Doubly stochastic models for spatio-temporal covariation of replicated point processes

Daniel Gervini
University of Wisconsin–Milwaukee

May 17, 2021

Abstract

This article proposes log-linear models for the latent intensity functions of replicated spatio-temporal point processes. By simultaneously fitting correlated spatial and temporal Karhunen–Loève expansions, these models produce spatial and temporal components that are usually easy to interpret and capture the main directions of spatio-temporal correlation. The asymptotic distribution of the estimators is derived, and their finite sample properties are studied by simulation. As an example of application, we analyze the spatio-temporal patterns of usage of a bike station in the Divvy bike-sharing system of the city of Chicago.

Key words: Bike-sharing system; Karhunen–Loève decomposition; latent-variable model; Poisson process.

1 Introduction

Point processes in time and space have a broad range of applications, in areas as diverse as neuroscience, ecology, finance, seismology, and others. Examples are given in classic texts like Baddeley (2007), Cox and Isham (1980), Diggle (2013), Møller and Waagepetersen (2004) and Streit (2010). Due to the prevailing types of data and applications, the point-process literature has mainly focused on single realizations of point processes. In particular, spatio-temporal processes have been widely studied (e.g. Ahn et al., 2014; Li and Guan, 2014; Shirota and Gelfand, 2017; Waagepetersen et al., 2016; Mortensen et al., 2018), but also in the context of single realizations. An exception has been spike-train data studies, where neural activity for several patients or for the same patient under different trials is observed (Brown et al., 2004); these are examples of replicated point processes, that is, processes observed for different subjects or units.

Among the few papers that have addressed replicated point processes we can cite Diggle et al. (1991), Baddeley et al. (1993), Diggle et al. (2000), Bell and Grunwald (2004), Landau et al. (2004), Wager et al. (2004), and Pawlas (2011). Because of the limited amount of data, these papers propose estimators only for summary statistics of the processes (the so-called F, G and K statistics) rather than for the more informative intensity functions. Some examples of intensity-function estimation for replicated point processes are presented in Baddeley et al. (2015, ch. 16).

The increasing availability of complex data has made replicated point processes more common in recent years. For example, bike sharing systems are becoming ubiquitous in large cities around the world (Shaheen et al., 2010). These systems provide short-term bicycle rental services at unattended stations distributed across the city. The Divvy system in the city of Chicago keeps records of every bike trip in the system and makes them publicly available at the Chicago Data Portal website (https://data.cityofchicago.org). In this article (Section 6) we analyze trips that took place between April 1 and November 30 of 2016. There were a total of 3,068,211 trips and 458 active bike stations in that period. Each bike trip can be represented by a point $(t, \mathbf{s})$ where t is the trip start time and $\mathbf{s}$ the destination. The 458 stations are so densely distributed in the city that, for practical purposes, we can regard $\mathbf{s}$ as a continuous spatial variable. Then, all bike trips on a given day constitute a single realization of a spatio-temporal process, and the 244 days between April 1 and

November 30 are the n replications of the process. We focus on trips originating at a single bike station, the one at the intersection of Wrightwood and Ashland Avenues, identified as station 166 in the system. We chose that station because it has the median annual trip count (4,304) among the 458 stations in the system. It is of interest to investigate patterns in the daily distributions of trip start times and destinations. For example, is bike demand uniformly distributed during the day, or does it peak at certain times? Is this pattern similar every day of the week or is it different on weekdays and weekends? Are trip destinations uniformly distributed in the vicinity of the bike station or do some specific locations tend to attract more trips? Are these temporal and spatial patterns related? For example, do days with a bike demand peak in the early morning also show trip destinations concentrated on a specific area, such as downtown? Answering these questions is important for an efficient administration of the system, because understanding the patterns of usage of each bike station helps correct imbalances in bike distribution that inevitably arise in these systems (Nair and Miller-Hooks, 2011).

The methods proposed in this article exploit the availability of replications, which allow 'borrowing strength' across several days. Otherwise, estimation of daily intensity functions would not be feasible for these data, since, on some days, only a dozen or so trips take place. Such low counts do not allow accurate estimation of temporal intensity functions, let alone spatial ones, if estimation is carried out separately for each day.

The idea of 'borrowing strength' across replications underlies most functional data methods (Ramsay and Silverman, 2005). However, functional data analysis has focused mostly on continuous processes; little work has been done on discrete point processes. The link between discrete and continuous time processes is provided by the underlying intensity functions, which can be modelled as realizations of latent continuous stochastic processes. This relationship has been exploited by some authors, but the literature is scant. We can mention Bouzas et al. (2006b, 2007) and Fernández-Alcalá et al. (2012), who have rather limited scopes since they only estimate the mean of temporal processes, not their variability. Wu et al. (2013) estimate the mean and the principal components of temporal processes but their kernel-based methods are not easy to extend to spatial domains. The same can be said about Bouzas et al. (2006a).

Recently proposed models for replicated temporal or spatial processes include

Gervini (2016) and Gervini and Khanal (2019), and for marked point processes with continuous marks, Gervini and Baur (2020), but joint models for spatio-temporal processes have not been proposed yet. This article proposes a log-linear model for the latent intensity functions of such processes. The model is based on the Karhunen–Loève expansion of stochastic processes. By simultaneously fitting correlated temporal and spatial Karhunen–Loève expansions, the model produces temporal and spatial components that are easy to interpret and capture the most important modes of variability and spatio-temporal correlation of the process. This method does not consist of simply fitting separate temporal and spatial models, as in Gervini (2016) and Gervini and Khanal (2019), and then computing cross-correlations. Components that are important for explaining variability in their temporal or spatial domains when taken separately, may not be optimal for explaining spatio-temporal cross-correlations. From that point of view, our method resembles multivariate canonical correlation analysis (Seber, 2004, ch. 5). Therefore, although there are similarities with Gervini and Khanal (2019), the joint spatio-temporal models proposed here are not trivial extensions of those models.

This article is organized as follows. The new model is presented in Section 2, and an estimation procedure in Section 3. Asymptotic results for statistical inference are derived in Section 4, and the finite-sample behavior of the method is studied by simulation in Section 5. As an example of application, the Divvy bike data is analyzed in Section 6.

2 Doubly stochastic spatio-temporal model

A spatio-temporal point process X is a random countable set in $\mathscr{S} = \mathbb{R} \times \mathbb{R}^2$ (Møller and Waagepetersen, 2004, ch. 2; Streit, 2010, ch. 2). A point process is locally finite if $\#(X \cap B) < \infty$ with probability one for any bounded $B \subseteq \mathscr{S}$, where $\#$ denotes the cardinality of a set. For a locally finite process we can define the count function $N(B) = \#(X \cap B)$, which characterizes the distribution of the process. A Poisson process is a locally finite process for which there exists a locally integrable function $\lambda : \mathscr{S} \to [0, \infty)$, called the intensity function, such that *(i)* $N(B)$ has a Poisson distribution with rate $\int_B \lambda(t, \mathbf{s})\, dt\, d\mathbf{s}$, and *(ii)* for disjoint sets $B_1, \ldots, B_k$ the random variables $N(B_1), \ldots, N(B_k)$ are independent. A consequence of *(i)* and *(ii)* is that the conditional distribution of the points in $X \cap B$ given $N(B) = m$ is the

distribution of m independent and identically distributed observations with density $\lambda(t, \mathbf{s}) / \int_B \lambda$, $(t, \mathbf{s}) \in B$.

It follows that, for a realization $x = \{(t_1, \mathbf{s}_1), \ldots, (t_m, \mathbf{s}_m)\}$ of a Poisson process X on a given bounded region $B = B_t \times B_s \subset \mathbb{R} \times \mathbb{R}^2$, the density function (in the sense of Proposition 3.1 of Møller and Waagepetersen, 2004) is

$$f(x) = \frac{\exp(-\int_B \lambda)}{m!} \prod_{j=1}^{m} \lambda(t_j, \mathbf{s}_j). \tag{1}$$

For replicated point processes, a single intensity function λ rarely provides an adequate fit for all replications; it is more reasonable to assume that λ itself is random. Such processes are called doubly stochastic (Møller and Waagepetersen, 2004, ch. 5; Streit, 2010, ch. 8). A doubly stochastic Poisson process is a pair (X, Λ) where $X \mid (\Lambda = \lambda)$ is a Poisson process with intensity function λ, and Λ is a stochastic process that takes values on the space $\mathscr{F}$ of non-negative locally integrable functions on $\mathscr{S}$.

We will assume $\Lambda(t, \mathbf{s})$ factorizes as

$$\Lambda(t, \mathbf{s}) = R\Lambda_t(t)\Lambda_s(\mathbf{s}), \tag{2}$$

where Λ_t is a temporal process, Λ_s a spatial process, and R a random scale factor. Identifiability constraints for this factorization are discussed below. For a common Poisson process with fixed $\lambda(t, \mathbf{s})$, the factorization in Equation (2) would imply spatio-temporal separability, that is, independence of the time points and the spatial points. But for doubly stochastic processes, this is not the case. We will assume that R, Λ_t and Λ_s are correlated and will model their interdependence, and this correlation will be reflected in a non-separable unconditional distribution of the spatio-temporal points.

The scale factor R and the latent processes Λ_t and Λ_s are non-negative. For simplicity, we assume they are strictly positive and model their logarithms,

$$\log R = \tau + Z, \tag{3}$$

where Z is a zero-mean random variable,

$$\log \Lambda_t(t) = \mu(t) + \sum_{k=1}^{p_1} U_k \phi_k(t), \tag{4}$$

and

$$\log \Lambda_s(\mathbf{s}) = \nu(\mathbf{s}) + \sum_{k=1}^{p_2} V_k \psi_k(\mathbf{s}), \tag{5}$$

where the ϕ_k's and the ψ_k's are orthonormal functions in $L^2(B_t)$ and $L^2(B_s)$, respectively, $E(U_k) = E(V_k) = 0$ for all k, and $\operatorname{cov}(U_k, U_{k'}) = \operatorname{cov}(V_k, V_{k'}) = 0$ for all $k \neq k'$. The terms in Equations (4) and (5) are arranged in decreasing order of variances $\sigma_{uk}^2 = \operatorname{var}(U_k)$ and $\sigma_{vk}^2 = \operatorname{var}(V_k)$. For any $\log \Lambda_t \in L^2(B_t)$ with $E(\| \log \Lambda_t \|^2) < \infty$ and $\log \Lambda_s \in L^2(B_s)$ with $E(\| \log \Lambda_s \|^2) < \infty$, expansions (4) and (5) always hold with possibly infinite p_1 and p_2, and are known as Karhunen–Loève expansions (Ash and Gardner, 1975, ch. 1.4). By taking finite p_1 and p_2 in (4) and (5) we do not lose much generality, in practice, since we are mostly interested in smooth processes where the first few components dominate.

Factorization (2) needs some identifiability constraints. It would seem natural to require that Λ_t and Λ_s integrate to one, in which case the overall rate of the process would be R, and Λ_t and Λ_s would be probability density functions. Unfortunately those constraints are not well adapted to the log-linear models (4) and (5). For computational simplicity, we will ask instead that $\log \Lambda_t$ and $\log \Lambda_s$ integrate to zero, for which it is sufficient to ask that μ, the ϕ_ks, ν and the ψ_ks integrate to zero. These constraints are computationally easier to handle. Under these conditions, we have

$$\log R = \frac{1}{|B|} \iint_B \log \Lambda(t, \mathbf{s}) \, dt \, d\mathbf{s},$$

$$\log \Lambda_t(t) = \frac{1}{|B_s|} \int_{B_s} \log \Lambda(t, \mathbf{s}) \, d\mathbf{s} - \frac{1}{|B|} \iint_B \log \Lambda(t, \mathbf{s}) \, dt \, d\mathbf{s},$$

and

$$\log \Lambda_s(\mathbf{s}) = \frac{1}{|B_t|} \int_{B_t} \log \Lambda(t, \mathbf{s}) \, dt - \frac{1}{|B|} \iint_B \log \Lambda(t, \mathbf{s}) \, dt \, d\mathbf{s},$$

where $|\cdot|$ denotes Lebesgue measure.

From Equations (4) and (5) it follows that the dependence structure between R, Λ_t, and Λ_s is determined by the dependence structure between Z, $\mathbf{U} = (U_1, \ldots, U_{p_1})^\top$,

and $\mathbf{V} = (V_1, \ldots, V_{p_2})^\top$. To model this dependence we collect these random effects into a single vector $\mathbf{W} = (Z, \mathbf{U}^\top, \mathbf{V}^\top)^\top$, and assume $\mathbf{W}$ follows a multivariate normal distribution with mean zero and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_z^2 & \boldsymbol{\sigma}_{zu}^\top & \boldsymbol{\sigma}_{zv}^\top \\ \boldsymbol{\sigma}_{zu} & \mathrm{diag}(\boldsymbol{\sigma}_u^2) & \boldsymbol{\Sigma}_{uv} \\ \boldsymbol{\sigma}_{zv} & \boldsymbol{\Sigma}_{uv}^\top & \mathrm{diag}(\boldsymbol{\sigma}_v^2) \end{pmatrix},$$

where $\boldsymbol{\sigma}_u^2 = (\sigma_{u1}^2, \ldots, \sigma_{up_1}^2)$, $\boldsymbol{\sigma}_v^2 = (\sigma_{v1}^2, \ldots, \sigma_{vp_2}^2)$, $\boldsymbol{\sigma}_{zu} = \mathrm{cov}(Z, \mathbf{U})$, $\boldsymbol{\sigma}_{zv} = \mathrm{cov}(Z, \mathbf{V})$, and $\boldsymbol{\Sigma}_{uv} = \mathrm{cov}(\mathbf{U}, \mathbf{V})$. The parameters of interest are the cross-covariances $\boldsymbol{\sigma}_{zu}$, $\boldsymbol{\sigma}_{zv}$ and $\boldsymbol{\Sigma}_{uv}$, since they determine the dependence or independence of the random effects Z, U_k's, and V_k's, and thus of the temporal and spatial points.

To facilitate estimation of the functional parameters μ, ν, ϕ_k's, and ψ_k's, we use semiparametric basis-function expansions. As basis functions for the temporal elements we use B-splines, and for the spatial elements we use normalized Gaussian radial kernels. But other families could be used, like simplicial bases for irregular spatial domains —our derivations in this article are not tied down to any particular bases. We call these families $\mathscr{B}_t$ and $\mathscr{B}_s$, respectively. Let $\boldsymbol{\beta}_t(t)$ be the vector of q_1 basis functions of $\mathscr{B}_t$ and $\boldsymbol{\beta}_s(\mathbf{s})$ the vector of q_2 basis functions of $\mathscr{B}_s$. Then we assume $\mu(t) = \mathbf{c}_0^\top \boldsymbol{\beta}_t(t)$, $\phi_k(t) = \mathbf{c}_k^\top \boldsymbol{\beta}_t(t)$, $\nu(\mathbf{s}) = \mathbf{d}_0^\top \boldsymbol{\beta}_s(\mathbf{s})$, and $\psi_k(\mathbf{s}) = \mathbf{d}_k^\top \boldsymbol{\beta}_s(\mathbf{s})$. The orthonormality constraints on the ϕ_k's can be expressed as $\mathbf{c}_k^\top \mathbf{J}_t \mathbf{c}_{k'} = \delta_{kk'}$, where $\delta_{kk'}$ is Kronecker's delta and $\mathbf{J}_t = \int_{B_t} \boldsymbol{\beta}_t(t) \boldsymbol{\beta}_t(t)^\top dt$, and similarly for the ψ_k's. The zero-integral constraints for μ and the ϕ_k's can be expressed as $\mathbf{a}_{t0}^\top \mathbf{c}_k = 0$ for $k = 0, \ldots, p_1$, where $\mathbf{a}_{t0} = \int_{B_t} \boldsymbol{\beta}_t(t) dt$, and similarly for ν and the ψ_k's. For some applications, such as the bike data mentioned in the Introduction, it is also natural to require that the temporal intensity functions and their derivatives match at the endpoints of B_t. So, if $B_t = [t_l, t_u]$, we would also have the constraints $\mu(t_l) = \mu(t_u)$, $\mu'(t_l) = \mu'(t_u)$, $\phi_k(t_l) = \phi_k(t_u)$ and $\phi_k'(t_l) = \phi_k'(t_u)$ for all k, which can be expressed as $\mathbf{A}_P \mathbf{c}_k = \mathbf{0}$ for $k = 0, \ldots, p_1$, with $\mathbf{A}_P = [\boldsymbol{\beta}_t(t_u) - \boldsymbol{\beta}_t(t_l), \boldsymbol{\beta}_t'(t_u) - \boldsymbol{\beta}_t'(t_l)]^\top$.

3 Parameter estimation

3.1 Penalized maximum likelihood estimation

For simplicity of notation we collect all parameters into a single vector

$$\boldsymbol{\theta} = (\boldsymbol{\sigma}_{zu}, \boldsymbol{\sigma}_{zv}, \operatorname{vec} \boldsymbol{\Sigma}_{uv}, \tau, \sigma_z^2, \mathbf{c}_0, \operatorname{vec} \mathbf{C}, \boldsymbol{\sigma}_u^2, \mathbf{d}_0, \operatorname{vec} \mathbf{D}, \boldsymbol{\sigma}_v^2), \tag{6}$$

where $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{p_1}]$ and $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_{p_2}]$. From the distributional assumptions in Section 2, the joint density of $(x, \mathbf{w})$ can be factorized as

$$f_{\boldsymbol{\theta}}(x, \mathbf{w}) = f_{\boldsymbol{\theta}}(x \mid \mathbf{w}) f_{\boldsymbol{\theta}}(\mathbf{w})$$

with $f_{\boldsymbol{\theta}}(x \mid \mathbf{w})$ as in Equation (1) and $f_{\boldsymbol{\theta}}(\mathbf{w})$ the multivariate normal density. Explicitly,

$$f_{\boldsymbol{\theta}}(x \mid \mathbf{w}) = \frac{\exp\{-r I_t(\mathbf{u}) I_s(\mathbf{v})\}}{m!}\ r^m \prod_{j=1}^{m} \lambda_t(t_j; \mathbf{u}) \prod_{j=1}^{m} \lambda_s(\mathbf{s}_j; \mathbf{v}),$$

where $r = \exp(\tau + z)$, $\lambda_t(t; \mathbf{u}) = \exp\{\mu(t) + \mathbf{u}^\top \boldsymbol{\phi}(t)\}$, $\lambda_s(\mathbf{s}; \mathbf{v}) = \exp\{\nu(\mathbf{s}) + \mathbf{v}^\top \boldsymbol{\psi}(\mathbf{s})\}$, $I_t(\mathbf{u}) = \int_{B_t} \lambda_t(t; \mathbf{u}) dt$, and $I_s(\mathbf{v}) = \int_{B_s} \lambda_s(\mathbf{s}; \mathbf{v}) d\mathbf{s}$. The marginal density for the observable datum x is

$$f_{\boldsymbol{\theta}}(x) = \int f_{\boldsymbol{\theta}}(x, \mathbf{w}) d\mathbf{w},$$

which has no closed form. We use Laplace's approximation for its evaluation, as explained in the Supplementary Material.

Given n independent realizations $x_1, \ldots, x_n$ of the process X, the maximum likelihood estimator of $\boldsymbol{\theta}$ would be the maximizer of $\sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(x_i)$. However, if basis families $\mathscr{B}_t$ and $\mathscr{B}_s$ of large dimensions are used, it is advisable to regularize the estimators by adding roughness penalties to the objective function. We then define the penalized log-likelihood function

$$\ell_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(x_i) - \xi_1 P_1(\mu) - \xi_2 \sum_{k=1}^{p_1} P_2(\phi_k) - \xi_3 P_3(\nu) - \xi_4 \sum_{k=1}^{p_2} P_4(\psi_k), \tag{7}$$

where the ξ's are nonnegative smoothing parameters and the P_i's are roughness penalty functions. For the temporal functions μ and ϕ_k's we use $P_i(f) = \int (f'')^2$, and for the spatial functions ν and ψ_k's we use $P_i(f) = \iint \{(\frac{\partial^2 f}{\partial s_1^2})^2 + 2(\frac{\partial^2 f}{\partial s_1 \partial s_2})^2 + (\frac{\partial^2 f}{\partial s_2^2})^2\}$.

The estimator of $\boldsymbol{\theta}$ is then defined as

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}),$$

where Θ is the parameter space that includes all constraints discussed in Section 2:

$$\begin{aligned}\Theta \;=\; & \{\boldsymbol{\theta} \in \mathbb{R}^r : h^C_{kl}(\boldsymbol{\theta}) = 0,\ k = 1,\dots,l,\ l = 1,\dots,p_1; \\ & h^D_{kl}(\boldsymbol{\theta}) = 0,\ k = 1,\dots,l,\ l = 1,\dots,p_2;\quad \mathbf{a}_{t0}^\top \mathbf{c}_k = 0,\ k = 0,\dots,p_1; \\ & \mathbf{a}_{s0}^\top \mathbf{d}_k = 0,\ k = 0,\dots,p_2;\quad \mathbf{A}_P \mathbf{c}_k = 0,\ k = 0,\dots,p_1;\quad \boldsymbol{\Sigma} > 0\},\end{aligned} \tag{8}$$

with r the dimension of $\boldsymbol{\theta}$, $h^C_{kl}(\boldsymbol{\theta}) = \mathbf{c}_k^\top \mathbf{J}_t \mathbf{c}_l - \delta_{kl}$, $h^D_{kl}(\boldsymbol{\theta}) = \mathbf{d}_k^\top \mathbf{J}_s \mathbf{d}_l - \delta_{kl}$, and $\boldsymbol{\Sigma} > 0$ denoting that $\boldsymbol{\Sigma}$ is symmetric positive definite. The periodicity constraints $\mathbf{A}_P \mathbf{c}_k = 0$ may not be present in every situation, but for the sake of generality we include them in all our derivations; the results below are still valid if these constraints are simply deleted when not used.

Once $\hat{\boldsymbol{\theta}}$ has been obtained, individual predictors of the latent random effects $\mathbf{w}$ can be obtained as $\hat{\mathbf{w}}_i = E_{\hat{\boldsymbol{\theta}}}(\mathbf{w} \mid x_i)$. The estimating equations for $\hat{\boldsymbol{\theta}}$ and an EM algorithm (Dempster et al., 1977) for its computation are derived in the Supplementary Material. Programs implementing these algorithms are available on the author's website.

3.2 Choice of meta-parameters

The proposed model has a number of tuning parameters that have to be chosen by the user: *(i)* the number of functional components p_1 and p_2, *(ii)* the basis families $\mathscr{B}_t$ and $\mathscr{B}_s$ and in particular their dimensions q_1 and q_2, and *(iii)* the smoothing parameters ξs. Regarding *(ii)* we can say that the overall dimensions q_1 and q_2 of the basis families are more important parameters than other specifications such as the precise knot placement or the degree of the spline family. The dimensions of $\mathscr{B}_t$ and $\mathscr{B}_s$ should be chosen relatively large in order to avoid bias; the variability of the estimators will be taken care of by the smoothing parameters ξs. As noted by Ruppert (2002, sec. 3), although optimal q_1 and q_2 could be chosen by cross-validation (Hastie et al., 2009, ch. 7), there is little improvement in goodness of fit after a minimum dimension has been reached and the smoothing parameters essentially determine the fit after that point.

The choice of smoothing parameters ξ is then more important. It can be done objectively by cross-validation. Leave-one-out cross-validation finds ξs that maximize

$$\mathrm{CV}(\xi_1, \xi_2, \xi_3, \xi_4) = \sum_{i=1}^{n} \log f_{\hat{\boldsymbol{\theta}}^{[-i]}}(x_i), \tag{9}$$

where $\hat{\boldsymbol{\theta}}^{[-i]}$ denotes the estimator obtained omitting observation x_i. A faster alternative is to use k-fold cross-validation, where the data is split into k subsets that are alternatively used as test data, $k = 5$ being a common choice. But full four-dimensional optimization of Equation (9) is too time consuming even for five-fold cross-validation. A workable alternative is sequential optimization, where each ξ_j is optimized in turn on a grid while the other ξs are kept fixed at an initial value chosen by the user. A faster but subjective alternative is to choose the ξs by visual inspection. Plots of the means and the components for different ξs can be inspected to see how curve features emerge or vanish as ξ decreases or increases. Curve shapes change smoothly with the ξs, so there is usually a relatively broad range of ξs that produce comparable and reasonable results; there is no need to select the precise optimum.

The choice of the number of components p_1 and p_2 can also be done either objectively or subjectively: the former by cross-validation or testing, the latter by taking into account the relative contributions of the new components on the total variances, $\sigma^2_{up_1}/(\sigma^2_{u1} + \cdots + \sigma^2_{up_1})$ and $\sigma^2_{vp_2}/(\sigma^2_{v1} + \cdots + \sigma^2_{vp_2})$.

4 Asymptotics

The asymptotic behavior of $\hat{\boldsymbol{\theta}}$ as $n \to \infty$ can be studied via empirical-process techniques (Pollard, 1984; Van der Vaart, 2000), since ℓ_n in Equation (7) is the average of independent identically distributed functions plus non-random roughness penalties, as in e.g. Knight and Fu (2000). We derive here 'parametric' asymptotics where the dimensions q_1 and q_2 of the basis families $\mathscr{B}_t$ and $\mathscr{B}_s$ are held fixed and the true functional parameters are assumed to belong to $\mathscr{B}_t$ and $\mathscr{B}_s$. A full nonparametric asymptotic analysis, where the functional parameters are assumed to belong to general Sobolev spaces and the basis dimensions q_1 and q_2 go to infinity with n, would perhaps be theoretically more interesting and provide more guarantees on the legitimacy

of the statistical procedures derived from the results below, but it is too complicated for this model. The 'parametric' asymptotic approach looks like a reasonable compromise and it has been followed by other authors in similar semiparametric contexts (e.g. Yu and Ruppert, 2002, and Xun et al., 2013).

The first result of this section, Theorem 1, establishes consistency of $\hat{\boldsymbol{\theta}}$. The proof is given in the Supplementary Material. For uniqueness of the true parameters, the indeterminate signs of the ϕ_k's and ψ_k's require special handling; we also need to assume that the components have multiplicity one. Our modified parameter space, then, will be

$$\begin{aligned}\Theta \;=\; \{&\boldsymbol{\theta} \in \mathbb{R}^r : h^C_{kl}(\boldsymbol{\theta}) = 0,\ k = 1, \ldots, l,\ l = 1, \ldots, p_1; \\ &h^D_{kl}(\boldsymbol{\theta}) = 0,\ k = 1, \ldots, l,\ l = 1, \ldots, p_2;\quad \mathbf{a}_{t0}^\top \mathbf{c}_k = 0,\ k = 0, \ldots, p_1; \\ &\mathbf{a}_{s0}^\top \mathbf{d}_k = 0,\ k = 0, \ldots, p_2;\quad \mathbf{A}_P \mathbf{c}_k = 0,\ k = 0, \ldots, p_1; \\ &\boldsymbol{\Sigma} > 0;\quad \sigma_{u1} > \cdots > \sigma_{up_1} > 0;\quad \sigma_{v1} > \cdots > \sigma_{vp_2} > 0; \\ &c_{k1} \geq 0,\ k = 1, \ldots, p_1;\quad d_{k1} \geq 0,\ k = 1, \ldots, p_2\}.\end{aligned} \tag{10}$$

We make the following assumptions:

A1 The signs of the $\hat{\phi}_k$s and $\hat{\psi}_k$s are specified so that the first non-zero basis coefficient of each $\hat{\phi}_k$ and $\hat{\psi}_k$ is positive (then $\hat{\boldsymbol{\theta}} \in \Theta$ for Θ defined in (10).)

A2 The true functional parameters μ_0, ν_0, ϕ_{k0}s and ψ_{k0}s of models (4) and (5) belong to the functional spaces $\mathscr{B}_t$ and $\mathscr{B}_s$, and their basis coefficients $c_{k1,0}$ and $d_{k1,0}$ are not zero. The signs of ϕ_{k0} and ψ_{k0} are then chosen so that $c_{k1,0} > 0$ and $d_{k1,0} > 0$; therefore there is a unique $\boldsymbol{\theta}_0$ in Θ such that $f_{\boldsymbol{\theta}_0}(x)$ is the true density of the data.

A3 $\boldsymbol{\xi}_n \to \mathbf{0}$ as $n \to \infty$, where $\boldsymbol{\xi}_n = (\xi_{1n}, \xi_{2n}, \xi_{3n}, \xi_{4n})^\top$ is the vector of smoothing parameters in Equation (7).

The requirement in assumption A2 that the first basis coefficients $c_{k1,0}$ and $d_{k1,0}$ of each ϕ_{k0} and ψ_{k0} are non-zero is somewhat artificial: Although the ϕ_{k0}'s and ψ_{k0}'s must have at least one non-zero basis coefficient, it need not be the first one. However, a condition like this one is necessary to uniquely identify a 'true' parameter $\boldsymbol{\theta}_0$, which would otherwise be unidentifiable because of sign ambiguity, and that condition has

to be consistent with a sign-specification rule that can be used in practice for the estimators, like the one in assumption A1.

Theorem 1 *Under assumptions A1–A3,* $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$ *as* $n \to \infty$.

To establish asymptotic normality of the estimators we use the results of Geyer (1994), which make use of the tangent cone of the parameter space. The definition and properties of tangent cones can be found in Rockafellar and Wets (1998, ch. 6). From Theorem 6.31 of Rockafellar and Wets (1998), the tangent cone of Θ at $\boldsymbol{\theta}_0$ is

$$\begin{aligned}\mathscr{T}_0 \;=\; & \{\boldsymbol{\delta} \in \mathbb{R}^r : \nabla h^C_{kl}(\boldsymbol{\theta}_0)^\top \boldsymbol{\delta} = 0,\ k = 1, \ldots, l,\ l = 1, \ldots, p_1; \\ & \nabla h^D_{kl}(\boldsymbol{\theta}_0)^\top \boldsymbol{\delta} = 0,\ k = 1, \ldots, l,\ l = 1, \ldots, p_2;\quad \mathbf{a}_{t0}^\top \mathbf{K}_{\mathbf{c}_k} \boldsymbol{\delta} = 0,\ k = 0, \ldots, p_1; \\ & \mathbf{a}_{s0}^\top \mathbf{K}_{\mathbf{d}_k} \boldsymbol{\delta} = 0,\ k = 0, \ldots, p_2;\quad \mathbf{A}_P \mathbf{K}_{\mathbf{c}_k} \boldsymbol{\delta} = 0,\ k = 0, \ldots, p_1\},\end{aligned}$$

where $\mathbf{K}_{\mathbf{d}_k}$ and $\mathbf{K}_{\mathbf{c}_k}$ are the 'extraction' matrices such that $\mathbf{d}_k = \mathbf{K}_{\mathbf{d}_k}\boldsymbol{\theta}$ and $\mathbf{c}_k = \mathbf{K}_{\mathbf{c}_k}\boldsymbol{\theta}$. The explicit forms of $\nabla h^C_{kl}(\boldsymbol{\theta})$ and $\nabla h^D_{kl}(\boldsymbol{\theta})$ are derived in the Supplementary Material. Let $\mathbf{A}$ be the $r_1 \times r$ matrix with rows $\nabla h^C_{kl}(\boldsymbol{\theta}_0)^\top$, $\nabla h^D_{kl}(\boldsymbol{\theta}_0)^\top$, $\mathbf{a}_{t0}^\top \mathbf{K}_{\mathbf{c}_k}$, $\mathbf{a}_{s0}^\top \mathbf{K}_{\mathbf{d}_k}$, and $\mathbf{A}_P \mathbf{K}_{\mathbf{c}_k}$, and let $\mathbf{B}$ be an orthogonal complement of $\mathbf{A}$, that is, an $(r - r_1) \times r$ matrix such that $\mathbf{A}\mathbf{B}^\top = \mathbf{O}$.

The next theorem gives the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. In addition to $\mathbf{B}$ above, it uses Fisher's information matrix,

$$\begin{aligned}\mathbf{F}_0 \;&=\; E_{\boldsymbol{\theta}_0}\{\nabla \log f_{\boldsymbol{\theta}_0}(X) \nabla \log f_{\boldsymbol{\theta}_0}(X)^\top\} \\ &=\; -E_{\boldsymbol{\theta}_0}\{\nabla^2 \log f_{\boldsymbol{\theta}_0}(X)\},\end{aligned}$$

where ∇ and ∇^2 are taken with respect to the parameter $\boldsymbol{\theta}$, and also $\mathsf{D}\mathbf{P}(\boldsymbol{\theta})$, the Jacobian matrix of the smoothness penalty vector $\mathbf{P}(\boldsymbol{\theta}) = (P(\mu), \sum_{k=1}^{p_1} P(\phi_k), P(\nu), \sum_{k=1}^{p_2} P(\psi_k))^\top$. Explicit expressions for these derivatives are given in the Supplementary Material. We also need an additional assumption:

A4 $\sqrt{n}\boldsymbol{\xi}_n \to \boldsymbol{\kappa}$ as $n \to \infty$, for a finite $\boldsymbol{\kappa}$.

Theorem 2 *Under assumptions A1–A4,* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathrm{N}(-\mathbf{V}\mathsf{D}\mathbf{P}(\boldsymbol{\theta}_0)^\top \boldsymbol{\kappa}, \mathbf{V})$ *as* $n \to \infty$, *with* $\mathbf{V} = \mathbf{B}^\top(\mathbf{B}\mathbf{F}_0\mathbf{B}^\top)^{-1}\mathbf{B}$.

Fisher's information matrix $\mathbf{F}_0$ can be estimated by

$$\hat{\mathbf{F}}_0 = \frac{1}{n}\sum_{i=1}^{n} \nabla \log f_{\hat{\boldsymbol{\theta}}}(x_i) \nabla \log f_{\hat{\boldsymbol{\theta}}}(x_i)^\top, \tag{11}$$

and $\mathbf{V}$ in Theorem 2 by $\hat{\mathbf{V}} = \mathbf{B}^\top(\mathbf{B}\hat{\mathbf{F}}_0\mathbf{B}^\top)^{-1}\mathbf{B}$. Because of the high dimensionality of $\boldsymbol{\theta}$, $\hat{\mathbf{F}}_0$ is often singular or nearly singular for small sample sizes, leading to unstable values of $\hat{\mathbf{V}}$. A practical alternative is to treat the functional parameters μ, ν, ϕ_k's and ψ_k's as fixed and known, reducing $\boldsymbol{\theta}$ to a more manageable $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\sigma}_{zu}, \boldsymbol{\sigma}_{zv}, \operatorname{vec} \boldsymbol{\Sigma}_{uv}, \tau, \sigma_z^2, \boldsymbol{\sigma}_u^2, \boldsymbol{\sigma}_v^2)$. Fisher's information matrix for $\tilde{\boldsymbol{\theta}}$, $\tilde{\mathbf{F}}_0$, is usually low-dimensional enough that it can be accurately estimated by the corresponding version of (11), $\widehat{\tilde{\mathbf{F}}}_0$, even for relatively small sample sizes. Since $\tilde{\boldsymbol{\theta}}$ is not subject to equality constraints or smoothness penalties, the asymptotic distribution of $\sqrt{n}(\widehat{\tilde{\boldsymbol{\theta}}} - \tilde{\boldsymbol{\theta}}_0)$ is simply $\mathrm{N}(\mathbf{0}, \tilde{\mathbf{F}}_0^{-1})$, according to the standard maximum likelihood asymptotics. The functional parameters μ, ν, ϕ_k's and ψ_k's still need to be estimated, of course, and plugged into $\widehat{\tilde{\mathbf{F}}}_0$. Despite it being an approximation, this 'reduced' or 'marginal' asymptotics still gives accurate variance estimators even for small sample sizes, as we show by simulation in Section 5.

5 Simulations

To assess the finite-sample behavior of the estimators, we simulated data from model (3)–(5) with $p_1 = p_2 = 2$. We took the interval $B_t = [0, 1]$ as the temporal domain, with functional parameters $\mu(t) = \sin \pi t - c_1$, $\phi_1(t) = (\sin \pi t - c_1)/c_2$ and $\phi_2(t) = \sqrt{2} \sin 2\pi t$, where c_1 and c_2 are standardizing constants. As the spatial domain we took the square $B_s = [0, 1] \times [0, 1]$, with functional parameters $\nu(s_1, s_2) = -(s_1 - .5)^2 - (s_2 - .5)^2 - c_3$, $\psi_1(s_1, s_2) = \{\sin \pi s_1 \sin \pi s_2 - c_4\}/c_5$ and $\psi_2(s_1, s_2) = 2 \sin 2\pi s_1 \sin 2\pi s_2$, where c_3, c_4 and c_5 are standardizing constants. The ϕ_ks and the ψ_ks satisfy the orthogonality constraints. For τ we used two different values, $\tau = \log 10$ and $\tau = \log 30$; the lower τ generates sparse data where intensity functions cannot be estimated by individual smoothing.

The variances were taken as $\sigma_{u1}^2 = .3^2 \times .7$, $\sigma_{u2}^2 = .3^2 \times .3$, $\sigma_{v1}^2 = .7^2 \times .7$ and $\sigma_{v2}^2 = .7^2 \times .3$. The cross-covariance parameters were set as $\boldsymbol{\sigma}_{zu} = \mathbf{0}$, $\boldsymbol{\sigma}_{zv} = \mathbf{0}$ and $\boldsymbol{\Sigma}_{uv}$ a diagonal matrix with elements $\boldsymbol{\Sigma}_{uv,11} = .7\sigma_{u1}\sigma_{v1}$ and $\boldsymbol{\Sigma}_{uv,22} = .7\sigma_{u2}\sigma_{v2}$, so

	τ							
	$\log 10$				$\log 30$			
	n				n			
Parameter	50	100	200	400	50	100	200	400
$\sigma_{zu,1}$	.024	.019	.016	.013	.015	.013	.012	.011
$\sigma_{zu,2}$	.016	.011	.009	.007	.012	.008	.006	.004
$\sigma_{zv,1}$	.044	.043	.037	.035	.039	.034	.031	.030
$\sigma_{zv,2}$	.030	.019	.013	.008	.022	.014	.009	.007
$\Sigma_{uv,11}$	.043	.031	.017	.014	.032	.021	.015	.010
$\Sigma_{uv,21}$	.040	.031	.018	.015	.024	.017	.011	.008
$\Sigma_{uv,12}$	.031	.019	.012	.008	.014	.013	.008	.006
$\Sigma_{uv,22}$	.026	.015	.011	.007	.013	.008	.007	.004
τ	.089	.092	.101	.103	.073	.066	.065	.065
μ	.145	.107	.074	.062	.117	.084	.072	.058
ϕ_1	.551	.425	.243	.180	.365	.249	.157	.116
ϕ_2	.724	.541	.395	.349	.451	.296	.199	.156
ν	.291	.244	.203	.188	.256	.215	.202	.178
ψ_1	.397	.295	.217	.165	.274	.213	.173	.150
ψ_2	.582	.424	.315	.249	.372	.267	.204	.171
σ_z	.068	.049	.037	.027	.051	.030	.022	.018
σ_{u1}	.045	.034	.021	.018	.036	.024	.017	.012
σ_{u2}	.040	.032	.026	.022	.030	.021	.016	.011
σ_{v1}	.060	.057	.050	.040	.062	.047	.034	.031
σ_{v2}	.047	.028	.021	.016	.037	.027	.022	.017

Table 1: Simulation Results. Root mean squared errors of parameter estimators.

U_1 and U_2 were correlated with V_1 and V_2, respectively. We considered four sample sizes: $n = 50$, $n = 100$, $n = 200$ and $n = 400$. Each scenario was simulated 500 times.

For estimation we used cubic B-splines with ten equally spaced knots as $\mathscr{B}_t$ and normalized Gaussian kernels with 25 uniformly spaced knots as $\mathscr{B}_s$. This gives dimensions $q_1 = 14$ and $q_2 = 25$, respectively. As smoothing parameters we took all ξ's equal to 10^{-5}, which produced reasonably smooth results.

As a measure of estimation error we used the root mean squared error. For scalar parameters, e.g. τ, we employed the usual definition, $\{E(\hat{\tau} - \tau)^2\}^{1/2}$. For functional parameters, e.g. $\mu(t)$, we used the root mean squared error defined in terms of the L^2-

	τ							
	$\log 10$				$\log 30$			
	n				n			
Variable	50	100	200	400	50	100	200	400
Z	.244	.242	.232	.230	.179	.174	.174	.170
U_1	.200	.180	.171	.167	.158	.144	.138	.135
U_2	.169	.157	.148	.144	.132	.121	.116	.114
V_1	.298	.282	.271	.265	.239	.215	.211	.196
V_2	.274	.247	.237	.230	.189	.173	.162	.157

Table 2: Simulation Results. Root mean squared errors of random-effect estimators.

norm, $\{E(\|\hat{\mu}-\mu\|^2)\}^{1/2}$. For random-effect estimators, e.g. the $\hat{u}_{i1}$'s, we defined it as $[E\{\sum_{i=1}^{n}(\hat{u}_{i1}-u_{i1})^2/n\}]^{1/2}$. The signs of the $\hat{\phi}_k$'s and the $\hat{\psi}_k$'s, which in principle are indeterminate, were chosen as the signs of the inner products $\langle\hat{\phi}_k,\phi_k\rangle$ and $\langle\hat{\psi}_k,\psi_k\rangle$, respectively; the signs of the $\hat{u}_{ik}$'s and $\hat{v}_{ik}$'s, and of the elements of $\hat{\boldsymbol{\sigma}}_{zu}$, $\hat{\boldsymbol{\sigma}}_{zv}$ and $\hat{\boldsymbol{\Sigma}}_{uv}$, were changed accordingly.

Table 1 shows that, as expected, estimation errors decrease as n increases, and they also decrease as the baseline rate, determined by τ, increases. Even in the sparse situation $\tau = \log 10$ we can see that the functional parameters are accurately estimated, showing the advantages of 'borrowing strength' across replications. Somewhat unusual is the case of $\hat{\tau}$, whose estimation errors do not decrease as functions of n as fast as they do for the other parameters. A more in-depth analysis reveals that this behaviour is due to bias. Nevertheless, τ is not a very important parameter for inferential purposes; more important are the cross-covariance parameters and the functional components, and they are accurately estimated.

The Supplementary Material shows plots of the simulated temporal mean and component estimators $\hat{\mu}$, $\hat{\phi}_1$ and $\hat{\phi}_2$, providing more detailed information than the overall error rates in Table 1. For example, it is clear that an increased baseline rate helps reduce border effects. This is noticeable for the second component ϕ_2, where, for example, for $n = 400$ there is a reduction in estimation error from .349 for $\tau = \log 10$ to .156 for $\tau = \log 30$; the respective plots in the Supplementary Material show that this is largely due to border effects.

Table 2 shows that random-effect estimation errors also decrease as n increases, but τ, which determines the number of observations per individual, has a larger influence than n does. The reason is that random-effect estimators can only be

computed from the observations available for each individual; 'borrowing strength' across replications is not possible for random effects.

Tables 3 and 4 compare the true finite-sample standard deviations of the estimators with their average asymptotic approximations. We use the 'reduced' asymptotics mentioned at the end of Section 4. The dimension of the full $\boldsymbol{\theta}$ is 131, whereas the dimension of the reduced $\tilde{\boldsymbol{\theta}}$ is 14, so it is clear that only the 'reduced' asymptotics are practical for these sample sizes. Tables 3 and 4 show that the true standard deviations of the estimators are accurately estimated, especially for $n \geq 100$. Even for $n = 50$, where the approximation is not as good for some parameters, the asymptotic standard deviations tend to overestimate the true standard deviations, which, for inferential purposes, is better than underestimating them. For $n \geq 200$ the approximation is extremely accurate for most parameters, even under the sparse scenario $\tau = \log 10$. The accuracy of the approximation does not change much with τ.

6 Application: Chicago's Divvy bike sharing system

In this section we analyze bike trips that took place between April 1 and November 31 of 2016 in Chicago's Divvy bike-sharing system. Specifically, we analyze trips originating at station 166, located at the intersection of Wrightwood and Ashland Avenues. For each bike trip we observe the time t when the bike was checked out and the spatial destination $\mathbf{s}$, so we can see $(t, \mathbf{s})$ as an observation of a spatio-temporal process. Strictly speaking, $\mathbf{s}$ is a discrete variable that can only take values on the lattice of 458 stations, but this grid is dense enough that for practical purposes we can take $\mathbf{s}$ as a continuous variable.

For estimation of the temporal functional parameters we used cubic B-splines with ten equally spaced knots in $B_t = [0, 24]$, so the family $\mathscr{B}_t$ has dimension $q_1 = 14$. The spatial domain B_s is more irregular. Since all trips from this station have destinations within the rectangle $[-87.840, -87.530] \times [41.800, 42.030]$ in longitude-latitude coordinates, we took as B_s the sector of the city included in this rectangle, which is basically the northern half of the city. As basis family $\mathscr{B}_s$ we used normalized Gaussian kernels with 43 equally spaced centroids (we created a grid of 100 equally

	n											
	50			100			200			400		
Parameter	True	Mean	Sd	True	Mean	Sd	True	Mean	Sd	True	Mean	Sd
$\mathrm{sd}(\hat{\sigma}_{zu,1})$	.211	.355	.082	.141	.206	.038	.104	.128	.015	.072	.084	.007
$\mathrm{sd}(\hat{\sigma}_{zu,2})$	.160	.291	.061	.115	.163	.023	.088	.100	.011	.066	.065	.006
$\mathrm{sd}(\hat{\sigma}_{zv,1})$	.327	.558	.125	.259	.329	.053	.184	.209	.024	.120	.142	.011
$\mathrm{sd}(\hat{\sigma}_{zv,2})$	.295	.420	.084	.191	.245	.033	.126	.153	.020	.081	.102	.007
$\mathrm{sd}(\hat{\Sigma}_{uv,11})$	.411	.490	.125	.310	.289	.051	.161	.177	.020	.124	.121	.012
$\mathrm{sd}(\hat{\Sigma}_{uv,21})$	.398	.377	.083	.311	.202	.035	.181	.124	.011	.151	.081	.007
$\mathrm{sd}(\hat{\Sigma}_{uv,12})$	.305	.333	.061	.192	.189	.024	.117	.119	.012	.084	.079	.006
$\mathrm{sd}(\hat{\Sigma}_{uv,22})$	.254	.299	.057	.153	.166	.022	.109	.104	.011	.070	.069	.005

Table 3: Simulation Results. Comparison of true standard deviations and asymptotic standard deviation estimators of parameter estimators ($\times 10$). For asymptotic standard deviation estimators, mean and standard deviations are reported. Results for simulations with $\tau = \log 10$.

	n											
	50			100			200			400		
Parameter	True	Mean	Sd	True	Mean	Sd	True	Mean	Sd	True	Mean	Sd
$\mathrm{sd}(\hat{\sigma}_{zu,1})$	.144	.207	.045	.095	.127	.018	.077	.086	.009	.056	.058	.005
$\mathrm{sd}(\hat{\sigma}_{zu,2})$	.118	.151	.034	.081	.092	.014	.058	.061	.006	.040	.041	.003
$\mathrm{sd}(\hat{\sigma}_{zv,1})$	.332	.401	.091	.200	.252	.037	.156	.169	.019	.110	.113	.008
$\mathrm{sd}(\hat{\sigma}_{zv,2})$	.217	.269	.056	.141	.169	.022	.092	.111	.012	.071	.076	.006
$\mathrm{sd}(\hat{\Sigma}_{uv,11})$	.322	.373	.099	.209	.223	.041	.147	.151	.020	.096	.101	.010
$\mathrm{sd}(\hat{\Sigma}_{uv,21})$	.240	.200	.037	.169	.121	.018	.108	.079	.009	.078	.054	.004
$\mathrm{sd}(\hat{\Sigma}_{uv,12})$	.139	.185	.041	.125	.110	.013	.082	.073	.006	.057	.050	.004
$\mathrm{sd}(\hat{\Sigma}_{uv,22})$	.130	.162	.033	.081	.100	.015	.066	.067	.007	.041	.045	.003

Table 4: Simulation Results. Comparison of true standard deviations and asymptotic standard deviation estimators of parameter estimators ($\times 10$). For asymptotic standard deviation estimators, mean and standard deviations are reported. Results for simulations with $\tau = \log 30$.

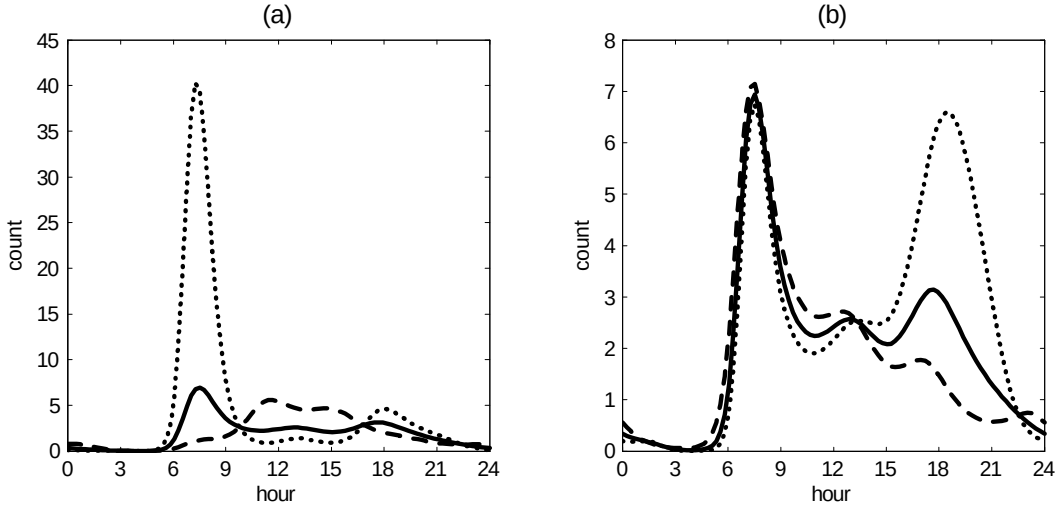

Figure 1: Divvy Data Analysis. Effect of the temporal components ϕ_k on the baseline intensity. Plot shows baseline intensities $\exp\{\hat{\mu}(t)\}$ (solid line), $\exp\{\hat{\mu}(t) - c\hat{\phi}_k(t)\}$ (dotted line) and $\exp\{\hat{\mu}(t) + c\hat{\phi}_k(t)\}$ (dashed line) for (a) first component ($k = 1$) and (b) second component ($k = 2$) .

spaced points in the rectangle $[-87.840, -87.530] \times [41.800, 42.030]$, and 43 of those ended up within the city boundaries). Then the family $\mathscr{B}_s$ has dimension $q_2 = 43$. As smoothing parameters we took all ξs equal to 10^{-5}, which provided smooth estimators while retaining a reasonable level of local detail.

We tried different combinations of numbers of components (p_1, p_2): $(1,1)$, $(2,2)$, $(3,2)$, $(3,3)$ and $(4,4)$. For each model we computed five-fold cross-validated mean log-likelihoods, obtaining 40.61, 41.23, 41.34, 41.46, and 41.50, respectively. A screeplot shows a big improvement from the $(1,1)$-model to the $(2,2)$-model, but practically no improvement from the $(3,3)$-model to the $(4,4)$-model. For the $(3,3)$-model the relative contribution of the variances of the spatial components are 82%, 16% and 2%, respectively, so the last component is rather superfluous. For this reason we opted for the $(2,2)$-model, where the relative variance proportions for the temporal components are 67% and 33%, and for the spatial components 75% and 25%, respectively.

To interpret the temporal components we plotted $\exp\{\hat{\mu}(t)\}$ versus $\exp\{\hat{\mu}(t) \pm c\hat{\phi}_k(t)\}$ for each $\hat{\phi}_k$, where c was an arbitrary constant conveniently chosen for visualization. Figure 1(a) shows that a negative score on the first component corresponds to a sharp morning peak around 7 am, while a positive score is associated with the

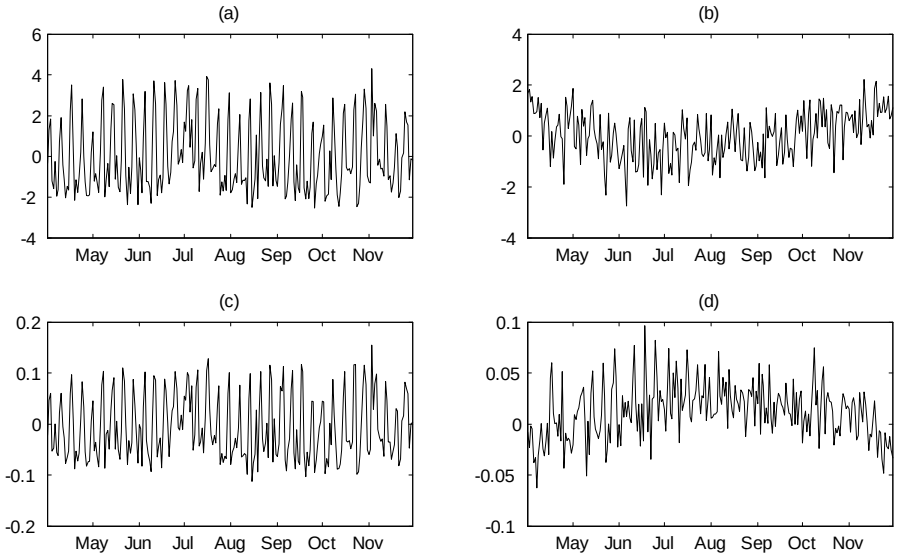

Figure 2: Divvy Data Analysis. Component scores for (a) first temporal component, (b) second temporal component, (c) first spatial component, and (d) second spatial component.

absence of a morning spike and a higher bike demand in the early afternoon. This component accounts for the difference between weekday and weekend patterns of demand. This is corroborated by a time series plot of the scores $\hat{u}_{i1}$, shown in Figure 2(a), which is strongly weekly periodic with peaks on Sundays and troughs on Thursdays and Wednesdays. Figure 1(b) shows that a negative score on the second component is associated with higher bike demand in the evening, around 6 pm, while a positive score is associated with lower demand at that time. The time series plot of the scores $\hat{u}_{i2}$ in Figure 2(b) shows a clear seasonal trend, with a minimum at the summer months. So, this component is associated with a seasonal pattern of demand.

Spatial components are harder to interpret from static plots, so we provide three-dimensional black-and-white surface plots here and colour contour plots in the Supplementary Material. Figure 3 corresponds to the first component. We see that a positive score corresponds to a sharp peak around the bike station, meaning that most trips are short and local. A negative score is associated with a higher proportion of trips to downtown. A time series plot of the $\hat{v}_{i1}$'s, shown in Figure 2(c),

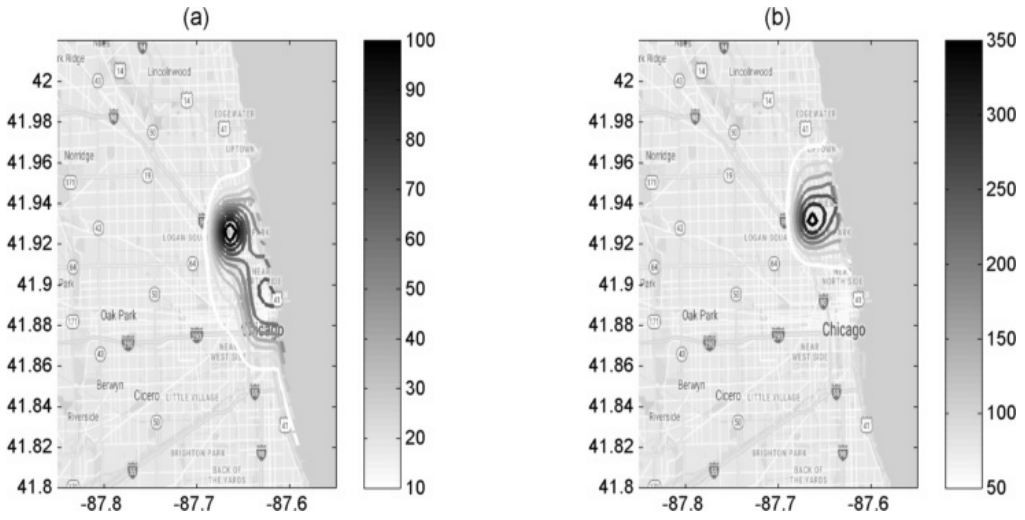

Figure 3: Divvy Data Analysis. Effect of the first spatial component ψ_1 on the baseline intensity, (a) $\exp\{\hat{\nu}(\mathbf{s}) - c\psi_1(\mathbf{s})\}$ and (b) $\exp\{\hat{\nu}(\mathbf{s}) + c\psi_1(\mathbf{s})\}$.

is strongly weekly periodical, indicating that this component is strongly associated with weekday versus weekend patterns of usage. For the second spatial component, Figure 4 shows that positive scores are associated with days when most trips stay within the neighbourhood or downtown, whereas negative scores correspond to days with a higher proportion of faraway trips. The component scores $\hat{v}_{i2}$'s, shown in Figure 2(d), show a clear seasonal trend that peaks at the summer months.

The estimated cross-correlations between temporal and spatial component scores are: $\text{corr}(U_1, V_1) = .90$, $\text{corr}(U_1, V_2) = .32$, $\text{corr}(U_2, V_1) = -.12$ and $\text{corr}(U_2, V_2) = .10$. The asymptotic standard deviations of these estimators, derived from the results in Section 4 using the Delta Method, are .10, .12, .19, and .19, respectively. Therefore only $\text{corr}(U_1, V_1)$ and $\text{corr}(U_1, V_2)$ are statistically significant. The high correlation between U_1 and V_1 is not surprising and is easy to interpret: On weekdays, there is a higher proportion of bike trips early in the morning with a downtown destination, suggesting that people use bikes for work commute; whereas on weekends, most bike trips take place in the afternoon and tend to stay within the neighbourhood.

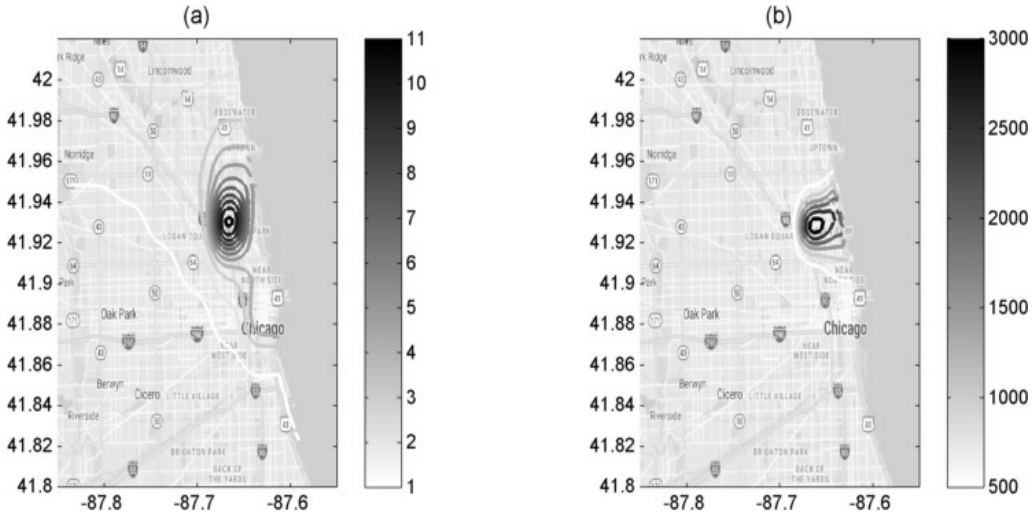

Figure 4: Divvy Data Analysis. Effect of the second spatial component ψ_2 on the baseline intensity, (a) $\exp\{\hat{\nu}(\mathbf{s}) - c\psi_2(\mathbf{s})\}$ and (b) $\exp\{\hat{\nu}(\mathbf{s}) + c\psi_2(\mathbf{s})\}$.

7 Discussion

In this article we presented a model that helps uncover spatio-temporal associations when several replications of a point process are available. In addition to the Divvy data analyzed above, other examples of data that can be modelled as replicated spatio-temporal point processes include ambulance demand (Zou et al., 2015), incidence of crime (Mohler et al., 2011; Mohler, 2014), and neuronal activity, in particular of place neurons associated with spatial memory (Kloosterman et al., 2014; Eden et al., 2018).

The proposed model is reminiscent of multivariate canonical correlation analysis, where orthogonal components of two different sets of variables are estimated and interpreted. Similarly, our model provides interpretable components for temporal and spatial variability, and estimators of their correlations. In our view, this approach is primarily an exploratory tool. Some assumptions, like independent and identically distributed replications, may not hold exactly in some cases: For example, there may be seasonal trends or mild correlations if the replications are themselves time-dependent, as in the Divvy data analyzed in Section 6.

It is possible to extend models (4) and (5) to incorporate both seasonal trends and correlations among replications. For example, by using covariates and/or autocorrelated models for the component scores U_k's and V_k's, rather than assuming

them independent and identically distributed. But this will be a matter for future research.

Acknowledgement

This research was partly supported by the US National Science Foundation, grant DMS 1505780.

References

Ahn, J., Johnson, T. D., Bhavnani, D., Eisenberg, J. N., and Mukherjee, B. (2014). A space-time point process model for analyzing and predicting case patterns of diarrheal disease in northwestern Ecuador. *Spatial and spatio-temporal epidemiology* **9** 23–35.

Ash, R.B. and Gardner, M.F. (1975). *Topics in stochastic processes*. Academic Press, New York.

Baddeley, A. (2007). Spatial point processes and their applications. In *Stochastic Geometry*, Lecture Notes in Mathematics 1892, pp. 1–75. Springer, New York.

Baddeley, A.J., Moyeed, R.A., Howard, C.V., and Boyde, A. (1993). Analysis of a three-dimensional point pattern with replication. *Applied Statistics* **42** 641–668.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC, Boca Raton, FL.

Bell, M.L., and Grunwald, G.K. (2004). Mixed models for the analysis of replicated spatial point patterns. *Biostatistics* **5** 633–648.

Bouzas, P.R., Aguilera, A., Valderrama, M., and Ruiz-Fuentes, N. (2006a). On the structure of the stochastic process of mortgages in Spain. *Computational Statistics* **21** 73–89.

Bouzas, P.R., Valderrama, M., Aguilera, A.M., and Ruiz-Fuentes, N. (2006b). Modelling the mean of a doubly stochastic Poisson process by functional data analysis. *Computational Statistics and Data Analysis* **50** 2655–2667.

Bouzas, P.R., Ruiz-Fuentes, N., and Ocaña, F.M. (2007). Functional approach to the random mean of a compound Cox process. *Computational Statistics* **22** 467–479.

Brown, E.N., Kass, R.E. and Mitra, P.P. (2004). Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience* **7** 456–461.

Cox, D.R., and Isham, V. (1980). *Point Processes.* Chapman and Hall/CRC, Boca Raton.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B* **39** 1–38.

Diggle, P.J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition.* Chapman and Hall/CRC, Boca Raton.

Diggle, P.J., Lange, N., and Beneš, F.M. (1991). Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *Journal of the American Statistical Association* **86** 618–625.

Diggle, P.J., Mateau, J., and Clough, H.E. (2000). A comparison between parametric and nonparametric approaches to the analysis of replicated spatial point patterns. *Advances in Applied Probability* **32** 331–343.

Eden, U.T., Frank, L.M., and Tao, L. (2018). Characterizing complex, multi-scale neural phenomena using state-space models. In *Dynamic Neuroscience*, pp. 29–52, Springer, NY.

Fernández-Alcalá, R.M., Navarro-Moreno, J., and Ruiz-Molina, J.C. (2012). On the estimation problem for the intensity of a DSMPP. *Methodology and Computing in Applied Probability* **14** 5–16.

Gervini, D. (2016). Independent component models for replicated point processes. *Spatial Statistics* **18** 474–488.

Gervini, D. and Khanal, M. (2019). Exploring patterns of demand in bike sharing systems via replicated point process models. *Journal of the Royal Statistical Society Series C: Applied Statistics* **68** 585–602.

Gervini, D., and Baur, T. (2020). Joint models for grid point and response processes in longitudinal and functional data. *Statistica Sinica* **30** 1905–1924.

Geyer, C.J. (1994). On the asymptotics of constrained M-estimation. *The Annals of Statistics* **22** 1993–2010.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition.* Springer, New York.

Kloosterman, F., Layton, S.P., Chen, Z., and Wilson, M.A. (2014). Bayesian decoding using unsorted spikes in the rat hippocampus. *Journal of Neurophysiology* **111** 217–227.

Knight, K., and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28** 1356–1378.

Landau, S., Rabe-Hesketh, S., and Everall, I.P. (2004). Nonparametric one-way analysis of variance of replicated bivariate spatial point patterns. *Biometrical Journal* **46** 19–34.

Li, Y., and Guan, Y. (2014). Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance. *Journal of the American Statistical Association* **109** 1205–1215.

Mohler, G.O. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting* **30** 491–497.

Mohler, G.O., Short, M.B., Brantingham, P.J., Schoenberg, F.P., and Tita, G.E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association* **106** 100–108.

Møller, J., and Waagepetersen, R.P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman and Hall/CRC, Boca Raton.

Mortensen, J.W., Heaton, M.J., and Wilhelmi, O.V. (2018). Urban heat risk mapping using multiple point patterns in Houston, Texas. *Applied Statistics* **67** 83–102.

Nair, R., and Miller-Hooks, E. (2011). Fleet management for vehicle sharing operations. *Transportation Science* **45** 524–540.

Pawlas, Z. (2011). Estimation of summary characteristics from replicated spatial point processes. *Kybernetika* **47** 880–892.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer, New York.

Ramsay, J.O., and Silverman, B.W. (2005). *Functional Data Analysis (second edition).* Springer, New York.

Rockafellar, R.T., and Wets, R.J. (1998). *Variational Analysis.* Springer, New York.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11** 735–757.

Seber, G.A.F. (2004). *Multivariate Observations.* Wiley, New York.

Shaheen, S., Guzman, S., and Zhang, H. (2010). Bike sharing in Europe, the Americas and Asia: Past, present and future. *Transportation Research Record: Journal of the Transportation Research Board* **2143** 159–167.

Shirota, S., and Gelfand, A.E. (2017). Space and circular time log Gaussian Cox processes with application to crime event data. *The Annals of Applied Statistics* **11** 481–503.

Streit, R.L. (2010). *Poisson Point Processes: Imaging, Tracking, and Sensing.* Springer, New York.

Van der Vaart, A. (2000). *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK.

Waagepetersen, R., Guan, Y., Jalilian, A., and Mateu, J. (2016). Analysis of multispecies point patterns by using multivariate log-Gaussian Cox processes. *Journal of the Royal Statistical Society Series C: Applied Statistics* **65** 77–96.

Wager, C.G., Coull, B.A., and Lange, N. (2004). Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging. *Journal of the Royal Statistical Society Series B* **66** 429–446.

Wu, S., Müller, H.-G., and Zhang, Z. (2013). Functional data analysis for point processes with rare events. *Statistica Sinica* **23** 1–23.

Xun, X., Cao, J., Mallick, B., Maity, A., and Carroll, R.J. (2013). Parameter estimation of partial differential equations. *Journal of the American Statistical Association* **108** 1009–1020.

Yu, Y., and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* **97** 1042–1054.

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