UC Irvine

Title

Fitting birth–death processes to panel data with applications to bacterial DNA fingerprinting

Authors

Doss, Charles R
Suchard, Marc A
Holmes, Ian
et al.

Peer reviewed

FITTING BIRTH–DEATH PROCESSES TO PANEL DATA WITH APPLICATIONS TO BACTERIAL DNA FINGERPRINTING

By Charles R. Doss[*], Marc A. Suchard[1,5,†], Ian Holmes[2,‡],
Midori Kato-Maeda[3,§] and Vladimir N. Minin[4,5,*]

*University of Washington, Seattle[*], University of California, Los Angeles[†],
University of California, Berkeley[‡], and University of California,
San Francisco[§]*

Continuous-time linear birth–death-immigration (BDI) processes are frequently used in ecology and epidemiology to model stochastic dynamics of the population of interest. In clinical settings, multiple birth–death processes can describe disease trajectories of individual patients, allowing for estimation of the effects of individual covariates on the birth and death rates of the process. Such estimation is usually accomplished by analyzing patient data collected at unevenly spaced time points, referred to as panel data in the biostatistics literature. Fitting linear BDI processes to panel data is a nontrivial optimization problem because birth and death rates can be functions of many parameters related to the covariates of interest. We propose a novel expectation–maximization (EM) algorithm for fitting linear BDI models with covariates to panel data. We derive a closed-form expression for the joint generating function of some of the BDI process statistics and use this generating function to reduce the E-step of the EM algorithm, as well as calculation of the Fisher information, to one-dimensional integration. This analytical technique yields a computationally efficient and robust optimization algorithm that we implemented in an open-source R package. We apply our method to DNA fingerprinting of *Mycobacterium tuberculosis*, the causative agent of tuberculosis, to study intrapatient time evolution of IS*6110* copy number, a genetic marker frequently used during estimation of epidemiological clusters of *Mycobacterium tuberculosis* infections.

1

arXiv:1009.0893v5 [stat.CO] 10 Jan 2014

Our analysis reveals previously undocumented differences in IS*6110*
birth–death rates among three major lineages of *Mycobacterium tu-
berculosis*, which has important implications for epidemiologists that
use IS*6110* for DNA fingerprinting of *Mycobacterium tuberculosis*.

1. Introduction. Linear birth–death-immigration (BDI) processes pro-
vide useful building blocks for modeling population dynamics in ecology
[Nee (2006)], molecular evolution [Thorne, Kishino and Felsenstein (1991)]
and epidemiology [Gibson and Renshaw (1998)], among many other areas.
Although Keiding (1975) has extensively studied inference for fully ob-
served continuous-time BDI processes, more often such processes are not
observed completely, posing challenging computational problems for statis-
ticians. Here, we use applied probability tools to develop a new, efficient
implementation of the expectation–maximization (EM) algorithm for fitting
discretely observed BDI processes.

We are interested in situations where we observe multiple independent
continuous-time BDI trajectories at fixed, possibly irregularly spaced, time
points. Such observations, called panel data, often arise in medical applica-
tions, with independent BDI trajectories corresponding to some stochastic
process recorded in different patients under study [Crespi, Cumberland and
Blower (2005)]. The birth and death rates can then be modeled as functions
of patient-specific covariates. This modeling framework is similar to the use
of continuous-time Markov chains (CTMCs) in multi-state disease progres-
sion models with a finite number of states [Kalbfleisch and Lawless (1985)].
Although established methods for fitting finite state CTMCs to panel data
exist [Kalbfleisch and Lawless (1985), Lange (1995), Jackson (2011)], less
attention has been paid to infinite state-space processes, such as BDI mod-
els.

Outside of medical applications, estimating parameters of discretely ob-
served BDI models is considered in the molecular evolution and bioinfor-
matics literature [Thorne, Kishino and Felsenstein (1991), Holmes (2005)].
For example, Holmes (2005) proposed an EM algorithm for discretely ob-
served BDI processes in the context of finding the most optimal alignment
of multiple genomic sequences. The author argues that the EM algorithm's
simplicity and robustness make this method attractive for large-scale bioin-
formatics applications. Unfortunately, implementation of the EM algorithm
by Holmes (2005) is applicable only to a very restricted class of BDI pro-
cesses. In this paper, we develop a more general EM algorithm that applies
to a large class of BDI models and is not restricted to molecular evolution
applications.

Computing expectations of the complete-data log-likelihood, needed for
executing an EM algorithm, can be challenging, especially if the complete-
data were generated by a continuous-time stochastic process. When the com-
plete data are generated by a finite state-space CTMC, these expectations

can be computed efficiently [Lange (1995), Holmes and Rubin (2002)]. Although the BDI process is also a CTMC, the infinite state-space of the process prohibits us from using these computationally efficient methods. Holmes (2005) considers a BDI model with the immigration rate either zero or proportional to the birth rate. Under this restriction, the complete-data likelihood belongs to the exponential family, which means that the complete-data log-likelihood is a linear function of sufficient statistics of the complete data. Making further stringent assumptions about the initial state of the process, Holmes (2005) computes expectations of these sufficient statistics by numerically solving a system of coupled nonlinear ordinary differential equations (ODEs). Working with this birth–death-restricted immigration (BDRI) model, but without any restrictions on the starting state of the process, we develop a new computationally efficient method for computing the expected sufficient statistics. Our method combines ideas from Kendall (1948) and Lange (1982) and reduces computations of the expected sufficient statistics to one-dimensional integration, a computational task that is much simpler than solving a system of nonlinear ODEs. We develop a similar integration method to compute the observed Fisher information matrix via Louis' formula [Louis (1982)] and use this matrix for calculation of confidence intervals and sets. In addition, when we have multiple BDRI trajectories observed, we allow the birth and death rates to be functions of trajectory-specific covariates.

We first test our EM algorithm on simulated data and then turn to a problem of estimating birth and death rates of the transposable element IS*6110* in *Mycobacterium tuberculosis*, the causative bacterial agent of most tuberculosis (TB) in humans. *Mycobacterium tuberculosis* genome carries multiple IS*6110* copies that get duplicated and deleted rapidly during replication. Estimating IS*6110* copy number birth (duplication) and death (loss) rates is an important task in TB molecular epidemiology because researchers use IS*6110* copy number to group infected individuals into epidemiological clusters [Small et al. (1994)]. In the United States, the resurgence of TB cases, attributed to significant changes in socioeconomic factors, started in the late 1980s, with the number of TB cases reaching its peak in 1991 and steadily declining since then [Cattamanchi et al. (2006)]. Since 1991, the University of California, San Francisco has been maintaining a database of TB cases reported to the San Francisco Department of Public Health. The database contains demographic and certain clinical information as well as *M. tuberculosis* genotypes (e.g., IS*6110* copy number) for each reported TB case [Jasmer et al. (1999)]. Rosenberg, Tsolaki and Tanaka (2003) used a subset of this database to estimate IS*6110* birth and death rates. These authors proposed an approximate likelihood method to accomplish this estimation. We revisit this problem using our EM algorithm and compare our results with the approximation of Rosenberg, Tsolaki and Tanaka (2003). Further,

we examine differences in birth and death rates among three main lineages of *M. tuberculosis* and find that the East-Asian *M. tuberculosis* is evolving at a slower rate than its European–American counterpart. This novel finding has serious implications on the definition of epidemiological clusters based on the IS*6110* copy number. To investigate the possibility of spurious effect of *M. tuberculosis* lineage on IS*6110* birth and death rates due to a confounding factor, we build a more complicated model for birth and death rates. In addition to the lineage, we include *M. tuberculosis* drug-resistance status and HIV infection status of each patient as birth and death rate covariates. We find that after including these covariates, the lineage remains the only variable that significantly affects IS*6110* birth and death rates.

2. BDRI process with covariates. We start with m independent continuous-time homogeneous linear BDRI processes $\{X_{p,t}\}$, for $p = 1, \ldots, m$, with corresponding per capita birth rates $\lambda_p \geq 0$, per capita death rates $\mu_p \geq 0$ and immigration rates $\nu_p = \beta\lambda_p$, where $\beta \geq 0$ is a known constant. Assuming that each process p has c_1 covariates related to the birth rates and c_2 covariates related to the death rates, collected into vectors $\mathbf{z}'_{p,\lambda} = (z_{p,\lambda,1}, \ldots, z_{p,\lambda,c_1}) \in \mathbb{R}^{c_1}$ and $\mathbf{z}'_{p,\mu} = (z_{p,\mu,1}, \ldots, z_{p,\mu,c_2}) \in \mathbb{R}^{c_2}$, we model birth and death rates as log-linear functions of these covariates:

$$(1) \qquad \log \lambda_p = \mathbf{z}'_{p,\lambda}\boldsymbol{\gamma}_\lambda \quad \text{and} \quad \log \mu_p = \mathbf{z}'_{p,\mu}\boldsymbol{\gamma}_\mu,$$

where $\boldsymbol{\gamma}'_\lambda = (\gamma_{\lambda,1}, \ldots, \gamma_{\lambda,c_1})$ and $\boldsymbol{\gamma}'_\mu = (\gamma_{\mu,1}, \ldots, \gamma_{\mu,c_2})$ are birth and death regression coefficients. Covariate vectors $\mathbf{z}_{p,\lambda}$ and $\mathbf{z}_{p,\mu}$ are assumed to be known and fixed for every process p. For example, if each BDRI process models a disease related trajectory for each patient, then covariates are usually composed of patient-specific clinical and demographic information (e.g., gender, medical history).

We assume that we observe the pth process at $n(p) + 1$ distinct times, $0 = t_{p,0} < t_{p,1} < \cdots < t_{p,n(p)}$. We denote our data vector by

$$\mathbf{Y} = (X_{1,t_{1,0}}, \ldots, X_{1,t_{1,n(1)}}, \ldots, X_{m,t_{m,0}}, \ldots, X_{m,t_{m,n(m)}})$$

and the parameter vector by $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_\lambda, \boldsymbol{\gamma}_\mu) \in \mathbb{R}^{c_1+c_2}$. We are interested in computing the parameter maximum likelihood estimates (MLEs), $\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} l_o(\mathbf{Y}; \boldsymbol{\gamma})$, where

$$(2) \qquad l_o(\mathbf{Y}; \boldsymbol{\gamma}) := \sum_{p=1}^{m} \sum_{i=0}^{n(p)-1} \log p_{X_{p,t_{p,i}}, X_{p,t_{p,i+1}}}(t_{p,i+1} - t_{p,i}; \lambda_p, \mu_p)$$

is the observed-data log-likelihood and $p_{i,j}(t; \lambda, \mu) = P_{\lambda,\mu}(X_t = j | X_0 = i)$, $i, j = 0, 1, \ldots$, are the transition probabilities of the BDRI process. These transition probabilities can be calculated either using the generating function

derived by Kendall ([1948](#)) or via the orthogonal polynomial representation of Karlin and McGregor ([1958](#)). Despite the explicit algebraic nature of the orthogonal polynomials, the latter method can be numerically unstable and the generating function method is often preferred [Sehl et al. ([2011](#))]. Although one can maximize the likelihood $l_o(\mathbf{Y}; \boldsymbol{\gamma})$ using standard off-the-shelf optimization algorithms, such generic algorithms can be problematic when the BDI rates are functions of a high-dimensional parameter vector, such as the vector of regression coefficients $\boldsymbol{\gamma}$ in our case. As an alternative to generic optimization, we develop an EM algorithm, known for its robustness and ability to cope with high-dimensional optimization [Dempster, Laird and Rubin ([1977](#))].

3. EM algorithm for the BDRI process. The complete data in our case consist of the BDRI trajectories $\{X_{p,t}\}$, observed continuously during the corresponding intervals $[0, t_{p,n(p)}]$, $p = 1, \ldots, m$. Let $\mathbf{X} = \{X_{p,t}\}_{p=1,\ldots,m}^{t \in [0, t_{p,n(p)}]}$ be the complete data and let $l_c(\mathbf{X}; \boldsymbol{\gamma})$ be the complete data log-likelihood. The EM algorithm starts by initializing the parameter vector to an arbitrarily chosen vector $\boldsymbol{\gamma}_0$. At the kth iteration of the algorithm we set

$$(3) \qquad \boldsymbol{\gamma}_k = \arg \max_{\boldsymbol{\gamma}} \mathrm{E}_{\boldsymbol{\gamma}_{k-1}}[l_c(\mathbf{X}; \boldsymbol{\gamma}) | \mathbf{Y}].$$

To accomplish the above maximization, we need to be able to evaluate the expectation in ([3](#)) for any vector $\boldsymbol{\gamma}$. Traditionally, a numerical procedure for computing such an expectation is called an E-step of the EM algorithm. The maximization of the expectation is called an M-step of the EM algorithm. Below, we develop efficient algorithms for implementing these E- and M-steps for the discretely observed BDRI process. As is often the case, we will see that to compute the needed expectations for all $\boldsymbol{\gamma} \in \mathbb{R}^{c_1+c_2}$, we need to compute only the expectations of certain statistics that do not depend on $\boldsymbol{\gamma}$.

3.1. *E-step.* Since our BDRI process is a CTMC, the log-likelihood of the complete data is

$$
\begin{aligned}
l_c(\mathbf{X}; \boldsymbol{\gamma}) = -\sum_{p=1}^{m} &\left[\sum_{i=0}^{\infty} d^p(i)[i(\lambda_p + \mu_p) + \nu_p] \right.\\
&\left. + \sum_{i=0}^{\infty} (n_{i,i+1}^p \log(i\lambda_p + \nu_p) + n_{i,i-1}^p \log(i\mu_p)) \right],
\end{aligned}
$$

(4)

where $d^p(i)$ is the total time spent by $X_{p,t}$ in state i and $n_{i,j}^p$ is the number of jumps from state i to state j during the interval $[0, t_{p,n(p)}]$ [Guttorp ([1995](#))]. Replacing ν_p with $\beta\lambda_p$ in the above equation, we arrive at a more compact

representation of the complete-data log-likelihood:

$$
l_c(\mathbf{X}; \boldsymbol{\gamma}) = \sum_{p=1}^{m} [-R_{p,t_{p,n(p)}}(\lambda_p + \mu_p) - t_{n(p)}\beta\lambda_p
$$

(5)

$$
+ N_{p,t_{p,n(p)}}^{+} \log \lambda_p + N_{p,t_{p,n(p)}}^{-} \log \mu_p] + \mathrm{const},
$$

where the number of jumps up $N_{p,t_{p,n(p)}}^{+} := \sum_{i\geq 0} n_{i,i+1}^p$, the number of jumps down $N_{p,t_{p,n(p)}}^{-} := \sum_{i\geq 0} n_{i,i-1}^p$, and the total particle-time

$$
R_{p,t_{p,n(p)}} := \int_{t_0}^{t_{p,n(p)}} X_s\, ds = \sum_{i=0}^{\infty} i d^p(i)
$$

for $p = 1, \ldots, m$, are the sufficient statistics. Equation (5) shows that, for the E-step, the only expectations we need are $E_{\tilde{\gamma}}[N_{p,t_{p,n(p)}}^{+}|\mathbf{Y}]$, $E_{\tilde{\gamma}}[N_{p,t_{p,n(p)}}^{-}|\mathbf{Y}]$ and $E_{\tilde{\gamma}}[R_{p,t_{p,n(p)}}|\mathbf{Y}]$ for all values $\tilde{\gamma}$. Using independence of the p BDRI processes, the Markov property and additivity of expectations, we break the desired expectations into sums of expectations of the numbers of jumps up and down and the total particle time during each time interval $[t_{p,k}, t_{p,k+1}]$, conditional on $X_{p,t_{p,k}}$ and $X_{p,t_{p,k+1}}$. By the homogeneity of each of the BDRI processes, in order to complete the E-step of the EM algorithm, we need to be able to calculate

$$
U_{i,j}(t) = U_{i,j}(t; \lambda, \mu) = \mathrm{E}(N_t^+|X_0 = i, X_t = j),
$$

(6)

$$
D_{i,j}(t) = D_{i,j}(t; \lambda, \mu) = \mathrm{E}(N_t^-|X_0 = i, X_t = j) \quad \text{and}
$$

$$
P_{i,j}(t) = P_{i,j}(t; \lambda, \mu) = \mathrm{E}(R_t|X_0 = i, X_t = j)
$$

for all nonnegative integers i and j.

Following Minin and Suchard (2008), we choose to work with restricted moments

$$
\tilde{U}_{i,j}(t) = \tilde{U}_{i,j}(t; \lambda, \mu) = \mathrm{E}(N_t^+ 1_{\{X_t=j\}}|X_0 = i),
$$

(7)

$$
\tilde{D}_{i,j}(t) = \tilde{D}_{i,j}(t; \lambda, \mu) = \mathrm{E}(N_t^- 1_{\{X_t=j\}}|X_0 = i) \quad \text{and}
$$

$$
\tilde{P}_{i,j}(t) = \tilde{P}_{i,j}(t; \lambda, \mu) = \mathrm{E}(R_t 1_{\{X_t=j\}}|X_0 = i),
$$

that we can divide by transition probabilities $p_{i,j}(t)$ to recover the conditional expectations (6),

$$
U_{i,j}(t) = \tilde{U}_{i,j}(t)/p_{i,j}(t),
$$

(8)

$$
D_{i,j}(t) = \tilde{D}_{i,j}(t)/p_{i,j}(t) \quad \text{and}
$$

$$
P_{i,j}(t) = \tilde{P}_{i,j}(t)/p_{i,j}(t).
$$

In order to compute the restricted moments, we first consider the joint generating function

(9) $$H_i(u, v, w, s, t) := \mathrm{E}(u^{N_t^+} v^{N_t^-} e^{-wR_t} s^{X_t} | X_0 = i),$$

where $0 \le u, v, s \le 1$ and $w \ge 0$. Partial derivatives of this function,

$$\left. \frac{\partial H_i(u, 1, 0, s, t)}{\partial u} \right|_{u=1} = \sum_{j=0}^{\infty} s^j \sum_{n=0}^{\infty} n \Pr_i(N_t^+ = n, X_t = j)$$

$$= \sum_{j=0}^{\infty} \tilde{U}_{i,j}(t) s^j,$$

$$\left. \frac{\partial H_i(1, v, 0, s, t)}{\partial v} \right|_{v=1} = \sum_{j=0}^{\infty} s^j \sum_{n=0}^{\infty} n \Pr_i(N_t^- = n, X_t = j)$$

(10) $$= \sum_{j=0}^{\infty} \tilde{D}_{i,j}(t) s^j \quad \text{and}$$

$$\left. \frac{\partial H_i(1, 1, w, s, t)}{\partial w} \right|_{w=0} = -\sum_{j=0}^{\infty} s^j \int_0^{\infty} x \, d\Pr_i(R_t \le x, X_t = j)$$

$$= -\sum_{j=0}^{\infty} \tilde{P}_{i,j}(t) s^j$$

are power series with coefficients $\tilde{U}_{i,j}(t)$, $\tilde{D}_{i,j}(t)$ and $-\tilde{P}_{i,j}(t)$, respectively, for $j = 0, 1, \ldots, \infty$, where $\Pr_i$ denotes probability conditional on $X_0 = i$. We will denote these power series by $G_i^+(t, s)$, $G_i^-(t, s)$ and $G_i^*(t, s)$, respectively. If we can compute $G_i^+(t, s)$, $G_i^-(t, s)$ and $G_i^*(t, s)$ for every possible t and s, then we should be able to recover coefficients of the corresponding power series via differentiation or integration. Numerical evaluation of the partial derivatives (10) is straightforward if we can compute finite differences of $H_i(u, v, w, s, t)$. Remarkably, $H_i(u, v, w, s, t)$ is available in closed form, as we demonstrate in the theorem below, so one can even obtain derivatives (10) analytically. Note that the theorem below applies to a general linear BDI process, not only to the BDRI processes.

THEOREM 1. *Let $\{X_t\}$ be a linear BDI process with parameters $\lambda \ge 0$, $\mu \ge 0$ and $\nu \ge 0$. Over the interval $[0, t]$, let N_t^+ be the number of jumps up, N_t^- be the number of jumps down and R_t be the total particle-time. Then $H_i(u, v, w, s, t) = E(u^{N_t^+} v^{N_t^-} e^{-wR_t} s^{X_t} | X_0 = i)$ satisfies the following partial differential equation:*

(11) $$\frac{\partial}{\partial t} H_i = [s^2 u \lambda - (\lambda + \mu + w)s + v\mu] \frac{\partial}{\partial s} H_i + \nu(us - 1) H_i,$$

subject to initial condition $H_i(u, v, w, s, 0) = s^i$. The Cauchy problem defined by equation (11) and the initial condition has a unique solution. When $\lambda > 0$, the solution is

$$H_i(u, v, w, s, t) = \left(\frac{\alpha_1 - \alpha_2(s - \alpha_1)e^{-\lambda(\alpha_2 - \alpha_1)ut}/(s - \alpha_2)}{1 - (s - \alpha_1)e^{-\lambda(\alpha_2 - \alpha_1)ut}/(s - a2)} \right)^i$$

(12)

$$\times \left(\frac{\alpha_1 - \alpha_2}{s - \alpha_2 - (s - \alpha_1)e^{-\lambda(\alpha_2 - \alpha_1)ut}} \right)^{\nu/\lambda} e^{-\nu(1 - u\alpha_1)t},$$

where $\alpha_1 = \frac{\lambda + \mu + w - \sqrt{(\lambda + \mu + w)^2 - 4\lambda\mu uv}}{2\lambda u}$ and $\alpha_2 = \frac{\lambda + \mu + w + \sqrt{(\lambda + \mu + w)^2 - 4\lambda\mu uv}}{2\lambda u}$. When $\lambda = 0$, the solution is

$$H_i(u, v, w, s, t) = \left(se^{-(\mu+w)t} - \frac{v\mu(e^{-(\mu+w)t} - 1)}{\mu + w} \right)^i$$

(13)

$$\times e^{\nu u[v\mu - (\mu+w)s](e^{-(\mu+w)t} - 1)/(\mu+w)^2 + \nu(uv\mu/(\mu+w) - 1)t}.$$

PROOF.   Our proof, detailed in Appendix A, is a generalization of Kendall's derivation of the generating function of X_t [Doss et al. (2013), Kendall (1948)]. $\square$

Having H_i in closed form gives us access to functions G_i^+, G_i^- and G_i^*, so we are left with the task of recovering coefficients of these power series. One way to accomplish this task is to differentiate the power series repeatedly, for example, $\tilde{U}_{i,j}(t) = \frac{1}{j!} \frac{\partial^j G_i^+(s,t)}{\partial s^j}|_{s=0}$. In Appendix C, we demonstrate that for the death-immigration model ($\lambda = 0$, $\nu \neq 0$, $\mu \neq 0$) and the BDRI model considered by Holmes (2005), these derivatives can be found analytically [Doss et al. (2013)]. In general, repeated differentiation of G_i^+, G_i^- and G_i^* needs to be done numerically, making this method impractical. Instead, we extend $G_i^+(t, \cdot)$, $G_i^-(t, \cdot)$ and $G_i^*(t, \cdot)$ to the boundary of a unit circle in the complex plane by the change of variables $s = e^{2\pi i z}$ (i in this context is the imaginary number $\sqrt{-1}$, not the initial state of the BDI process). For example,

$$G_l^+(t, e^{2\pi i z}) = \sum_{j=0}^{\infty} \tilde{U}_{l,j}(t) e^{2\pi i j z}$$

is a periodic function in z, which means that $\tilde{U}_{l,j}(t)$ are Fourier coefficients of this periodic function. Therefore, we can use the Riemann approximation to the Fourier transform integral to obtain

$$\tilde{U}_{l,j}(t) = \int_0^1 G_l^+(t, e^{2\pi i s}) e^{-2\pi i j s} \, ds \approx \frac{1}{K} \sum_{k=0}^{K-1} G_l^+(t, e^{2\pi i k/K}) e^{-2\pi i j k/K}$$

for some suitably large K. The Fast Fourier Transform (FFT) [Henrici (1979)] can be applied to quickly compute multiple Fourier coefficients [Lange (1982), Dorman, Sinsheimer and Lange (2004), Suchard, Lange and Sinsheimer (2008)]. We do not, however, use the FFT in our algorithm because, for a particular time interval length t, we almost always need to compute $\tilde{U}_{i,j}(t)$, $\tilde{D}_{i,j}(t)$, $\tilde{P}_{i,j}(t)$ for only one value of j.

Now, we can put the pieces together to compute $\mathrm{E}_{\tilde{\gamma}}[l_c(\mathbf{X};\boldsymbol{\gamma})|\mathbf{Y}]$. As mentioned above, $N^+_{p,t_{p,n(p)}}$ equals the sum of the number of jumps up over the disjoint intervals $[t_{p,i-1}, t_{p,i})$, $i = 1,\ldots,n(p)$. The Markov property says that the conditional expectations of the number of jumps up of $X_{p,t}$ over $[t_{p,i-1}, t_{p,i})$ given $\mathbf{Y}$ is equal to the conditional expectation of the number of jumps up over $[t_{p,i-1}, t_{p,i})$ given just $X_{p,t_{p,i-1}}$ and $X_{p,t_{p,i}}$. Using similar logic for $N^-_{p,t_{p,n(p)}}$ and $R_{p,t_{p,n(p)}}$, this gives for $p = 1,\ldots,m$,

$$\mathrm{E}_{\tilde{\gamma}_p}[N^+_{p,t_{p,n(p)}}|\mathbf{Y}] = \sum_{i=1}^{n(p)} U_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p),$$

$$(14) \quad \mathrm{E}_{\tilde{\gamma}_p}[N^-_{p,t_{p,n(p)}}|\mathbf{Y}] = \sum_{i=1}^{n(p)} D_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p) \quad \text{and}$$

$$\mathrm{E}_{\tilde{\gamma}_p}[R_{p,t_{p,n(p)}}|\mathbf{Y}] = \sum_{i=1}^{n(p)} P_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p),$$

where $\log \tilde{\lambda}_p = \mathbf{z}'_{p,\lambda}\tilde{\gamma}_{p,\lambda}$ and $\log \tilde{\mu}_p = \mathbf{z}'_{p,\mu}\tilde{\gamma}_{p,\mu}$. Thus, by (5), (8) and (14), we see that, up to an additive constant, $\mathrm{E}_{\tilde{\gamma}}[l_c(\mathbf{X};\boldsymbol{\gamma})|\mathbf{Y}]$ is equal to

$$\sum_{p=1}^{m}\Bigg\{-t_{n(p)}\beta\lambda_p$$
$$+ \sum_{i=1}^{n(p)}\Bigg(-\frac{\tilde{P}_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p)}{p_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p)}(\lambda_p + \mu_p)$$
$$+ \frac{\tilde{U}_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p)}{p_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p)}\log \lambda_p$$
$$+ \frac{\tilde{D}_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p)}{p_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p)}\log \mu_p\Bigg)\Bigg\},$$

where the transition probabilities $p_{X_{p,t_{p,i-1}},X_{p,t_{p,i}}}(t_{p,i} - t_{p,i-1}; \tilde{\lambda}_p, \tilde{\mu}_p)$ can be calculated by using the (known) generating function for the BDI process, as is described in Appendix A [Doss et al. (2013)].

3.2. *M-step.* To complete the M-step for each iteration of the EM algorithm, we use a Newton–Raphson algorithm to maximize

$$f(\boldsymbol{\gamma}) = \mathrm{E}_{\tilde{\boldsymbol{\gamma}}}[l_c(\mathbf{X}; \boldsymbol{\gamma})|\mathbf{Y}].$$

In each Newton–Raphson step, we update $\boldsymbol{\gamma}$ via the following recursion:

$$\boldsymbol{\gamma}_{\text{new}} = \boldsymbol{\gamma}_{\text{cur}} - [\mathbf{H}f(\boldsymbol{\gamma}_{\text{cur}})]^{-1}\boldsymbol{\nabla}f(\boldsymbol{\gamma}_{\text{cur}}),$$

where $\boldsymbol{\nabla}f(\boldsymbol{\gamma}_{\text{cur}})$ is the gradient vector and $\mathbf{H}f(\boldsymbol{\gamma}_{\text{cur}})$ is the Hessian matrix of the function $f(\boldsymbol{\gamma})$. If we collect the observation times into a vector $\mathbf{T}' = (t_{1,n(1)}, \ldots, t_{m,n(m)})$, the expectations of the sufficient statistics into vectors

$$\mathbf{U}' = (E_{\tilde{\boldsymbol{\gamma}}}[N^+_{1,t_{1,n(1)}}|\mathbf{Y}], \ldots, E_{\tilde{\boldsymbol{\gamma}}}[N^+_{m,t_{m,n(m)}}|\mathbf{Y}]),$$

(15) $$\mathbf{D}' = (E_{\tilde{\boldsymbol{\gamma}}}[N^-_{1,t_{1,n(1)}}|\mathbf{Y}], \ldots, E_{\tilde{\boldsymbol{\gamma}}}[N^-_{m,t_{m,n(m)}}|\mathbf{Y}]),$$

$$\mathbf{P}' = (E_{\tilde{\boldsymbol{\gamma}}}[R_{1,t_{1,n(1)}}|\mathbf{Y}], \ldots, E_{\tilde{\boldsymbol{\gamma}}}[R_{m,t_{m,n(m)}}|\mathbf{Y}]),$$

and the process-specific birth and death rates into vectors

$$\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_m) \quad \text{and} \quad \boldsymbol{\mu}' = (\mu_1, \ldots, \mu_m),$$

then after defining covariate matrices

$$\mathbf{Z}'_\lambda = (\mathbf{z}_{1,\lambda}, \ldots, \mathbf{z}_{m,\lambda}) \quad \text{and} \quad \mathbf{Z}'_\mu = (\mathbf{z}_{1,\mu}, \ldots, \mathbf{z}_{m,\mu}),$$

the gradient and the Hessian can be compactly expressed in matrix form as

(16) $$\boldsymbol{\nabla}f(\boldsymbol{\gamma}) = (\mathbf{Z}'_\lambda[-\operatorname{diag}(\mathbf{P} + \beta\mathbf{T})\boldsymbol{\lambda} + \mathbf{U}], \mathbf{Z}'_\mu[-\operatorname{diag}(\mathbf{P})\boldsymbol{\mu} + \mathbf{D}]),$$

(17) $$\mathbf{H}f(\boldsymbol{\gamma}) = \begin{pmatrix} -\mathbf{Z}'_\lambda \operatorname{diag}(\mathbf{P} + \beta\mathbf{T})\operatorname{diag}(\boldsymbol{\lambda})\mathbf{Z}_\lambda & \mathbf{0} \\ \mathbf{0} & -\mathbf{Z}'_\mu \operatorname{diag}(\mathbf{P})\operatorname{diag}(\boldsymbol{\mu})\mathbf{Z}_\mu \end{pmatrix},$$

which we show in Appendix B; see (S-4), (S-6) and (S-9) [Doss et al. (2013)]. Notice that the algebraic separation of the birth and the death components in the complete-data likelihood results in blocks—corresponding to $\boldsymbol{\gamma}_\lambda$ and $\boldsymbol{\gamma}_\mu$—in the above formulae. The fact that the gradient and Hessian of $f(\boldsymbol{\gamma})$ is available analytically results in fast execution of Newton–Raphson updates. In our experience, the Newton–Raphson algorithm in our M-step converges after only 3–5 iterations. However, we also note that it is not critical to achieve convergence of this algorithm since even a single Newton–Raphson update within the M-step is enough to guarantee the usual convergence properties of the EM algorithm [Lange (1995)].

We obtain the observed Fisher information via Louis' formula:

$$\hat{I}_{\mathbf{Y}}(\hat{\boldsymbol{\gamma}}) = \mathrm{E}_{\hat{\boldsymbol{\gamma}}}[-\mathbf{H}l_c(\mathbf{X}; \hat{\boldsymbol{\gamma}})|\mathbf{Y}] - \mathrm{E}_{\hat{\boldsymbol{\gamma}}}[\boldsymbol{\nabla}l_c(\mathbf{X}; \hat{\boldsymbol{\gamma}})\boldsymbol{\nabla}l_c(\mathbf{X}; \hat{\boldsymbol{\gamma}})'|\mathbf{Y}],$$

where $\boldsymbol{\nabla}l_c$ is the gradient and $\mathbf{H}l_c$ is the Hessian of the complete-data log-likelihood [Louis (1982)]. This requires calculation of the conditional cross-product means, $\mathrm{E}[N^+_t N^-_t|\mathbf{Y}]$, $\mathrm{E}[N^+_t R_t|\mathbf{Y}]$, $\mathrm{E}[N^-_t R_t|\mathbf{Y}]$, and the conditional

TABLE 1
Summary statistics for the simulated and M. tuberculosis IS6110 data

Value	Simulated data	IS6110 data
Number of intervals	387	252
Average interval length	5	0.35
Number of individuals	100	196
Number of intervals with an increase	78	14
Average increase given an increase	1.5	1
Number of intervals with a decrease	190	14
Average decrease given a decrease	2.5	1.2
Number of intervals with no change	119	224
Mean starting state	5.5	11
Standard deviation of starting state	3.8	5.3
Total length of time	1947	89

second moments of N_t^+, N_T^- and R_t. The derivation of the information in terms of these moments is in Appendix B [Doss et al. (2013)]. These conditional second- and cross-moments, as well as $\mathbf{P}$ and $\mathbf{D}$, can be computed in analogous fashion to $\mathbf{U}$ above, using the joint generating function (12). We use the information matrix to compute approximate standard errors of $\hat{\gamma}$ and use these standard errors together with asymptotic normality of maximum likelihood estimators to form confidence intervals and sets for our model parameters.

4. Results.

4.1. *Simulations.* To test our methods, we simulate data from the BDRI model with $\lambda = 0.07$, $\mu = 0.12$ and $\beta = 1.2$, where β is assumed to be known, leaving us with only two parameters to estimate: λ and μ. We choose these parameters to resemble, but not exactly match, the dynamics of our biological example, discussed in the next subsection. We simulate 100 independent processes starting from initial states drawn uniformly between 1 and 15. From each process we collect at least two observations. We place observation times uniformly between 0 and 30. Table 1 gives some summary statistics for the simulated data.

We test our EM algorithm and confidence interval calculations on these simulated data with initial parameter values of 0.2 for both λ and μ. We considered other choices of starting values, but the algorithm was not sensitive to them. Notice that this is the simplest parameterization of our BDRI model, where both $\mathbf{z}_\lambda$ and $\mathbf{z}_\mu$ are vectors of ones. We estimate 0.067 with a 95% confidence interval of $(0.052, 0.081)$ for λ and $0.12, (0.1, 0.14)$ for μ, indicating that our algorithm successfully recovered these BDRI model parameters. We also conduct a similar simulation study for the BDRI model

with covariates, successfully estimating parameters of this model as well, but omit detailed results of this simulation for brevity.

4.2. *Comparison with the frequent monitoring method.* We compare our EM algorithm for computing the actual MLE to the frequent monitoring (FM) method of Rosenberg, Tsolaki and Tanaka (2003) for computing the MLE of an approximate likelihood. In the FM method, Rosenberg, Tsolaki and Tanaka (2003) assume that if the starting and ending values of the birth–death process are equal for a particular interval, then no jumps occurred in this interval. Further, if the difference between the starting and ending values is -1 or 1, then exactly one jump up or exactly one jump down must have occurred, respectively. The authors exclude all observed intervals, for which starting and ending values differ by more than one unit. Let i be the starting state for an interval, t the length of the interval and $\lambda_i = i(\lambda + \mu)$. Then the corresponding probabilities for the three possible events are $e^{-\lambda_i u}$, $\frac{i\lambda}{\lambda_i}(1 - e^{-\lambda_i u})$ and $\frac{i\mu}{\lambda_i}(1 - e^{-\lambda_i u})$, respectively. Rosenberg, Tsolaki and Tanaka (2003) use this FM method to estimate rates in what is effectively a multi-state branching process, but we will compare the two methods on our BDRI model with the immigration rate β constrained to be 0. We again simulate an underlying BD process using $\lambda = 0.07$ and $\mu = 0.12$. To compare the two methods, we generate three different sets of data. In each set, we generate observed states of the BD process at a fixed constant distance dt apart. This distance varies across the data sets, taking the values $0.2, 0.4$ and 0.6, respectively. We repeat this procedure 200 times and compute birth and death rate estimates and corresponding 95% confidence intervals using the EM algorithm and FM approximation method. We show box plots of the resulting estimates for λ and μ in Figure 1. As expected, the FM estimates behave reasonably when interval lengths are small, but the approximation becomes poor as we increase the interval length. The FM method always underestimates the parameters since the method effectively undercounts the number of unobserved jumps in the BD process. We also compute Monte Carlo estimates of coverage probabilities of the two methods, shown above the box plots in Figure 1. Not surprisingly, coverage of the 95% confidence intervals computed under the proper BD model likelihood are very close to the promised value of 0.95. In contrast, the FM approximation-based 95% confidence intervals contain the true parameter value less than 95% for all three simulation scenarios.

4.3. *Mycobacterium tuberculosis IS6110 transposon.* We apply our EM algorithm to estimation of birth and death rates of the transposon IS*6110* in *M. tuberculosis* [McEvoy et al. (2007)]. A transposon, or transposable element, is a genetic sequence that can duplicate, remove itself and jump to a
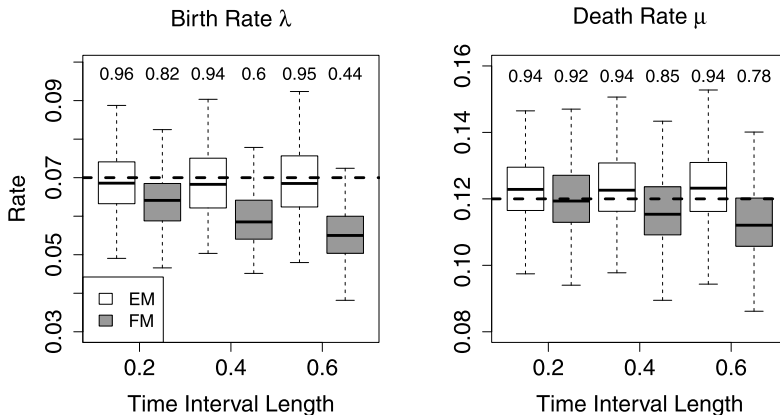
FIG. 1. *Box plots of birth (left panel) and death (right panel) rate estimates, obtained from 200 simulated data sets using the EM algorithm and frequent monitoring (FM) method. The true parameter values, used in data simulations, are marked by the horizontal dashed lines. Above the box plots, we show Monte Carlo estimates of coverage probabilities of the 95% confidence intervals.*

new location in the genome. IS*6110* is a transposon that plays an important role in epidemiological studies of tuberculosis. More specifically, the number and locations of IS*6110* elements in the *M. tuberculosis* form a genetic signature or genotype of the mycobacterium, allowing epidemiologists to draw inference about disease transmission when the same genotype is observed among patients with active tuberculosis [van Embden et al. (1993)]. Such genotypic comparison can translate into meaningful epidemiological inference only if the dynamics of IS*6110* evolution are well understood. Therefore, accurate estimation of rates of changes of IS*6110*-based genotypes is critical for using these genotypes in epidemiological studies [Tanaka and Rosenberg (2001)].

We analyze data from an ongoing population-based study that includes all tuberculosis cases reported to the San Francisco Department of Public Health [Cattamanchi et al. (2006)]. Our data include patients with more than one *M. tuberculosis* isolate from specimens sampled more than 10 days apart and genotyped with IS*6110* restriction fragment length polymorphism. We ignore genomic locations of IS*6110* and assume that the transposon counts are discretely observed realizations of a BDRI process, with no immigration ($\beta = 0$); in particular, we assume that patients are not reinfected with a different strain of the bacteria in the period between observations. The third column in Table 1 gives summary statistics for the data.

We first use a simple model with one single birth rate and one single death rate of the IS*6110* for all patients. In the analysis presented, we start the EM algorithm with parameter guesses of 0.05 and 0.05 for λ and μ, respectively, and their MLEs are 0.0176 and 0.0207, respectively. The start-

ing values for the EM do not affect these results. Our estimate and 95% confidence interval for λ, 0.0176 and $(0.0082, 0.027)$, are consistent with the corresponding quantities, 0.0188 and $(0.0085, 0.0291)$, from Rosenberg, Tsolaki and Tanaka (2003). Although the authors' confidence interval for μ, $(0.0057, 0.0237)$, overlaps with ours, $(0.011, 0.031)$, our estimate for μ, 0.0207, is noticeably higher than Rosenberg, Tsolaki and Tanaka's (2003) estimate of 0.0147. Note from Table 1 that among the intervals with a decrease, the average count drop is by more than 1; there are 3 intervals where IS*6110* counts drop by 2, whereas there are no intervals that experience an increase by more than 1. Thus, we would expect our estimate for μ to increase over Rosenberg, Tsolaki and Tanaka's (2003) approximation, whereas that of λ should be similar between the two methods. We also point out that we analyze an updated version of the data analyzed by Rosenberg, Tsolaki and Tanaka (2003). Moreover, Rosenberg, Tsolaki and Tanaka (2003) use a slightly more complicated model for IS*6110* evolution, which takes into account shifts in transposon location. We conclude that estimates of birth and death rates of IS*6110* do not vary dramatically when estimation methods and data collection are altered. We now turn to more complicated BDRI models that have not been applied before to the *M. tuberculosis* IS*6110* copy number evolution. These models will take into account potential dependence of IS*6110* birth and death rates on patient-specific covariates.

4.3.1. *Mycobacterium tuberculosis lineage comparison.* In addition to estimation of the global birth and death rates, we separately estimate these parameters in each of the three lineages of *M. tuberculosis* observed in San Francisco. Based on genomic sequence similarity, *M. tuberculosis* is divided into six main lineages: Euro-American, East-Asian, Indo-Oceanic, East-African–Indian, West-African I and West-African II [Gagneux et al. (2006)]. In our lineage-specific analysis, we consider 109 individuals infected with Euro-American (EU) lineage strains, 54 individuals infected with East-Asian (EA) lineage strains and 25 individuals infected with Indo-Oceanic (IND) lineage strains. One simple way to accommodate this lineage effect is to build a log-linear model for birth and death rates with two categorical covariates:

$$\log \lambda_p = \gamma_{\lambda,1} + \gamma_{\lambda,2} \, \mathrm{EU}_p + \gamma_{\lambda,3} \, \mathrm{IND}_p, \log \mu_p = \gamma_{\mu,1} + \gamma_{\mu,2} \, \mathrm{EU}_p + \gamma_{\mu,3} \, \mathrm{IND}_p,$$

where $\mathrm{EU}_p = 1$ if patient p is infected with the EU strain and 0 otherwise, and $\mathrm{IND}_p = 1$ if patient p is infected with the IND strain and 0 otherwise. The intercepts, $\gamma_{\lambda,1}$ and $\gamma_{\mu,1}$, correspond to birth and death of the EA strain. We transform the coefficients $(\gamma_{\lambda,1}, \gamma_{\lambda,2}, \gamma_{\lambda,3})$ and $(\gamma_{\mu,1}, \gamma_{\mu,2}, \gamma_{\mu,3})$ into the *M. tuberculosis* lineage-specific birth and death rates and show these estimates together with their corresponding confidence in the first column of Figure 2. Most notably, there appears to be a substantial difference between death rates of the Euro-American and East-Asian lineages. We report regression
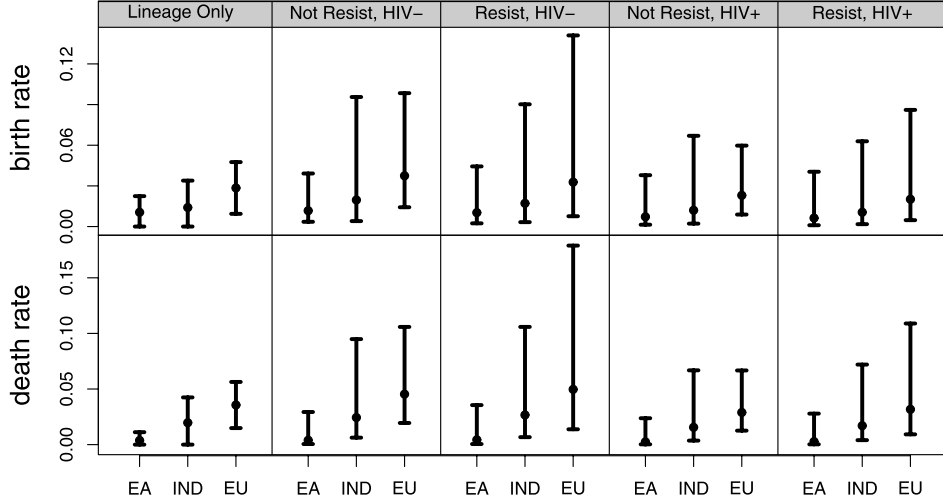
FIG. 2. *Point estimates and 95% confidence intervals for birth and death rates of the IS6110 transposable element obtained by separately analyzing three M. tuberculosis lineages: European–American (EU), Indo-Oceanic (IND) and East Asian (EA) (leftmost column) and by fitting the log-linear model with lineage, drug resistance and HIV status as covariates. For the latter model, the estimated regression coefficients are transformed into four sets of lineage-specific birth and death rates (last four columns).*

coefficients on the multiplicative scale [e.g., $\exp(\gamma_{\lambda,1})$] with their corresponding 95% confidence intervals in the lineage model columns of Table 2. In this table the highlighted EU rate multiplier shows that the death rate of IS*6110* copy number is estimated to be approximately ten times higher than the corresponding death rate in the EA lineage. The confidence interval of the EU rate multiplier does not contain one, indicating that EA and EU lineages have different death rates of the IS*6110* transposon.

Since this is a novel result that has implications for monitoring tuberculosis with molecular genotyping, we examine the difference in death rates between the three lineages more closely. More specifically, we add two binary covariates to our log-linear model: *M. tuberculosis* drug resistance (DR) and HIV infection status of each patient (HIV^{+}). Our new model for birth and death rates becomes

$$\log \lambda_p = \gamma_{\lambda,1} + \gamma_{\lambda,2}\,\mathrm{EU}_p + \gamma_{\lambda,3}\,\mathrm{IND}_p + \gamma_{\lambda,5}\,\mathrm{DR}_p + \gamma_{\lambda,4}\,\mathrm{HIV}_p^{+},$$

$$\log \mu_p = \gamma_{\mu,1} + \gamma_{\mu,2}\,\mathrm{EU}_p + \gamma_{\mu,3}\,\mathrm{IND}_p + \gamma_{\mu,5}\,\mathrm{DR}_p + \gamma_{\mu,4}\,\mathrm{HIV}_p^{+},$$

where $\mathrm{DR}_p = 1$ if patient p is infected with a drug resistant strain *M. tuberculosis* and 0 otherwise, and $\mathrm{HIV}_p^{+} = 1$ if patient p is infected with HIV and 0 otherwise. Parameter estimates of this full model and their corresponding 95% confidence intervals are reported in the full model columns of Table 2. The HIV infection and drug resistance appear to have no effect on the birth

*Results of the two log-linear models for birth and death rates of IS6110. The lineage model includes only effects of M. tuberculosis lineages (EA, EU, IND). The full model combines the effects of lineages, HIV infection status (*HIV$^+$*) and drug resistance status (DR). The birth and death rate multiplier estimates for the EU lineage are highlighted in bold to indicate that the confidence intervals for these parameters are above one*

		Lineage model		Full model	
Coefficient		MLE	CIs	MLE	CIs
EA birth rate,	$\exp(\gamma_{\lambda,1})$	0.011	(0.003, 0.034)	0.012	(0.006, 0.025)
EU multiplier,	$\exp(\gamma_{\lambda,2})$	2.63	(0.689, 10.0)	**3.2**	**(1.1, 9.4)**
IND multiplier,	$\exp(\gamma_{\lambda,3})$	1.40	(0.229, 8.53)	1.7	(0.29, 9.7)
DR multiplier,	$\exp(\gamma_{\lambda,4})$	–	–	0.88	(0.36, 2.1)
HIV$^+$ multiplier,	$\exp(\gamma_{\lambda,5})$	–	–	0.61	(0.28, 1.3)
EA death rate,	$\exp(\gamma_{\mu,1})$	0.004	(0.0005, 0.028)	0.004	(0.0005, 0.031)
EU multiplier,	$\exp(\gamma_{\mu,2})$	**9.32**	**(1.19, 72.8)**	**11**	**(1.2, 114)**
IND multiplier,	$\exp(\gamma_{\mu,3})$	5.40	(0.553, 52.6)	6.2	(0.36, 1.1)
DR multiplier,	$\exp(\gamma_{\mu,4})$	–	–	1.1	(0.52, 2.3)
HIV$^+$ multiplier,	$\exp(\gamma_{\mu,5})$	–	–	0.64	(0.36, 1.1)

and death rates of IS*6110* transposon. IS*6110* copy number variation may have an impact on functions of neighboring genes in the *M. tuberculosis* genome [Alonso et al. (2011)]. Therefore, IS*6110* copy number can potentially interact with other *M. tuberculosis* phenotypes, such as drug resistance and adaptation to HIV and antiviral treatment, with the help of selection [McEvoy et al. (2007)]. However, we do not expect to see association between IS*6110* copy number and *M. tuberculosis* phenotypes within one patient because selection is unlikely to play a role on such a short time scale. Hence, we view our estimated small effects of HIV infection and drug resistance on IS*6110* copy number as biologically plausible. The EU lineage effect on the death rate remains statistically significant even after controlling for the two additional covariates. Interestingly, the EU lineage effect on the birth rate also becomes statistically significant in the full model. Effect sizes for both birth and death rates increase and the confidence intervals include larger values in the full model over the lineage-only model. This indicates that the full model tends to find more differences in rates between the lineages than the lineage-only model does. While more data are certainly needed to confirm that EU lineage birth rate effect is not 1, the full model may be capturing information the simpler lineage-only model does not, which, in the face of limited data, is valuable. For practical considerations, the fact that our most parameter rich full model results in significant effects of EU lineage on IS*6110* birth and death rates suggests that *M. tuberculosis* lineage has to be taken into consideration when IS*6110* genotype data are used to uncover the history of *M. tuberculosis* transmission.
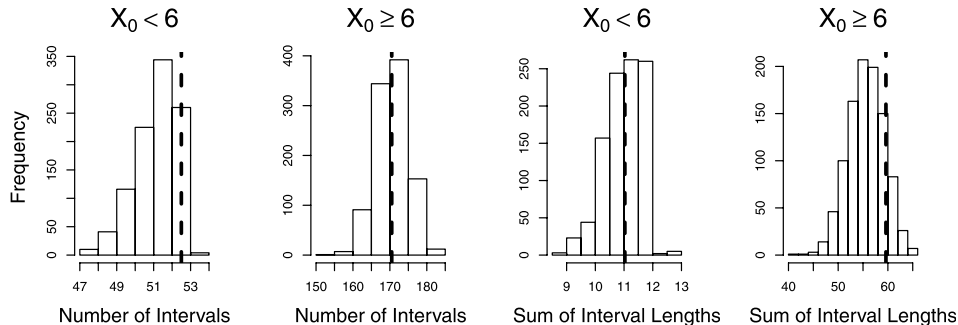
FIG. 3. *Low- vs high-count genotype analysis. Histograms of simulated numbers of intervals and sums of interval lengths are plotted for intervals with starting values less than six and greater or equal to six. The vertical dashed lines indicate the observed values of the four statistics.*

4.3.2. *IS6110 counts.* The initial number of IS*6110* elements is a potential confounder in our analysis because patients infected with Euro-American and East-Asian differ drastically in the number of IS*6110* elements at the beginning of the observation period. The isolates from the Euro-American lineage have between 2 and 17 IS*6110* elements, with 41 out of 109 patients having the first recorded IS*6110* count less than 6, while IS*6110* counts vary between 6 and 22 for the East-Asian isolates. Warren et al. (2002) suggest that IS*6110* genotypes with fewer than six elements have a very low rate of change, because in their data cases with no observed changes in the genotype are dominated by such low-count genotypes. However, our birth–death model very well predicts the conclusion of Warren et al. (2002) that low-count genotypes evolve slower than high-count genotypes. To demonstrate this, we simulate 1000 data sets using our global birth and death rates and observed initial IS*6110* counts for each patient. We record the number of intervals with equal starting and ending values less than six, $n_{0,<6}$, and equal starting and ending values greater or equal to six, $n_{0,\geq6}$. We also recorded the length sum of both kinds of intervals: $t_{0,<6}$ and $t_{0,\geq6}$. In our data, $n_{0,<6}^{\mathrm{obs}} = 53$ and $n_{0,\geq6}^{\mathrm{obs}} = 171$ with $n_{0,<6}^{\mathrm{obs}}/t_{0,<6}^{\mathrm{obs}} = 4.6 > 2.8 = n_{0,\geq6}^{\mathrm{obs}}/t_{0,\geq6}^{\mathrm{obs}}$, in agreement with Warren et al.'s (2002) analysis. Histograms of simulated values of the four statistics, $n_{0,<6}$, $n_{0,\geq6}$, $t_{0,<6}$ and $t_{0,\geq6}$, shown in Figure 3, demonstrate that our birth–death model replicates well the observed dynamics of low-count and high-count IS*6110* genotypes. We conclude that our data do not provide evidence that evolutionary dynamics of low-count genotypes differ from high-count genotype dynamics. Therefore, it is unlikely that a high percentage of low-count genotypes in the Euro-American lineage isolates causes our estimated discrepancy between death rates of Euro-American and East-Asian *M. tuberculosis* lineages.

5. Discussion. In this paper we present a novel EM algorithm for fitting birth–death processes to panel data. We allow logarithms of birth and death rates to be linear combinations of individual-level covariates. Such birth–death models with covariates share analogy with covariate-dependent CTMC models on finite state spaces—a widely used class of models in medical statistics [Kalbfleisch and Lawless (1985)]. To our knowledge, there is no established and well tested method for fitting birth–death processes, considered in this paper, to panel data. We hope that by filling this void with our new EM algorithm, accompanied by an open-source R package DOBAD (available at http://cran.r-project.org), we will stimulate statistical applications of birth–death processes, at least in the context of panel data.

We illustrate the applicability of birth–death models by analyzing the evolutionary dynamics of the IS*6110* transposon—an important genetic marker that serves as a genetic signature of the *M. tuberculosis* bacterium. By building realistic models for IS*6110* dynamics, we uncover differences in IS*6110* birth and death rates among major lineages of *M. tuberculosis*, while controlling for other clinical covariates. This novel result is important because IS*6110* copy number is used as a genetic marker to create DNA fingerprints of *M. tuberculosis* using the restriction fragment length polymorphism technology [van Embden et al. (1993), Kato-Maeda, Metcalfe and Flores (2011)]. Strains that have the same IS*6110* counts and in which the IS*6110* element is located in DNA fragments of similar size are considered identical. When such identical strains are found in community-based studies, the strains are clustered and patients carrying these strains are inferred to belong to the same *M. tuberculosis* transmission chain [Kato-Maeda, Metcalfe and Flores (2011)]. However, if some *M. tuberculosis* lineages evolve at much slower rates than others, as we discover in our analysis, then using the same notion of similarity between IS*6110* counts for these slow-evolving lineages could be highly misleading. Therefore, we suggest that when using IS*6110* genotypes, *M. tuberculosis* lineage effect should be included explicitly in statistical protocols of estimating tuberculosis epidemiological clusters.

Although in our *M. tuberculosis* fingerprinting example we do not consider the possibility of immigration, we include immigration in our methodological developments. More specifically, our EM algorithm and the accompanying software package allow for immigration to occur at a rate proportional to the birth rate. We have two reasons for including this generalization. First, this limited form of immigration complicates neither our mathematical developments nor computational tractability of the EM algorithm. Second, incorporating immigration makes our EM algorithm more transferable to other domains of application of birth–death processes. For example, our methodological developments directly apply to modeling the evolution of insertions and deletions in molecular sequences, where immigration is needed to prevent molecular sequences contracting to length zero [Thorne, Kishino and

Felsenstein ([1991](#)), Holmes ([2005](#))]. Moreover, as we show in Appendix C, for this particular application, the E-step of our EM algorithm is available in closed form, eliminating the need for numerical integration [Doss et al. ([2013](#))]. Another example of potential transferability of our EM algorithm is for hidden death-immigration models for recurrent medical conditions, such as that considered by Crespi, Cumberland and Blower ([2005](#)). Although our EM algorithm does not apply directly to the application these authors consider, because the states of the immigration-death process are only partially observed at discrete time points, our mathematical results remain useful here. More specifically, one can use our mathematical developments in the context of continuous-time hidden Markov models [Roberts and Ephraim ([2008](#))] in order to develop an EM algorithm, akin to a classical Baum–Welch algorithm [Baum et al. ([1970](#))]. As in the aforementioned insertion-deletion model, Appendix C demonstrates that the expectations of complete data sufficient statistics for the death-immigration model are available in closed form [Doss et al. ([2013](#))]. We note that because our Theorem [1](#) applies to general linear BDI models, we are able to use this theorem to study properties of a death-immigration model, which is not a BDRI model—the main focus of this manuscript.

Finally, we would like to point out that the generating functions derived in Theorem [1](#) are useful not only for developing EM algorithms for birth–death models, but also for probabilistic characterization of birth–death trajectories in general. For example, we are not aware of analytic formulae for expectations of the sufficient statistics that do not involve the ending state of the process at time t: $E(N_t^+|X_0 = i)$, $E(N_t^-|X_0 = i)$ and $E(R_t^+|X_0 = i)$. These expectations, useful for prediction purposes, arise analytically from the generating functions in Theorem [1](#) [e.g., $E(N_t^+|X_0 = i) = \partial H_i(u, 1, 0, 1, t)/\partial u|_{u=1}$].

SUPPLEMENTARY MATERIAL

Further mathematical details (DOI: [10.1214/13-AOAS673SUPP](#); .pdf). Appendices referenced in Sections [2](#) and [5](#) are available in the supplementary material [Doss et al. ([2013](#))].

REFERENCES

Alonso, H., Aguilo, J. I., Samper, S., Caminero, J. A., Campos-Herrero, M. I., Gicquel, B., Brosch, R., Martín, C. and Otal, I. (2011). Deciphering the role of IS*6110* in a highly transmissible *Mycobacterium tuberculosis* Beijing strain, GC1237. *Tuberculosis* **91** 117–126.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171. [MR0287613](#)

CATTAMANCHI, A., HOPEWELL, P. C., GONZALEZ, L. C., OSMOND, D. H., MASAE, KAWA-
    MURA, L., DALEY, C. L. and JASMER, R. M. (2006). A 13-year molecular epidemiolog-
    ical analysis of tuberculosis in San Francisco. *The International Journal of Tuberculosis
    and Lung Disease* **10** 297–304.
CRESPI, C. M., CUMBERLAND, W. G. and BLOWER, S. (2005). A queueing model
    for chronic recurrent conditions under panel observation. *Biometrics* **61** 193–198.
    MR2135860
DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from
    incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **39** 1–38.
    MR0501537
DORMAN, K. S., SINSHEIMER, J. S. and LANGE, K. (2004). In the garden of branching
    processes. *SIAM Rev.* **46** 202–229 (electronic). MR2114452
DOSS, C. R., SUCHARD, M. A., HOLMES, I., KATO-MAEDA, M. and MININ, V. N. (2013).
    Supplement to "Fitting birth–death processes to panel data with applications to bac-
    terial DNA fingerprinting." DOI:10.1214/13-AOAS673SUPP.
GAGNEUX, S., DERIEMER, K., VAN, T., KATO-MAEDA, M., DE JONG, B. C.,
    NARAYANAN, S., NICOL, M., NIEMANN, S., KREMER, K., GUTIERREZ, M. C.,
    HILTY, M., HOPEWELL, P. C. and SMALL, P. M. (2006). Variable host-pathogen com-
    patibility in *Mycobacterium tuberculosis*. *Proc. Natl. Acad. Sci. USA* **103** 2869–2873.
GIBSON, G. J. and RENSHAW, E. (1998). Estimating parameters in stochastic compart-
    mental models using Markov chain methods. *IMA Journal of Mathematics Applied in
    Medicine & Biology* **15** 19–40.
GOLINELLI, D. (2000). Bayesian inference in hidden stochastic population processes. Ph.D.
    thesis, Univ. Washington, Seattle, WA.
GUTTORP, P. (1995). *Stochastic Modeling of Scientific Data*. Chapman & Hall, London.
    MR1358359
HENRICI, P. (1979). Fast Fourier methods in computational complex analysis. *SIAM Rev.*
    21 481–527. MR0545882
HOLMES, I. (2005). Using evolutionary expectation maximization to estimate indel rates.
    Bioinformatics **21** 2294–2300.
HOLMES, I. and RUBIN, G. M. (2002). An expectation maximization algorithm for training
    hidden substitution models. *Journal of Molecular Biology* **317** 753–764.
JACKSON, C. H. (2011). Multi-state models for panel data: The msm package for R.
    Journal of Statistical Software **38** 1–29.
JASMER, R. M., HAHN, J. A., SMALL, P. M., DALEY, C. L., BEHR, M. A., MOSS, A. R.,
    CREASMAN, J. M., SCHECTER, G. F., PAZ, E. A. and HOPEWELL, P. C. (1999).
    A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991–1997.
    Annals of Internal Medicine **130** 971–978.
KALBFLEISCH, J. D. and LAWLESS, J. F. (1985). The analysis of panel data under a
    Markov assumption. *J. Amer. Statist. Assoc.* **80** 863–871. MR0819585
KARLIN, S. and MCGREGOR, J. (1958). Linear growth birth and death processes. *J. Math.
    Mech.* **7** 643–662. MR0098435
KATO-MAEDA, M., METCALFE, J. Z. and FLORES, L. (2011). Genotyping of Mycobac-
    terium tuberculosis: Application in epidemiologic studies. *Future Microbiol.* **6** 203–216.
KEIDING, N. (1975). Maximum likelihood estimation in the birth-and-death process. *Ann.
    Statist.* **3** 363–372. MR0362773
KENDALL, D. G. (1948). On the generalized "birth-and-death" process. *Ann. Math.
    Statist.* **19** 1–15. MR0024091
LANGE, K. (1982). Calculation of the equilibrium distribution for a deleterious gene by
    the finite Fourier transform. *Biometrics* **38** 79–86.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57** 425–437. MR1323348

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **44** 226–233. MR0676213

McEvoy, C. R. E., Falmer, A. A., van Pittius, N. C. G., Victor, T. C., van Helden, P. D. and Warren, R. M. (2007). The role of IS*6110* in the evolution of *Mycobacterium tuberculosis*. *Tuberculosis* **87** 393–404.

Minin, V. N. and Suchard, M. A. (2008). Counting labeled transitions in continuous-time Markov models of evolution. *J. Math. Biol.* **56** 391–412. MR2358440

Nee, S. (2006). Birth–death models in macroevolution. *Annual Review of Ecology, Evolution, and Systematics* **37** 1–17.

Roberts, W. J. J. and Ephraim, Y. (2008). An EM algorithm for ion-channel current estimation. *IEEE Trans. Signal Process.* **56** 26–33. MR2439810

Rosenberg, N. A., Tsolaki, A. G. and Tanaka, M. M. (2003). Estimating change rates of genetic markers using serial samples: Applications to the transposon IS*6110* in Mycobacterium tuberculosis. *Theoretical Population Biology* **63** 347–363.

Sehl, M., Zhou, H., Sinsheimer, J. S. and Lange, K. L. (2011). Extinction models for cancer stem cell therapy. *Math. Biosci.* **234** 132–146. MR2907020

Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., Schecter, G. F., Daley, C. L. and Schoolnik, G. K. (1994). The epidemiology of tuberculosis in San Francisco. A population-based study using conventional and molecular methods. *New England Journal of Medicine* **330** 1703–1709.

Suchard, M. A., Lange, K. and Sinsheimer, J. S. (2008). Efficiency of protein production from mRNA. *J. Stat. Theory Pract.* **2** 173–182. MR2524460

Tanaka, M. M. and Rosenberg, N. A. (2001). Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. *Stat. Med.* **20** 2409–2420.

Thorne, J. L., Kishino, H. and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33** 114–124.

van Embden, J. D., Cave, M. D., Crawford, J. T., Dale, J. W., Eisenach, K. D., Gicquel, B., Hermans, P., Martin, C., McAdam, R., Shinnick, T. M. et al. (1993). Strain identification of Mycobacterium tuberculosis by DNA fingerprinting: Recommendations for a standardized methodology. *J. Clin. Microbiol.* **31** 406–409.

Warren, R. M., van der Spuy, G. D., Richardson, M., Beyers, N., Booysen, C., Behr, M. A. and van Helden, P. D. (2002). Evolution of the IS6110-based restriction fragment length polymorphism pattern during the transmission of Mycobacterium tuberculosis. *J. Clin. Microbiol.* **40** 1277–1282.

C. R. Doss
V. N. Minin
Department of Statistics
University of Washington
Seattle, Washington 98195
USA
E-mail: cdoss@uw.edu
        vminin@uw.edu

I. Holmes
Department of Bioengineering
  and Biophysics Graduate Group
University of California, Berkeley
Berkeley, California 94720
USA
E-mail: ihh@berkeley.edu

M. A. Suchard
Departments of Biomathematics,
  Biostatistics and Human Genetics
University of California, Los Angeles
Los Angeles, California 90095
USA
E-mail: msuchard@ucla.edu

M. Kato-Maeda
Department of Medicine
San Francisco General Hospital
University of California, San Francisco
San Francisco, California 94143
USA
E-mail: Midori.Kato-Maeda@ucsf.edu