Inferring HIV Transmission Patterns from Viral Deep-Sequence Data via Latent Spatial Poisson Processes

Fan Bu[1,3], Oliver Ratmann[2], and Jason Xu[3]

[1]Department of Biostatistics, University of California, Los Angeles
[2]Department of Mathematics, Imperial College London
[3]Department of Statistical Science, Duke University

Abstract

Viral deep-sequencing technologies play a crucial role toward understanding disease transmission network flows, because the higher resolution of these data compared to standard Sanger sequencing provide evidence into the direction of infectious disease transmission. To more fully utilize these rich data and account for the uncertainties in phylogenetic analysis outcomes, we propose a spatial Poisson process model to uncover HIV transmission flow patterns at the population level. We represent pairings of two individuals with viral sequence data as typed points, with coordinates representing covariates such as sex and age, and the point type representing the unobserved transmission statuses (linkage and direction). Points are associated with observed scores on the strength of evidence for each transmission status that are obtained through standard deep-sequenece phylogenetic analysis. Our method is able to jointly infer the latent transmission statuses for all pairings and the transmission flow surface on the source-recipient covariate space. In contrast to existing methods, our framework does not require pre-classification of the transmission statuses of data points, instead learning them probabilistically through a fully Bayesian inference scheme. By directly modeling continuous spatial processes with smooth densities, our method enjoys significant computational advantages compared to previous methods that rely on discretization of the covariate space. We demonstrate that our framework can capture age structures in HIV transmission at high resolution, and bring valuable insights in a case study on viral deep-sequencing data from Southern Uganda.

1 Introduction

As a decades-long global pandemic, the human immunodeficiency virus (HIV) has most severely affected Africa with 1 in every 25 adults living with the HIV virus, accounting for more than two-thirds of infections worldwide (Eisinger and Fauci, 2018; Fauci and Lane, 2020). International public health organizations have proposed to combat HIV by targeting intervention efforts toward high-risk populations. To achieve this, it is important to understand and characterize transmission patterns between different groups and sub-populations, which requires the inference of mostly unobserved transmission flows between infected individuals. Facilitated by recent advances in phylogenetic analysis technologies for processing viral RNA sequences collected from persons living HIV, researchers have gained insights about individual-level transmission links, as well as population-level transmission flows and patterns

(Romero-Severson et al., 2016; Leitner and Romero-Severson, 2018; Rasmussen et al., 2018; Ratmann et al., 2020; Bbosa et al., 2020; Scire et al., 2020; Hall et al., 2021; Zhang et al., 2020). Such phylogenetic analysis outputs, when jointly analyzed with demographic information from population-based studies, continue to improve our understanding of HIV transmission structure between high-risk groups. However, such joint analysis is challenging as phylogenetic inference outputs and demographic information take different data modalities, and there is high uncertainty associated with phylogenetic analysis that is difficult to quantify statistically. Most existing approaches often neglect the phylogenetic analysis uncertainty by arbitrarily thresholding the outputs, and also discretize on demographic covariates which can lead to reduced inference resolutions and heavy computational burdens. It is thus desirable to develop a principled statistical approach that respects and accounts for uncertainty, maintains the high resolution of demographic information, and improves on scalability.

The overarching goal of this paper is to develop such statistical methods to better leverage both phylognetic analyses and population demographics toward understanding the age structure of heterosexual HIV transmissions using a fully stochastic generative model. In particular, we seek to infer the relative densities of transmissions between different age groups from demographic information together with phylogenetic analysis summaries of viral deep-sequencing data. Naturally, the direction of events, i.e., "who infected whom", informs our understanding of the age-structured transmission patterns, but these events are not observable in practice. Instead, they can be indirectly informed by the viral sequences sampled from infected individuals. Deep-sequence phylogenetic analysis tools such as *Phyloscanner* (Wymant et al., 2018) or *QUENTIN* (Skums et al., 2018) compare the HIV molecular genetic diversity in virus variants of each individual sampled and generate probabilistic summaries about the transmission relationship between pairs of deep-sequenced individuals — how likely that they shared a transmission link, and how probable that one infected the other or vice versa. Using such probabilistic summary statistics, we model the underlying ground truth transmission structures as latent surfaces in a plane whose axes denote covariates of the individuals, and we focus here on the ages of the source and recipient, respectively. In doing so, we represent the transmission structure as an interpretable and continuous latent "spatial" variable.

This scientific task can be cast as the challenge of making inference on the latent transmission structure from marked points. Each point corresponds to a pairing of two individuals, and here we focus on describing the coordinates of the point in terms of the sex and continuous age of the two individuals. The "type" (or "mark") of the point describes whether transmissions occurred between pairs and the direction of transmission. The type of each point is unknown but phylogenetic data provides strength of evidence for each type. Under this perspective, hierarchical spatial Poisson process models on typed point patterns provide a natural choice for inference of population-level transmission flows. A key advantage of our point-pattern approach lies in its use of continuous spatial surfaces, which is able to preserve richer information and underlying patterns that a discretized representation may obscure approach. Moreover, our continuous formulation avoids the need to keep track of discrete grids based on pre-specified age groups (such as 1-year or 5-year age groups), a common choice in epidemiological studies (e.g., Hyman et al. (1994); Heuveline (2004); Sharrow et al. (2014)). Such discretization requires heuristic choices *a priori* and leads to computationally intensive downstream analysis. Compared to recent work in Xi et al. (2022) that introduced a semi-parametric Poisson flow model on discrete age strata and thus tracks a discrete latent grid of

transmission intensities, our continuous modeling approach significantly improves scalability, and also offers the user flexibility to discretize or summarize output at any resolution post-hoc. Another important advantage of our model over existing work is that it no longer requires specifying ad hoc thresholds on the phylogenetic data summaries because the model infers the unknown linkage and transmission direction status of each pairing simultaneously. As a result, we appropriately propagate the uncertainty in transmission links and directions as captured in the phylogenetic scores.

Prior work Spatial Poisson process models have been widely applied to the study of point-referenced two-dimensional data (Banerjee et al., 2003; Huber, 2011; Cressie, 2015). Heterogeneity in spatial point patterns is often modelled through non-homogeneous Poisson processes (NHPPs), where the structure of the intensity function can be described using various choices of Bayesian mixture models. NHPP intensity functions have been parameterized through Markov random fields for piecewise constant functions based on Voronoi tessellations (Heikkinen and Arjas, 1998), weighted Gamma process mixtures of non-negative kernel functions (Lo and Weng, 1989; Wolpert and Ickstadt, 1998; Ishwaran and James, 2004), Gaussian process mixtures of log-transformed components (Møller et al., 1998; Brix and Diggle, 2001; Adams et al., 2009), and Dirichlet process mixtures (Ji et al., 2009; Zhou et al., 2015; Zhao and Kottas, 2021). Kim and Kottas (2022) offer an excellent summary. Our framework builds on previous developments in Dirichlet process mixtures that focus on learning a normalized functional form of the NHPP intensity (Kottas and Sansó, 2007; Kottas et al., 2008; Taddy et al., 2012), which we will see admits tractable and efficient inference procedures.

In recent years, Poisson process models have been extended to study spatial point patterns that are latent or partially observed, in that only certain indirect "signals" associated with the underlying spatial pattern are observed, with uncertainty about the quantity and locations of the latent points (Vedel Jesen and Thorarinsdottir, 2007; Ji et al., 2009). Joint modelling of the "signals" and the latent point process through Bayesian data augmentation has been successful, thanks to the ease of incorporating missing information as latent variables in a Bayesian inference framework (Givens et al., 1997). However, to our knowledge, much of the existing work in spatial Point process models focuses on *one* set or type of spatial points, rather than a combination of multiple types of latent point patterns whose "types" are associated with practical interpretation, but are not observed. Our framework bridges this methodological gap by exploiting "signals" that inform the latent type labels in a marked spatial Poisson process model.

Our continuous spatial process approach in modelling disease transmissions marks a contrast to the body of work that relies on count data in discretized spatial areas (Berke, 2004; Best et al., 2005; Wakefield, 2007; Gschlößl and Czado, 2008; Mohebbi et al., 2014; Bauer et al., 2016; Johnson et al., 2019). These discrete or areal spatial models often involve the use of Gaussian Markov random fields (GMRF) (Rue and Held, 2005) or other Gaussian-based models to handle spatial dependence structures. Due to the high computational cost associated with Gaussian covariance matrix operations, inference has to be performed through numerical approximation techniques such as integrated nested Laplace approximations (INLA) (Rue et al., 2009) and still entails intensive computation. These computational challenges are inevitable when only aggregate data are available (Gschlößl and Czado, 2008; Mohebbi et al., 2008), but given access to point-level data, formulating a continuous spatial model becomes more appealing than a discretized one. This is because (1) directly modeling a continuous

spatial process avoids expensive spatial smoothing techniques such as the GMRF, and (2) a continuous point pattern surface can induce a discrete spatial model at any desired resolution, avoiding the need for manual discretization before analysis (van de Kassteele et al., 2017; Xi et al., 2022). In the context of our application, where age is known for every surveyed individual and can be treated as a continuous variable, employing a continuous spatial model provides a more natural and efficient approach.

In the rest of this paper, we first provide an overview of the motivating data and develop a model framework in Section 2. Our likelihood-based Poisson process model is outlined in Section 3. In Section 4, we show results from simulation experiments, and in Section 5 we present an application to large-scale HIV deep sequence data from the Rakai Community Cohort Study in Southern Uganda that was collected between 2010 to 2015. Section 6 presents our conclusions.

2 Data and Model

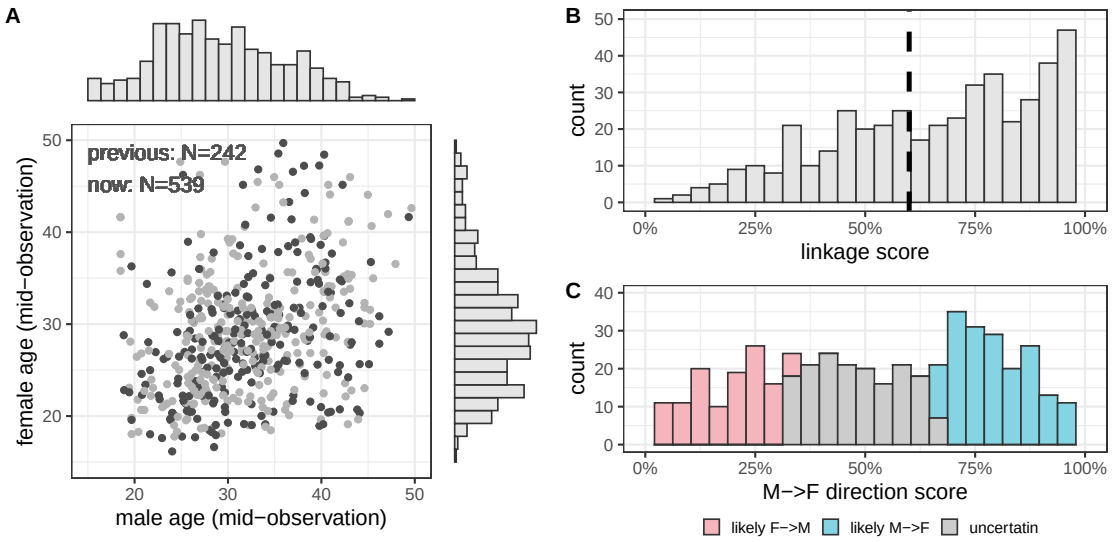

Figure 1: The data. Panel A: paired ages of heterosexual patients that are considered potential HIV transmission pairs; our method aims to analyzing *all* 539 potential transmission pairs (both gray and dark dots), whereas previous analysis with existing methods would pre-exclude "low-confidence" pairs and only include 242 pairs (in dark dots). Panels B and C: the linkage scores (ℓ_i's) and direction scores (d_i's) for the paired individuals, as pre-processing results from phylogenetic analysis using *phyloscanner*; previous analysis would only consider pairs with linkage scores $> 60\%$ (indicated by dark dashed line in panel B) and direction scores below 33.3% or above 66.7% while discarding pairs with scores in the "uncertain" region.

HIV deep-sequencing data were obtained from blood samples of study participants with HIV of the longitudinal, open, population-based Rakai Community Cohort Study in Southern Uganda between August 2011 and January 2015 (Ratmann et al., 2019). A total of 25,882 individuals participated during the study period, among whom 5142 were HIV seropositive. Based on exclusion criteria on HIV viral load, viral sequence read depth and length (Wymant et al.,

2018; Ratmann et al., 2019), virus from 2652 participants was deep-sequenced. Detailed demographic, behavioural and healthcare data are available for all participants, including their sex and exact age. The cohort study also obtained data on names of cohabitating sexual partners which can be used to validate sexual interactions that could contribute to disease transmission.

The data that motivate our model consist of a subset of 539 heterosexual pairs of participants considered as potential HIV transmission pairs based on phylogenetic evidence extracted using Phyloscanner and epidemiological data (see illustration in Figure 1). Although these pairs are considered as likely transmission pairs, there is high uncertainty about the existence of transmission links and directions between each pair of individuals, and there is substantial variability in the strengths of phylogenetic evidence in support of pairwise transmission relations. There are two main aspects of the data: the first facet consists of the set of these pairs of individuals. This set is represented by their respective ages, denoted $\mathbb{S} = \{\mathbf{s}_i = (a_{i1}, a_{i2})^T\}_{i=1}^N$, where the 2-dimensional vector (a_{i1}, a_{i2}) records the male's age a_{i1} and female's age a_{i2} in the ith pair. In our application, sample size $N = 539$. Our model will later envision these points of paired ages as an observation from a spatial process describing the transmission structure.

The second aspect consists of two scores (in the range of $[0, 1]$) that are outputs from phylogenetic analysis of HIV deep-sequencing data. For each pair i of heterosexual individuals, the phylogenetic software *phyloscanner* produces two scores — a linkage score and a direction score — by statistically comparing the patristic distances and topological configurations of the viral reads in deep-sequence phylogenies, repeatedly over a large set of sliding, overlapping genomic windows across the HIV genome (Ratmann et al., 2019). The linkage score ℓ_i represents the posterior probability of the pair sharing a transmission link in the transmission process under a Binomial count model of window-specific linkage classifications (Ratmann et al., 2019). The direction score d_i, on the other hand, measures the posterior probability of transmission taking place from the male to the female in this pair under a similar count model. We collectively denote the phylogenetic scores for the ith pair by $\mathbf{x}_i = (\ell_i, d_i)^T$.

Our goal is to make inference about the population-level transmission flows during the observation period in the study communities by age and sex strata, using the combined demographic and deep-sequence phylogenetic data. However, the linkage and direction scores produced by deep-sequence phylogenetic analysis do not directly classify the data points. They instead reflect uncertainties about individual transmission events; we therefore model $\mathbf{x}_i$ as a "signal" that comes with each point $\mathbf{s}_i$, and call the set $\{\mathbf{x}_i | i = 1, \ldots, N\}$ the "marks" associated with the point pattern. Whether a transmission occurs between each pair and the direction of the potential transmission are unknown, and will be accounted for as latent variables that connect $\mathbf{x}_i$ and $\mathbf{s}_i$ in the stochastic model described below. Doing so accounts for these uncertainties in a more statistically principled way, departing from conventions in existing approaches that essentially filter for high-confidence pairs only, which not only results in deterministic *a priori* choices, but entails throwing away much of the data.

2.1 Stochastic model framework

We now introduce our model for the point data $\mathbb{S} = \{\mathbf{s}_i = (a_{i1}, a_{i2})^T\}_{i=1}^N$ with associated phylogenetic transmission and direction scores (i.e., marks) $\mathbf{x}_i = (\ell_i, d_i)^T$. To each point we introduce a categorical latent "type" c_i that corresponds to three possible events: transmission did not occur between the two individuals that define the point (denoted by $c_i = 0$), trans-

mission occurred from the male to the female individual (denoted by $c_i = 1$), or transmission occurred from the second to the first individual (denoted by $c_i = -1$). Since the marks $\mathbf{x}_i$ partially inform the likelihood of transmissions and their directions, it is natural to consider a joint model for the point pattern $\mathbb{S} = \{(a_{i1}, a_{i2})^T\}_{i=1}^N$ and observed signals $\{\mathbf{x}_i\}_{i=1}^N$ connected through c_i. Our framework can be thought of as a marked spatial Poisson process, where the mark distribution at each point location depends on the latent type label of each observed point. In the rest of this subsection, we outline two key components of this framework: a spatial Poisson process with a density function modelled by Dirichlet process Gaussian mixtures, and a type-dependent distribution for the marks. We note that this framework, though motivated by our application, can be applied to many similar data sets of spatial patterns with latent labels and associated signals.

Suppose we have a spatial point pattern defined on a 2-dimensional space $\mathcal{S} \times \mathcal{S}$, where $\mathbb{S} = \{\mathbf{s}_i\}_{i=1}^N$ is the set of all points, and each point $\mathbf{s}_i = (s_{i1}, s_{i2})^T$ is represented as a point in this plane. The observed signal (i.e., marks) corresponding to each point $\mathbf{s}_i$ is denoted by $\mathbf{x}_i \in \mathbb{R}^d$,which follows some distribution determined by the type label c_i ($\in \mathcal{K}$) of the point $\mathbf{s}_i$. Note that in the context of our application, $d = 2$ as we consider phylogenetic scores. In the context of our application, $\mathcal{S}$ is the continuous age of individuals under study, $\mathcal{S} = [15, 50)$; each point $\mathbf{s}_i = (a_{i1}, a_{i2})^T$ corresponds to the ages of the two individuals forming a pair, ordered by gender; and the latent types are $\mathcal{K} = \{-1, 0, 1\}$, corresponding to female-to-male transmission, no transmission and male-to-female transmission.

I: the spatial point process We model the observed points in $\mathbb{S}$ as a realization of a 2D Poisson process on $\mathcal{S} \times \mathcal{S}$:

$$\mathbb{S} \sim PP(\boldsymbol{\lambda}). \tag{1}$$

Following Kottas and Sansó (2007) we decompose the intensity function $\boldsymbol{\lambda}$ into a scale component γ and a density function $f(\cdot)$

$$\boldsymbol{\lambda}(\cdot) = \gamma f(\cdot), \tag{2}$$

so that $f(\cdot)$ satisfies $\int_{\mathcal{S}\times\mathcal{S}} f(s_1, s_2) ds_1 ds_2 = 1$. This decomposition separates the intensity function into two terms which are simpler to write out in the likelihood function and make inference computationally tractable. We next model the density function $f(\cdot)$ as a mixture of K different "types":

$$f(\cdot) = \sum_{k\in\mathcal{K}} p_k f_k(\cdot), \tag{3}$$

where p_k is the probability of points belonging to type k, and $f_k(\cdot)$ is the spatial density function for type k. In our application, $\mathcal{K} = \{-1, 0, 1\}$ and, for instance p_1 corresponds to the proportion of male-to-female transmission events among all pairs of individuals being considered, and $f_1(\cdot)$ corresponds to the 2D function that captures the across-age transmission pattern with male sources and female recipients.

There are various choices to model the structure of the density functions $f_k(\cdot)$. To balance simplicity and flexibility, we choose a Dirichlet process (DP) Gaussian mixture model, abbreviated "DPGMM", consisting of bivariate Gaussian components $f_k(\cdot)$. Specifically, for each point $\mathbf{s}_i$, if its type label $c_i = k$, then

$$\mathbf{s}_i \mid c_i = k \sim N(\theta_{ki}, \Sigma_{ki}), \quad (\theta_{ki}, \Sigma_{ki}) \sim G_k, \quad G_k \sim DP(\alpha_k, G_0). \tag{4}$$

Here G_k represents the (infinite) mixture of bivariate normals model for type k, and θ_{ki} and Σ_{ki} are the mean vector and covariance matrix for the bivariate Gaussian component that $\mathbf{s}_i$ belongs to. In practice, Dirichlet process mixtures are often treated as a finite mixture but with a flexible number of components. Indeed, the above defined model may be expressed equivalently in the following manner in terms of each density function $f_k(\cdot)$

$$f_k(\cdot) = \sum_{h=1}^{H_k} w_{kh} \text{BVN}(\cdot; \theta_{kh}, \Sigma_{kh}), \tag{5}$$

where H_k denotes the number of "active" components, or total number of unique components generated by the DP, and each $(\theta_{kh}, \Sigma_{kh})$ is a *unique* Gaussian component for the type-k density. Here, $\text{BVN}((s_1, s_2); \theta, \Sigma)$ denotes the probability density of a bivariate normal distribution with mean θ and covariance Σ.

II: the signal distribution We next model the signal distributions that connect the observed data to the latent process described above. We view the signal $\mathbf{x}_i$ as a "mark" associated with each point $\mathbf{s}_i$ providing information on its true type c_i. Naturally, the probability distribution of $\mathbf{s}_i$ should then depend on its latent label c_i: that is, conditional on c_i, the general form of the probability density (or mass function) for $\mathbf{x}_i$ can be written as

$$p(\mathbf{x}_i \mid c_i = k) = \phi_k(\mathbf{x}_i). \tag{6}$$

This implies that our framework can be generally applied to modeling any spatial point patterns with associated signals that have signal probabilities dependent on the (latent) properties of the spatial points, as long as the density or probability function $\phi_k(\cdot)$ is well-chosen and well-defined.

Here, we adopt logit-normal distributions for the signal scores $\mathbf{x}_i = (\ell_i, d_i)^T$ defined on $(0, 1)$. Specifically, given the true labels c_i, we model the signals through

$$\text{logit}(\ell_i) \mid c_i \sim N(\tilde{\mu}_{\ell,i}, \sigma_\ell^2), \tag{7}$$

$$\text{logit}(d_i) \mid c_i \sim N(\tilde{\mu}_{d,i}, \sigma_d^2), \tag{8}$$

where

$$\tilde{\mu}_{\ell,i} = \mu_\ell \mathbb{1}\left[c_i \neq 0\right], \tag{9}$$

$$\tilde{\mu}_{d,i} = \mu_d \mathbb{1}\left[c_i = 1\right] + \mu_{-d} \mathbb{1}\left[c_i = -1\right]. \tag{10}$$

Following the construction of the linkage and direction scores, it is more probable for the linkage score ℓ_i to be larger for an actual transmission event and the direction score d_i to be larger for a male-to-female transmission. We thus effectively posit a mixture model for the scores: with $\mu_\ell > 0$, (9) implies that the linkage score ℓ_i is likely to exceed 0.5 for a real transmission event ($c_i \neq 0$); with $\mu_{-d} < 0 < \mu_d$, (10) implies that d_i is likely larger than 0.5 for a male-to-female event ($c_i = 1$) but smaller than 0.5 for a female-to-male event ($c_i = -1$). Note that such design uses the property $\text{logit}(0.5) = 0$.

2.2 The complete data likelihood

From the descriptions in the previous section, we see that the two key components in the model—the spatial process and the emission or signal distribution —are linked through the latent types c_i. A point i with true label $c_i = k$ contributes the term $\gamma p_k f_k(\mathbf{s}_i) \times \phi_k(\mathbf{x}_i)$ to the data likelihood. Here the first term $\gamma p_k f_k(\mathbf{s}_i)$ comes from the intensity of the spatial process (model for the age structure), and the second term $\phi_k(\mathbf{x}_i)$ comes from the signal distribution (model for the phylogenetic scores), both conditioned on $c_i = k$. Naturally, because all the c_i's are unknown, evaluating the likelihood function based on *observed* data only would require considering and marginalizing over all possible values of k that each c_i may take. Thus, we can instead construct the likelihood function given the "complete" data, which include the coordinates of all N points in the set $\mathbb{S}$, the observed signals $\mathbf{x}_i$ as well as the type labels c_i for all $i = 1, \ldots, N$:

$$L(\Theta; \{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\}) = \gamma^N \frac{e^{-\gamma}}{N!} \prod_{k \in \mathcal{K}} \prod_{i: c_i = k} p_k f_k(\mathbf{s}_i) \phi_k(\mathbf{x}_i) \tag{11}$$

$$= \prod_{i=1:N} \phi(x_{i1} \mid \tilde{\mu}_{\ell,i}, \sigma_\ell^2) \phi(x_{i2} \mid \tilde{\mu}_{d,i}, \sigma_d^2) \tag{12}$$

$$\times \gamma^N \frac{e^{-\gamma}}{N!} \prod_{k \in \mathcal{K}} \prod_{i: c_i = k} \left(p_k \sum_{h=1}^{H_k} w_{kh} \mathrm{BVN}((s_{i1}, s_{i2}); \theta_h, \Sigma_h) \right).$$

Here $\phi(\cdot; \mu, \sigma^2)$ is the normal p.d.f. with mean μ and variance σ^2, and the model parameters are $\Theta = \{\gamma, \mathbf{p}, \boldsymbol{\mu}, \sigma_\ell^2, \sigma_d^2, \{(\theta_{kh}, \Sigma_{kh})\}, \{\alpha_k\}\}$ (let $\mathbf{p} = (p_{-1}, p_0, p_1)^T$, $\boldsymbol{\mu} = (\mu_\ell, \mu_d, \mu_{-d})^T$). The following section discusses our approach to inference in practice when information such as c_i is missing.

3 Bayesian inference with data augmentation

We now derive an efficient Bayesian inference scheme for estimating model parameters Θ based on the likelihood function (12). Note that given known type labels c_i, we no longer need to infer the transmission link and direction from the scores, and thus the age structures for each transmission direction can be learned straightforwardly by independently estimating the density function f_k for each type k, which is standard for a Gaussian mixture model with Dirichlet process priors (Rasmussen, 1999).

When the type label c_i's are unknown, however, parameter estimation is complicated as deriving the marginal likelihood of the observed data is nontrivial. Basing inference on the marginal likelihood would entail integrating over all possible configurations of the unknown labels c_i's of all the spatial configurations. Instead of a direct approach based on the observed data likelihood, we exploit the complete data likelihood and adopt a data augmentation framework for inference. That is, we treat the unobserved labels c_i's as latent variables, and "augment" the observed data by sampling candidate values of c_i in each iteration of the sampler. These settings for c_i within each iteration of the sampler "complete" the data, enabling the use of (12) and thus rendering updates for the other parameters tractable.

With appropriate prior choices $p_0(\Theta)$ on all parameters Θ, we can derive the joint posterior

density for all unknown quantities, including parameters Θ and type labels $\{c_i\}$'s:

$$p(\Theta, \{c_i\} \mid \{\mathbf{x}_i\}, \{\mathbf{s}_i\}) \propto L(\Theta; \{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\}) p_0(\Theta). \tag{13}$$

The data-augmented inference framework is employed through a Bayesian Markov chain Monte Carlo (MCMC) sampler, which can be roughly divided into two major components in each iteration: (1) sample or update parameters Θ conditioned on configurations of the $\{c_i\}$'s from $p(\Theta \mid \{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\})$; and (2) sample c_i for each i given values of Θ from $p(c_i \mid \Theta, \mathbf{x}_i, \mathbf{s}_i)$ utilizing the factorized form of the complete data likelihood in (12).

To improve efficiency of the MCMC sampler, we prescribe conjugate or semi-conjugate priors whenever possible to enable straightforward Gibbs sampling exploiting full conditional posterior densities that exist for almost all parameters. Below we detail these prior choices and discuss each step of the MCMC sampler. We also provide a summary of our sampling algorithm in online Supporting Information (see Web Algorithm 1).

Scale parameter γ: Due to the scale decomposition in (2), sampling the parameter γ is straightforward and can in fact be done independently of the Markov chain that samples the remaining parameters. That is, if we assume a Gamma prior $\gamma \sim Ga(a_0, b_0)$, we may directly draw samples using

$$\gamma \mid \{\mathbf{x_i}\}, \{\mathbf{s_i}\} \sim Ga(\alpha_0 + N, \beta_0 + 1).$$

Signal distribution parameters $\boldsymbol{\mu}$, σ_ℓ^2, and σ_d^2: Conditioned on type labels $\{c_i\}$, the signal distributions (7) and (8) are simple 1D normal models. Assuming diffuse priors $\mu_\ell, \mu_d \sim \text{Unif}((0, \infty))$, $\mu_{-d} \sim \text{Unif}((-\infty, 0))$ and inverse-Gamma priors $\sigma_\ell^2, \sigma_d^2 \sim \text{inv-Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$, parameter updates only require straightforward Gibbs steps using the full conditionals. For example, for the linkage score parameters μ_ℓ and σ_ℓ^2, we draw

$$\mu_\ell \mid \sigma_\ell^2, \{\ell_i\}, \{c_i\} \sim N_{(0,\infty)} \left(\sum_{i:c_i \neq 0} \text{logit}(\ell_i), \sigma_\ell^2 / N_+ \right);$$

$$\sigma_\ell^2 \mid \mu_\ell, \{\ell_i\}, \{c_i\} \sim \text{inv-Gamma} \left(\frac{\nu_0 + N_+}{2}, \frac{\nu_0\sigma_0^2 + \sum_{i:c_i \neq 0} (\text{logit}(\ell_i) - \mu_\ell)^2}{2} \right).$$

Here $N_+ = \sum_{i=1}^N \mathbb{1}(c_i \neq 0)$ is the total number of real transmissions given the $\{c_i\}$ configurations, and $N_{(0,\infty)}$ denotes a normal distribution truncated on the positive real line. For the direction score parameters μ_d, μ_{-d} and σ_d^2, sampling steps are almost exactly the same.

Type probability p: Assuming a Dirichlet prior for the vector $\mathbf{p} = (p_{-1}, p_0, p_1)$, $\mathbf{p} \sim \text{Dir}(q_{-1}, q_0, q_1)^T$, given configurations for $\{c_i\}$, we can draw

$$\mathbf{p} \mid \{c_i\} \sim Dir\left(q_{-1} + N_{-1}, q_0 + N_0, q_1 + N_1\right),$$

where $N_k = \sum_{i=1}^N \mathbb{1}(c_i = k)$ is the total number of data points belonging to type k.

DP precision parameter α_k's and component weights w_{kh}'s: For the precision parameter α_k and weights w_{kh} for each type k, we adopt a truncated DP mixture model (with a large maximum number of mixtures) to approximate the (infinite) DP mixture, which is a technique described in Section 3.2 and the Appendix in Ji et al. (2009). More specifically, we use the auxiliary sampling trick introduced in Escobar and West (1995) to update the precision parameter α_k's ($k = 0, -1, 1$); here, by introducing an additional auxiliary parameter to sample

along with each α_k, the conditional posterior distribution for α_k can be reduced to a mixture of two Gamma distributions, which conveniently transforms its sampling step to a simple Gibbs step. Exact technical details are provided in Ji et al. (2009) and Escobar and West (1995).

BVN mixture components $(\theta_{kh}, \Sigma_{kh})$'s: For the bivariate normal mixture model of each type k, we introduce a component latent indicator z_i for each data point i (that belongs to type k) such that $z_i = h$ indicates point i belongs to component h. Assuming semi-conjugate priors $\theta_{kh} \sim \text{BVN}(\theta_0, \Sigma_0)$ and $\Sigma_{kh} \sim \text{inv-Wishart}(\nu, S_0)$, we then iteratively update z_i's and $(\theta_{kh}, \Sigma_{kh})$'s using

$$Pr(z_i = h \mid w_{kh}, \theta_{kh}, \Sigma_{kh}, \mathbf{s_i}) \propto w_{kh}\varphi(\mathbf{s}_i \mid \theta_{kh}, \Sigma_{kh});$$

$$\theta_{kh} \mid \Sigma_{kh}, \{z_i\}, \{\mathbf{s_i}\} \sim \text{BVN}\left((m_h\Sigma_{kh}^{-1} + \Sigma_0^{-1})^{-1}\left(\sum_{i:z_i=h}\Sigma_{kh}^{-1}\mathbf{s_i} + \theta_0\Sigma_0^{-1}\theta_0\right), (m_h\Sigma_{kh}^{-1} + \Sigma_0^{-1})^{-1}\right);$$

$$\Sigma_{kh} \mid \theta_{kh}, \{z_i\}, \{\mathbf{s_i}\} \sim \text{inv-Wishart}\left(\nu + m_h, \left(S_0^{-1} + \sum_{i:z_i=h}(\mathbf{s_i} - \theta_{kh})(\mathbf{s_i} - \theta_{kh})^T\right)^{-1}\right).$$

Here $\varphi(\cdot \mid \theta, \Sigma)$ is the density function of a bivariate normal with mean θ and covariance matrix Σ, and $m_h = \sum_{i=1}^{N} \mathbb{1}(z_i = h)$ is the total number of data points belonging to spatial component h.

Type labels c_i's: The type label c_i can be sampled for each data point i conditioned on all other parameter values via $Pr(c_i = k \mid \Theta) \propto p_k f_k(\mathbf{s}_i)\phi_k(\mathbf{x}_i)$. A pseudocode summary of these updates appears in the Supplement.

4 Simulation studies

In this section, we validate the model framework through simulation experiments. In particular, we assess whether the inferential procedure can successfully identify different underlying patterns of HIV transmission flows, in terms of both gender and age structures. Moreover, we explore the power and precision of the identification of underlying patterns given various sample sizes.

We focus on comparing two pairs of different scenarios that are of epidemiological interest:

1. proportions of male-to-female (MF) and female-to-male (FM) transmission events. One general finding from HIV transmission flow studies is that there tends to be more male-to-female than female-to-male transmissions(Hall et al., 2021; Bbosa et al., 2020; Ratmann et al., 2020), and thus it would be important for a modeling framework to recognize such a pattern if there is indeed a difference in the proportions of transmission directions. We consider two scenarios here: (1) "**MF 50-50**", where male-to-female (MF) and female-to-male (FM) events take up equal proportions among all the real transmissions; (2) "**MF 60-40**", where MF transmissions occur more frequently and constitute about 60% of all transmissions. Again, relevant model parameters for each scenario are detailed in the Online Supporting Information.

2. age structure of male sources for infections in young women. There is a general interest in the age distribution of the male sources of HIV infections, especially for young women aged 15 to 24, such as the ratio between younger men (aged around 25, within about ± 5 years age difference of female recipients) and older men (aged around 35, > 5 years age difference) as infection sources. Under our framework, this can be simulated via the

BVN mixture model of the density function $f_1(\cdot)$ (the density function for MF transmissions). We consider two scenarios with respect to the age structure of male sources: (1) "**same age**" scenario, where younger male sources ($\sim$ 25 y.o.) contribute to about 60% total infections, and older male sources contribute around 30% (with the other 10% attributed to other age groups); (2) "**discordant age**" scenario with the ratio reversed, i.e., older male contribute about 60% infections while younger male only contribute 30%. The relevant model parameters for each scenario are detailed in the Online Supporting Information.

We explore 5 different sample sizes (i.e., numbers of likely transmission pairs) with $N = 100, 200, 400, 600$ and 800. 100 independent simulations are run for each scenario and each sample size N. For brevity, we include all the parameter and prior choices in the simulation study in the Online Supporting Information.

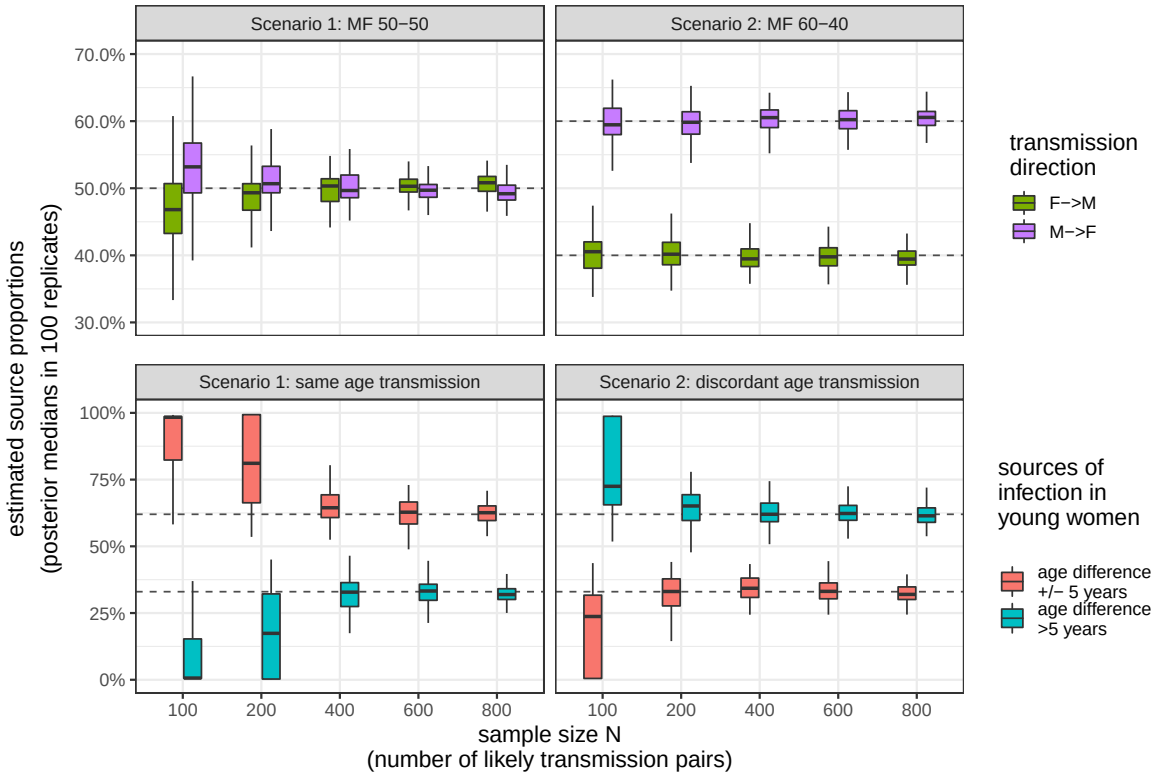

Figure 2: Model performance in latent transmission pattern identification with various event sizes. Top: posterior mean proportions of infections from younger men v.s. older men in Scenarios 1 and 2. Bottom: posterior mean proportions of MF and FM transmissions in Scenarios 1 and 2. The dashed lines mark the true proportion values, and the short solid line in each box marks the median of posterior means. Results are plotted using 100 simulation runs for each scenario and event size.

The simulation experiment results are summarized in Figure 2. Each box in the boxplots profiles the distribution of posterior median proportions (of a transmission direction or a male source age group) across 100 simulations with a specific event size N and a certain scenario.

In the top row, we show results for the proportions of male-to-female (MF) and female-to-male (FM) transmission events. The proportions are expected to be equal in Scenario "**MF 50-50**" (left) and MF events proportion is higher in Scenario "**MF 60-40**" (right), with true values marked with horizontal dashed lines. As sample size N increases, the posterior means concentrate more tightly around the true values. Even with moderate small sample size (like $N = 200$), we have relatively accurate inference results.

In the bottom row, we illustrate inferences of proportions of younger (red) versus older (blue) male sources of infections, with scenario "**same age**" on the left and scenario "**discordant age**" on the right. Similarly, inference accuracy gets improved with a larger sample sample size. Even with a small sample (like $N = 100$), the relative relationship between the two proportions is inferred correctly, although the difference between them seems to be over-estimated. Such over-estimation is likely due to the parsimony behavior of DP priors that tend to assign data points to the biggest existing clusters when there are not enough data to admit a new cluster, and so more points would be attributed to the component with the highest weight when N is small; this effect is mitigated as N increases. With a moderate sample size of $N = 400$ (smaller than the size of the real dataset), we already have satisfactory estimation precision for these mixture weight parameters. Also, in comparison, inference for source age proportions is harder (thus less accurate) than for the direction proportions (in top role), as there are fewer proportion parameters (only 3 entries in $\mathbf{p}$) to infer for the transmission direction than for the spatial patterns (6 BVN mixture components in this simulation study).

We can also inspect the posterior credible intervals as well as inference properties such as MCMC convergence. In Web Appendix B of the supporting information, we include additional plots to show that 95% credible intervals for the quantities of interest provide satisfactory coverage and have decreasing widths as N increases. In addition, by examining traceplots (also in Web Appendix B), we can check for convergence of the sampling algorithm.

5 Case study: inferring transmission flows by age from viral sequencing data

Note by authors: In order to preserve data autonomy of African countries and protect private health data of study participants from Uganda, results of the real data analysis have to be redacted from this online version. The full analysis results are under extensive, months-long review of the data consortium due to data regulations. The full manuscript, however, can be provided and shared privately upon request.

In this section, we analyze data containing demographic information and viral sequencing data collected in the in the Rakai Community Cohort Study between August 2011 and January 2015 (Ratmann et al., 2019, 2020; Xi et al., 2022). Recall for a point $\mathbf{s}_i$ on the age-by-age surface that denotes the ages for a pair of individuals, we are uncertain about its type label c_i, which could be 0 (no-event), 1 (male-to-female transmission), or -1 (female-to-male transmission). In the subsequent analyses, we aim to address such uncertainties with our model.

Using our framework, two modes of inference will be conducted: first, we fit the full model which does *not* require any fixed thresholding or pre-classification of the data. Second, and to illustrate the effect of learning the transmission statuses, we utilize a pre-specified

point classification, and fit the remaining parameters to learn the continuous spatial process component describing age structures in transmission flow. The second analysis is more similar to that in Xi et al. (2022) while operating in a continuous framework rather than one requiring discretization. Moreover, we will compare results from our two analyses in order to highlight the new insights revealed by the more flexible joint estimation framework.

Settings and data processing. After thresholding paired samples with linkage scores smaller than 0.2 as a preprocessing step to remove highly unlikely true transmission pairs, our analysis focuses on the 526 remaining pairs. These 526 filtered pairs are potential candidates for transmission events, but our model will probabilistically learn the likelihoods of their transmission links from information provided by data. Even for a pair that is highly likely to be linked through disease transmission, we do not have direct knowledge about the transmission direction and the model will also probabilistically characterize the likelihoods of each transmission direction.

In both analyses, we adopt the same priors as described in Section 4 and run the MCMC algorithm for 3000 iterations with 1000 burn-in steps. In lieu of a detailed report of runtime, we note that the full MCMC inference algorithm on a laptop with a standard 4-core Intel CPU takes less than 10 minutes, which is a drastic improvement of efficiency compared to prior work, where 4000 iterations would take about 30 hours. Despite the complexity and flexibility of the model, we see that posterior inference is efficient, opening up the possibility of analyzing larger-scale datasets.

Full analysis: with flexible point types ("Model"). We first apply the full model to learn transmission events and their directions probabilistically. This means that we can reduce the amount of prior information and data pre-processing work needed to specify the types of pairs deterministically, and we can address the intrinsic uncertainties in the data by allowing the phylogenetic analysis outputs (linkage and direction scores) to quantitatively inform inference. More specifically, instead of assuming a hard threshold and allocating points to fixed type labels, we adopt a "soft-thresholding" method through the score distributions defined in Section 2: If the linkage score ℓ_i is not close enough to 1, then pair i may not be linked in a transmission event, and if the direction score d_i is near 0.5, then pair i might represent either a male-to-female or a female-to-male transmission event. We accommodate such uncertainty in our analysis by not fixing the point types, but instead only specifying the parameters μ_d and μ_{-d}, which are the centers of MF transmission direction scores and FM transmission direction scores, respectively. All the other parameters in the full model are assumed unknown and need to be learned. This means that we need to run all the steps in the inference algorithm, but simply with μ_d and μ_{-d} fixed. In this case study, we choose $\mu_d = 1.5$ and $\mu_d = -1.5$, which implies that the d_i's with $i = 1$ are centered around 0.817 and the d_i's with $i = -1$ are centered around 0.182. We note that results are not sensitive to changes in these chosen values within a reasonable range; additional case study details can be found in the Online Supporting Information.

Partial analysis: with fixed point types ("Fixed"). In the partial analysis, we use a rule-of-thumb criterion suggested by domain experts (similar to that used by Xi et al. (2022)) to determine the occurrence and direction of a transmission event: for a pair i, if $\ell_i > 0.6$, we believe that a real transmission event took place between i_1 and i_2, and further, if $d_i > 0.5$ the transmission was male-to-female (MF), and otherwise it was female-to-male (FM). Following this heuristic to allocate point types, we may consider all the type indicators c_i's as known and fixed, and only focus on learning the spatial patterns for each type separately. As a result,

inference reduces to inference for the DPGMMs (the spatial component) only, as all remaining parameters are either fixed or can be sampled directly in a Monte Carlo step.

In the remainder of this section, we will first present our main results and discussion of the learned age structures of HIV transmissions. We then take a closer look at our inference results and examine the inferred transmission age patterns; we will discuss similarities and differences between the full and partial analyses, and meanwhile demonstrate how our method is able to leverage more data information with more consideration of uncertainty. We shall point out that all analyses presented in this section are based on data collected between 2011 and 2015, with transmission dynamics possibly different from patterns in more recent years.

5.1 Main results: identified transmission events and learned age structures in transmissions

Full results redacted due to data regulations. Available to interested parties upon request.

5.2 Transmission age structure in more detail, with uncertainty

Full results redacted due to data regulations. Available to interested parties upon request.

6 Conclusion

In this paper, we develop a Bayesian hierarchical spatial Poisson process model to learn disease transmission structures that are not directly observed, and apply it to analyzing HIV viral deep-sequencing data to uncover the transmission flow between different age groups at the population level. Our framework is novel in that it does not require any fixed thresholds or pre-specified classification on the data points about the transmission relationships between potential pairs of sources and recipients. We can probabilistically learn such unobserved relationships — whether or not there is transmission between a pair and in which direction the transmission occurs — with a fully Bayesian inference algorithm. Moreover, our method is based on a continuous spatial process that, unlike previous work (Xi et al., 2022), does not require discretization of the feature space and thus avoids keeping track of all cells in a large transmission flow matrix in computation (our method only needs to track all the data points). This advantage by construction has made our method much more computationally efficient. Our more flexible, generic framework allows inclusion of more data in analysis while accounting for the intrinsic uncertainties associated with outcomes of deep-sequence phylogenetic analyses. In our simulations and real data case study, we demonstrate that our new method is able to effectively exploit richer information from data and bring new insights into valuable epidemiological questions. Most phylodynamic analyses into the age-specific drivers of HIV transmission are limited to analyses of relatively coarse age bands, either for technical or computational reasons (De Oliveira et al., 2017; Le Vu et al., 2019; Bbosa et al., 2020). In this context, the model proposed here further extends ongoing and epidemiologically important research to enable inferences of high-resolution transmission flows.

There are several future directions that one may take based on our proposed framework. It may be of interest to incorporate individual-level covariates into the spatial process, either as additional marks or latent effects of the Poisson process (e.g., Hu and Bradley (2018)) or as additional covariates in the signal distributions. Also, extensions to non-normal components in the spatial density function mixture model can help relax certain assumptions entailed by a normal mixture model; for example, if similarity of transmission behavior is not necessarily dependant on spherical spatial proximity (an implicit assumption of the normal model), then some other kernels (e.g., bivariate Beta, as in Kottas and Sansó (2007)) could be considered.

Acknowledgements

This work was partially supported by NSF DMS-2030355. We would like to acknowledge the Rakai Health Sciences Program, particularly the Rakai Community Cohort Study and its participants and the PANGEA HIV consortium. We also thank Mike West for helpful comments and discussion.

Supporting Information

All Supporting Information and Web Appendices referenced in the main text is available in the supplementary document available at https://fanbu1995.github.io/Documents/HIV_transmission_supp.pdf.

We have made all code and a fully anonymized dataset available on `GitHub` at https://github.com/fanbu1995/HIV-transmission-PoissonProcess.

References

Adams, R. P., I. Murray, and D. J. MacKay (2009). Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 9–16.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Bauer, C., J. Wakefield, H. Rue, S. Self, Z. Feng, and Y. Wang (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in medicine 35*(11), 1848–1865.

Bbosa, N., D. Ssemwanga, A. Ssekagiri, X. Xi, Y. Mayanja, U. Bahemuka, J. Seeley, D. Pillay, L. Abeler-Dörner, T. Golubchik, et al. (2020). Phylogenetic and demographic characterization of directed hiv-1 transmission using deep sequences from high-risk and general population cohorts/groups in uganda. *Viruses 12*(3), 331.

Berke, O. (2004). Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics 3*(1), 1–11.

Best, N., S. Richardson, and A. Thomson (2005). A comparison of bayesian spatial models for disease mapping. *Statistical methods in medical research 14*(1), 35–59.

Brix, A. and P. J. Diggle (2001). Spatiotemporal prediction for log-gaussian cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(4), 823–841.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

De Oliveira, T., A. B. Kharsany, T. Gräf, C. Cawood, D. Khanyile, A. Grobler, A. Puren, S. Madurai, C. Baxter, Q. A. Karim, et al. (2017). Transmission networks and risk of hiv infection in kwazulu-natal, south africa: a community-wide phylogenetic study. *The lancet HIV 4*(1), e41–e50.

Eisinger, R. W. and A. S. Fauci (2018). Ending the hiv/aids pandemic. *Emerging infectious diseases 24*(3), 413.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association 90*(430), 577–588.

Fauci, A. S. and H. C. Lane (2020). Four decades of hiv/aids—much accomplished, much to do. *New England Journal of Medicine 383*(1), 1–4.

Givens, G. H., D. Smith, and R. Tweedie (1997). Publication bias in meta-analysis: a bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science 12*(4), 221–250.

Gschlößl, S. and C. Czado (2008). Modelling count data with overdispersion and spatial effects. *Statistical papers 49*(3), 531–552.

Hall, M., T. Golubchik, D. Bonsall, L. Abeler-Dorner, M. Limbada, B. Kosloff, A. Schaap, M. de Cesare, G. Mackintyre-Cockett, W. Probert, et al. (2021). Demographic characteristics of sources of hiv-1 transmission in zambia. *medRxiv*.

Heikkinen, J. and E. Arjas (1998). Non-parametric bayesian estimation of a spatial poisson intensity. *Scandinavian Journal of Statistics 25*(3), 435–450.

Heuveline, P. (2004). Impact of the hiv epidemic on population and household structure: the dynamics and evidence to date. *AIDS (London, England) 18*(0 2), S45.

Hu, G. and J. Bradley (2018). A bayesian spatial–temporal model with latent multivariate log-gamma random effects with application to earthquake magnitudes. *Stat 7*(1), e179.

Huber, M. (2011). Spatial point processes. *Handbook of Markov Chain Monte Carlo*, 253–278.

Hyman, J. M., J. Li, and E. A. Stanley (1994). Threshold conditions for the spread of the hiv infection in age-structured populations of homosexual men. *Journal of theoretical biology 166*(1), 9–31.

Ishwaran, H. and L. F. James (2004). Computational methods for multiplicative intensity models using weighted gamma processes: proportional hazards, marked point processes, and panel count data. *Journal of the American Statistical Association 99*(465), 175–190.

Ji, C., D. Merl, T. B. Kepler, and M. West (2009). Spatial mixture modelling for unobserved point processes: Examples in immunofluorescence histology. *Bayesian analysis (Online) 4*(2), 297.

Johnson, O., P. Diggle, and E. Giorgi (2019). A spatially discrete approximation to log-gaussian cox processes for modelling aggregated disease count data. *Statistics in medicine 38*(24), 4871–4887.

Kim, H. and A. Kottas (2022). Erlang mixture modeling for poisson process intensities. *Statistics and Computing 32*(1), 1–15.

Kottas, A., J. A. Duan, and A. E. Gelfand (2008). Modeling disease incidence data with spatial and spatio temporal dirichlet process mixtures. *Biometrical Journal: Journal of Mathematical Methods in Biosciences 50*(1), 29–42.

Kottas, A. and B. Sansó (2007). Bayesian mixture modeling for spatial poisson process intensities, with applications to extreme value analysis. *Journal of Statistical Planning and Inference 137*(10), 3151–3163.

Le Vu, S., O. Ratmann, V. Delpech, A. E. Brown, O. N. Gill, A. Tostevin, D. Dunn, C. Fraser, E. M. Volz, and U. H. D. R. Database (2019). Hiv-1 transmission patterns in men who have sex with men: insights from genetic source attribution analysis. *AIDS research and human retroviruses 35*(9), 805–813.

Leitner, T. and E. Romero-Severson (2018). Phylogenetic patterns recover known hiv epidemiological relationships and reveal common transmission of multiple variants. *Nature microbiology 3*(9), 983–988.

Lo, A. Y. and C.-S. Weng (1989). On a class of bayesian nonparametric estimates: Ii. hazard rate estimates. *Annals of the Institute of Statistical Mathematics 41*(2), 227–245.

Mohebbi, M., M. Mahmoodi, R. Wolfe, K. Nourijelyani, K. Mohammad, H. Zeraati, and A. Fotouhi (2008). Geographical spread of gastrointestinal tract cancer incidence in the caspian sea region of iran: spatial analysis of cancer registry data. *BMC cancer 8*(1), 1–12.

Mohebbi, M., R. Wolfe, and A. Forbes (2014). Disease mapping and regression with count data in the presence of overdispersion and spatial autocorrelation: a bayesian model averaging approach. *International journal of environmental research and public health 11*(1), 883–902.

Møller, J., A. R. Syversveen, and R. P. Waagepetersen (1998). Log gaussian cox processes. *Scandinavian journal of statistics 25*(3), 451–482.

Rasmussen, C. (1999). The infinite gaussian mixture model. *Advances in neural information processing systems 12*.

Rasmussen, D. A., E. Wilkinson, A. Vandormael, F. Tanser, D. Pillay, T. Stadler, and T. de Oliveira (2018, 12). Tracking external introductions of HIV using phylodynamics reveals a major source of infections in rural KwaZulu-Natal, South Africa. *Virus Evolution 4*(2). vey037.

Ratmann, O., M. K. Grabowski, M. Hall, T. Golubchik, C. Wymant, L. Abeler-Dörner, D. Bonsall, A. Hoppe, A. L. Brown, T. de Oliveira, et al. (2019). Inferring hiv-1 transmission networks and sources of epidemic spread in africa with deep-sequence phylogenetic analysis. *Nature communications 10*(1), 1–13.

Ratmann, O., J. Kagaayi, M. Hall, T. Golubchick, G. Kigozi, X. Xi, C. Wymant, G. Nakigozi, L. Abeler-Dörner, D. Bonsall, et al. (2020). Quantifying hiv transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in rakai, uganda. *The Lancet HIV 7*(3), e173–e183.

Romero-Severson, E. O., I. Bulla, and T. Leitner (2016). Phylogenetically resolving epidemiologic linkage. *Proceedings of the National Academy of Sciences 113*(10), 2690–2695.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology) 71*(2), 319–392.

Scire, J., J. Barido-Sottani, D. Kühnert, T. G. Vaughan, and T. Stadler (2020). Improved multi-type birth-death phylodynamic inference in beast 2. *bioRxiv*.

Sharrow, D. J., S. J. Clark, and A. E. Raftery (2014). Modeling age-specific mortality for countries with generalized hiv epidemics. *PloS one 9*(5), e96447.

Skums, P., A. Zelikovsky, R. Singh, W. Gussler, Z. Dimitrova, S. Knyazev, I. Mandric, S. Ramachandran, D. Campo, D. Jha, et al. (2018). Quentin: reconstruction of disease transmissions from viral quasispecies genomic data. *Bioinformatics 34*(1), 163–170.

Taddy, M. A., A. Kottas, et al. (2012). Mixture modeling for marked poisson processes. *Bayesian Analysis 7*(2), 335–362.

van de Kassteele, J., J. van Eijkeren, and J. Wallinga (2017). Efficient estimation of age-specific social contact rates between men and women. *The Annals of Applied Statistics 11*(1), 320–339.

Vedel Jesen, E. B. and T. L. Thorarinsdottir (2007). A spatio-temporal model for functional magnetic resonance imaging data–with a view to resting state networks. *Scandinavian journal of statistics 34*(3), 587–614.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics 8*(2), 158–183.

Wolpert, R. L. and K. Ickstadt (1998). Poisson/gamma random field models for spatial statistics. *Biometrika 85*(2), 251–267.

Wymant, C., M. Hall, O. Ratmann, D. Bonsall, T. Golubchik, M. de Cesare, A. Gall, M. Cornelissen, C. Fraser, S.-H. Consortium, T. M. P. Collaboration, and T. B. Collaboration (2018). Phyloscanner: inferring transmission from within-and between-host pathogen genetic diversity. *Molecular biology and evolution 35*(3), 719–733.

Xi, X., S. E. Spencer, M. Hall, M. K. Grabowski, J. Kagaayi, and O. Ratmann (2022). Inferring the sources of hiv infection in africa from deep sequence data with semi-parametric bayesian poisson flow models. *Journal of the Royal Statistical Society Series C (Applied Statistics)*.

Zhang, Y., C. Wymant, O. Laeyendecker, M. K. Grabowski, M. Hall, S. Hudelson, E. Piwowar-Manning, M. McCauley, T. Gamble, M. C. Hosseinipour, N. Kumarasamy, J. G. Hakim, J. Kumwenda, L. A. Mills, B. R. Santos, B. Grinsztejn, J. H. Pilotto, S. Chariyalertsak, J. Makhema, Y. Q. Chen, M. S. Cohen, C. Fraser, and S. H. Eshleman (2020, 02). Evaluation of Phylogenetic Methods for Inferring the Direction of Human Immunodeficiency Virus (HIV) Transmission: HIV Prevention Trials Network (HPTN) 052. *Clinical Infectious Diseases 72*(1), 30–37.

Zhao, C. and A. Kottas (2021). Modelling for poisson process intensities over irregular spatial domains. *arXiv preprint arXiv:2106.04654*.

Zhou, Z., D. S. Matteson, D. B. Woodard, S. G. Henderson, and A. C. Micheas (2015). A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association 110*(509), 6–15.