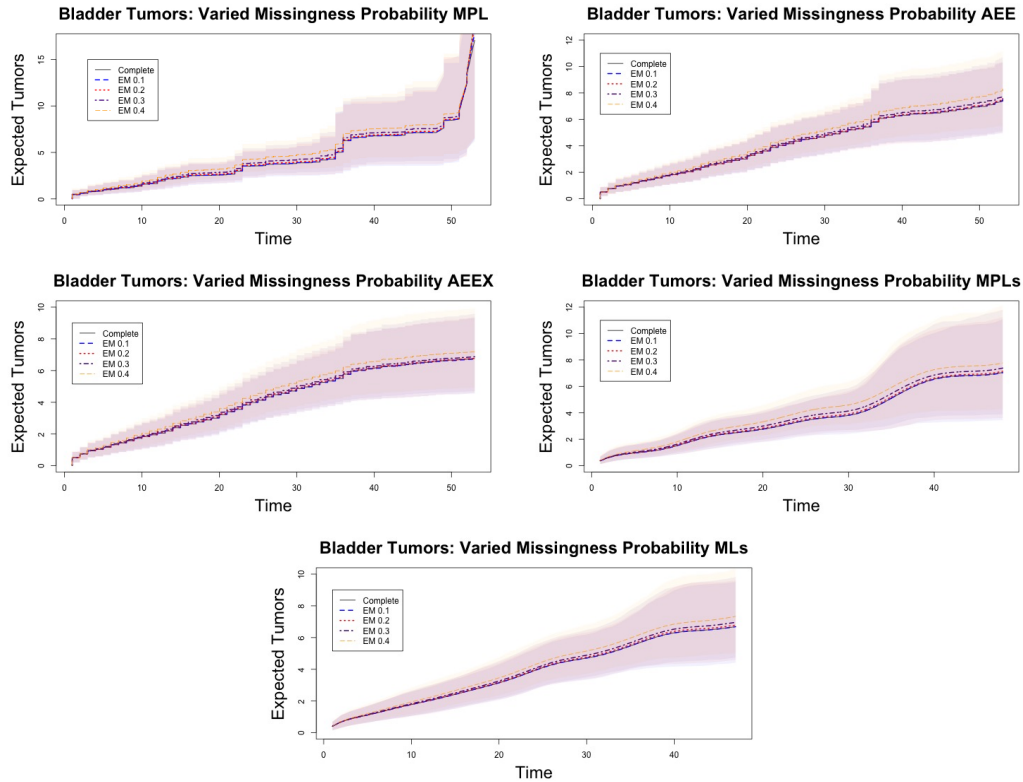


Figure 5: results for mean and 95% CIs for 1000 bootstrap replicates for bladder tumor dataset, with the missingness probability  $\epsilon$  set to  $\epsilon = 0.1, 0.2, 0.3, 0.4$ . We see that bias is very low in all settings, although it increases with the missingness probability. The methods wrapped with our method are (a) maximum pseudo-likelihood (MPL) of [31] (b) augmented estimating equations (AEE) of [29] (c) augmented estimating equations with informative censoring (AEEX) of [29] (d) maximum pseudo-likelihood splines (MPLs) of [19] note that for this we need to reduce the time axis slightly as spline models cannot interpolate far past the region where they have values in learning (e) maximum likelihood splines (MLs) of [19], same issue.

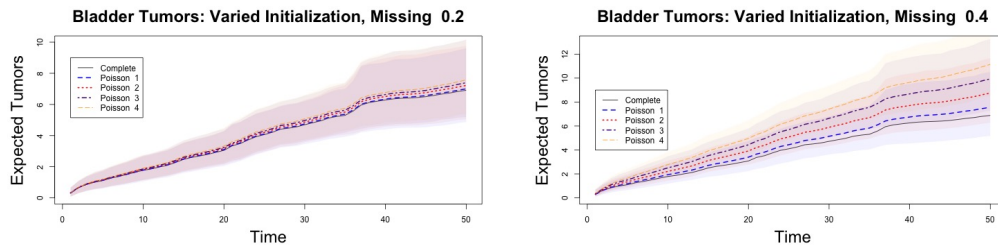


Figure 6: Here we vary the initialization by initializing increasingly far from the mean observed under complete data. The true sample mean of all intervals is 0.44. We initialize to Poisson random variables with means 1 to 4. We see that initializing further from the truth increases the bias, and that it is worse for the higher missingness probability of 0.4. Our theory requires good initialization and the missingness probability to be sufficiently low, so this is not surprising.

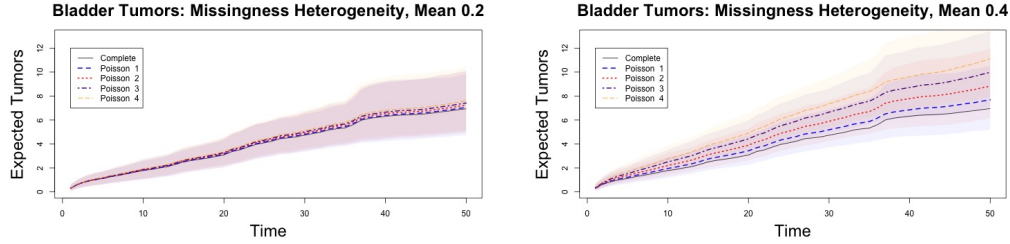


Figure 7: Here we let the missingness probability vary per participant while still being MCAR by multiplying the mean missingness by some value drawn from uniform $[0, 2]$ . We again initialize to Poisson(1) to Poisson(4) for a) mean missingness 0.2 and b) mean missingness 0.4.

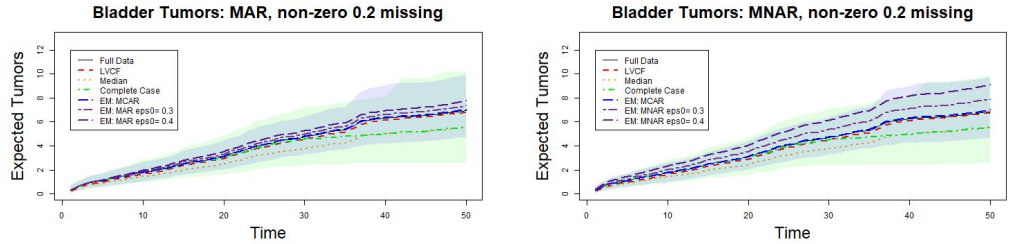


Figure 8: a) Missing at random (MAR). The time-varying missingness probability is  $\epsilon_j = 0.3, 0.4$  if the previous  $\Delta N_{j-1} = 0$  (approximately 84% of observations) and  $\epsilon_j = 0.2$  otherwise. b) Missing not at random (MNAR): same but instead of using the lagged term  $\Delta N_{j-1} = 0$  we use the current  $\Delta N_j = 0$ . We show confidence intervals for EM and complete case only in order to illustrate the high variance of complete case analysis. Note that for all of these the LVCF, median, and complete case analysis are under an MCAR missingness mechanism.

## E Further Exploratory Analysis of the Smoking Cessation Study and Dataset

In this section we plot several histograms and boxplots to get insight into behavior variability in the study. We plot cigarettes since the last assessment and number of days between EMAs, where for both we aggregate both between and within subjects. We also plot the percentage of long intervals between subjects. That is, we take the percentage for each participant and then take the boxplots and histograms with the summary statistic for each participant as a data point.

Figure 9 shows the plots for number of cigarettes since the last assessment. We see that most of the time they don't smoke any cigarettes. The histogram looks like a geometric distribution for the number of cigarettes. The mean number of cigarettes is 1.76, the median is 0, and the 25 and 75 percentiles are 0 and 2, respectively. Thus, in at least half of the intervals they don't smoke, and in 75% percent of them they smoke at most two cigarettes, but in some cases they smoke some huge number: in one case someone smoked over 40 cigarettes. That said, the number of cigarettes also will depend on the length of time elapsed since the last assessment, and the intensity function is time-varying: particularly, we noted in Figure 2b) that they tend to smoke more frequently in the pre-quit period.

Figure 10 plots the days between EMAs. Here we see that most observations are under one day, but there are a substantial number of outliers. In a number of cases the time between EMAs is over two days, and in several cases it is over six days. The mean time between observations is 0.34 days or approximately eight hours, the median is 0.17 or approximately four hours, and the 25 and 75 percentile are approximately two and 11 hours, respectively. Summarizing, most inter-EMA durations are under one day, but a few are very long.

Finally, also relevant is whether there is heterogeneity in the proportion of long intervals between subjects. Figure 11 investigates this. We see that participants vary between having no unreliable intervals and in one case having all unreliable intervals. The mean is 11%, the median is 6%, and the

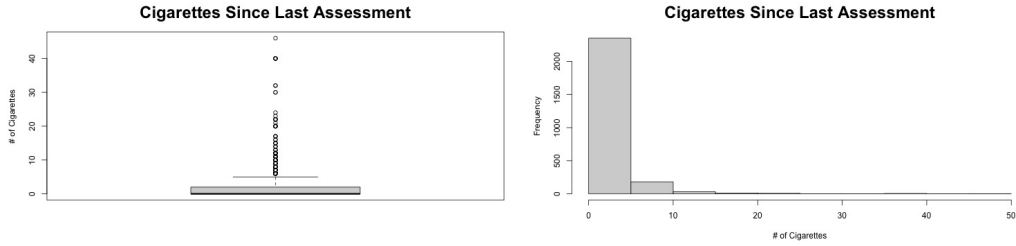


Figure 9: a) boxplot b) histogram for number of cigarettes since the last assessment. We see that the median number of cigarettes is 0, and that there are many outliers.

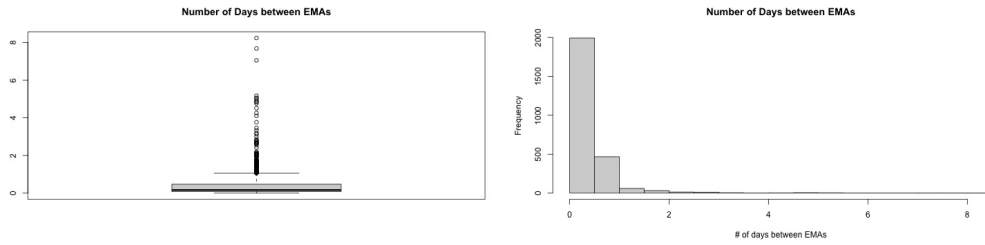


Figure 10: a) boxplot b) histogram for number of days between EMAs. Recall that we treat intervals over one day as missing/unreliable. From the boxplot we see that the mean and 75% percentile are under one day, but that there are a substantial number of observations over one day. The histogram suggests a similar finding.

25 and 75 percentile are 1.8% and 11.4%, respectively. Thus there is substantial heterogeneity. While our theory does not currently account for this and we leave it to future work, the experiments from the previous section where each subject had a different missingness probability (Figure 7) suggests that our model can handle this issue well.

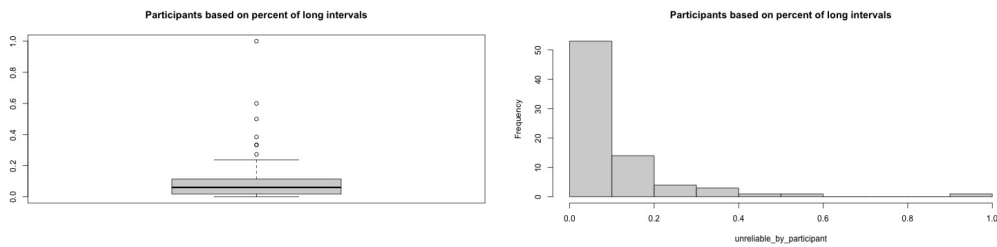


Figure 11: a) boxplot b) histogram for percent of long intervals between subjects. We see that most participants have a small proportion of long intervals (the median is 6.1%), while a few have a much larger proportion of long intervals.