





Bayesian Nonparametric Joint Mixture Model for Clustering Spatially Correlated Time Series

Youngmin Lee & Heeyoung Kim


To cite this article: Youngmin Lee & Heeyoung Kim (2020) Bayesian Nonparametric Joint Mixture Model for Clustering Spatially Correlated Time Series, *Technometrics*, 62:3, 313-329, DOI: [10.1080/00401706.2019.1635532](https://doi.org/10.1080/00401706.2019.1635532)


To link to this article: <https://doi.org/10.1080/00401706.2019.1635532>

 View supplementary material 



 Published online: 22 Jul 2019.

 Submit your article to this journal 

 Article views: 980

 View related articles 

 View Crossmark data 

 Citing articles: 1 View citing articles 



Bayesian Nonparametric Joint Mixture Model for Clustering Spatially Correlated Time Series

Youngmin Lee and Heeyoung Kim

Department of Industrial and Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

ABSTRACT

We develop a Bayesian nonparametric joint mixture model for clustering spatially correlated time series based on both spatial and temporal similarities. In the temporal perspective, the pattern of a time series is flexibly modeled as a mixture of Gaussian processes, with a Dirichlet process (DP) prior over mixture components. In the spatial perspective, the spatial location is incorporated as a feature for clustering, like a time series being incorporated as a feature. Namely, we model the spatial distribution of each cluster as a DP Gaussian mixture density. For the proposed model, the number of clusters does not need to be specified in advance, but rather is automatically determined during the clustering procedure. Moreover, the spatial distribution of each cluster can be flexibly modeled with multiple modes, without determining the number of modes or specifying spatial neighborhood structures in advance. Variational inference is employed for the efficient posterior computation of the proposed model. We validate the proposed model using simulated and real-data examples. Supplementary materials for the article are available online.

ARTICLE HISTORY

Received June 2018

Accepted June 2019

KEYWORDS

Clustering; Dirichlet process; Gaussian process; Mixture model; Spatio-temporal data

1. Introduction

In many real-world problems, datasets are obtained as collections of time series that are spatially indexed, and these are often spatially correlated (e.g., similar annual rainfall patterns in closely located areas or similar fMRI signal patterns in neighboring voxels). The clustering of spatially correlated time series has attracted increasing interest in various areas, such as climatology, social economics, and medical science (Sahu, Gelfand, and Holland 2007; Huang, Wu, and Barry 2010; Berrocal, Gelfand, and Holland 2012; Li and Guan 2014; Zhou et al. 2015), where a major goal is to group spatially correlated time series based on both spatial and temporal similarities (Wu, Zurita-Milla, and Kraak 2015). For example, in brain imaging, regions of brain activity can be detected by clustering a collection of functional magnetic resonance imaging (fMRI) time series according to their temporal behaviors and spatial locations (Zhang et al. 2014, 2016).

Previous methods for clustering spatially correlated time series can be broadly categorized into discriminative and generative (or model-based) approaches (Banfield and Raftery 1993). Discriminative methods, which are based on certain clustering metrics, typically impose a spatial penalty on a temporal dissimilarity measure. For example, in clustering fMRI data to detect brain functional activation, Liao et al. (2008) used a hierarchical clustering method with a metric that accounts for dissimilarities in temporal patterns of fMRI time series and spatial distances between voxels. Coppi, D'Urso, and Giordani (2010) developed a fuzzy *c*-means method by adding a spatial penalty term to a

measure of dissimilarity between time series, where the spatial penalty depends on spatial proximities. Giraldo, Delicado, and Mateu (2012) performed hierarchical clustering using a measure of temporal dissimilarity that is weighted by a multivariate variogram function, calculated using the distance between two observed sites of time series. Haggarty, Miller, and Scott (2015) used a similar hierarchical clustering method, but they used the stream distance between two stations in a river network, rather than the Euclidean distance, for weighting the temporal dissimilarity measure.

Model-based methods assume that data is generated by a mixture of distributions, each of which represents a unique cluster (Banfield and Raftery 1993). In the problem of clustering spatially correlated time series, only a few model-based methods have been proposed. Blekas et al. (2007) proposed a spatially constrained polynomial regression mixture model to segment the anatomies of the heart in a cardiac perfusion magnetic resonance imaging sequence. They imposed Markov random field (MRF) priors on the mixture parameters, under the assumption that contiguous pixels most likely belong to the same class. Jiang and Serban (2012) proposed a probabilistic clustering model to classify service accessibility patterns over time for the financial services industry. They also used MRF priors to consider spatial proximities, but for the cluster assignment variables rather than the mixture parameters as in Blekas et al. (2007).

In general, model-based methods can incorporate various model structures into a probabilistic framework, allowing the obtained partition to be interpreted from a statistical point of view, and avoiding some of the arbitrariness of the heuristic

approaches of discriminative methods (Bouveyron and Brunet-Saumard 2014). However, the existing model-based methods for clustering spatially correlated time series still suffer from issues in determining the correct number of clusters (or mixture components), as well as the basis functions or the degree of polynomials in representing time series. The determination of such components is usually referred to as a model selection problem. The most prevalent approaches to this problem are to use the Akaike information criterion (AIC) (Akaike 1974) or Bayesian information criterion (BIC) (Schwarz 1978). However, these model selection approaches can lead to a high experimental cost, particularly when dealing with large datasets, and furthermore they may be inappropriate to be used when the model complexity is high.

Another problem with the model-based methods described above is that even if the spatial dependency can be adjusted using some parameters of MRF priors, the correlation structure of the MRF is almost dominated by the structure of predetermined neighborhoods. Therefore, the specification of neighborhoods should be carefully determined according to the application under study. However, it is difficult to specify appropriate neighborhoods in certain situations, for example, when a geographical space contains highly irregular regions. Some authors have pointed out the same problem in conditionally autoregressive models or Gaussian MRF models, which are widely employed in spatial statistics (Earnest et al. 2007; Assunção and Krainski 2009).

To address the above issues, we develop a new model-based method for clustering spatially correlated time series, called a Bayesian nonparametric joint mixture model. Unlike previous methods that use classical parametric models, we adopt a Bayesian nonparametric approach to the modeling of time series patterns that are spatially correlated. Specifically, we model the time series patterns as a mixture of Gaussian processes (GP), with a Dirichlet process (DP) prior over mixture components. Thus, the number of clusters does not need to be determined in advance, but is automatically determined during the clustering procedure. In this manner, by employing the GP priors over time series, we can also avoid the model selection problem of choosing the appropriate degree of polynomial models or number of knots for basis functions in modeling the time series structure.

Furthermore, unlike in previous methods that use MRF priors, we consider spatial dependencies by jointly modeling the spatial distribution of each cluster. The proposed clustering model includes spatial location information as a feature, in a similar manner to a time series being included as a feature. This model structure allows clustering to be performed by considering the similarity in terms of two features: the spatial location and time series pattern. More specifically, we model the spatial distribution of each cluster as a DP Gaussian mixture density. This approach allows the spatial distribution of each cluster to be modeled with multiple, but unknown, modes. Therefore, using the proposed method, it is possible to cluster multiple subgroups that have similar temporal patterns but are located far away from each other into a single group. In contrast, if the previous discriminative methods based on a pairwise spatial similarity metric are used, each of these subgroups would be identified as a unique cluster because of spatial dissimilarity, even though they

form a single cluster in the truth. We discuss this advantage of the proposed method in more detail using simulated examples in Section 5. Moreover, unlike previous models that use MRF priors to consider spatial dependencies, our approach does not require the specification of neighborhood structures in advance. For the posterior computation of our model, variational inference (Jordan et al. 1999) is applied, which is widely employed for learning Bayesian models, and this exhibits a fast convergence speed.

The remainder of this article is organized as follows. In Section 2, we briefly review the GP and DP mixture model. Section 3 describes the proposed model. The inference details for the proposed model are presented in Section 4. In Section 5, simulated data examples are presented. In Section 6, real examples in mobility networks and brain imaging are presented. Finally, we conclude with discussions in Section 7.

2. Background

2.1. Gaussian Process

A stochastic process f with an index $t \in \mathcal{T}$ is called a GP if for any finite subset of \mathcal{T} , $\mathbf{t} = (t_1, t_2, \dots, t_k)^\top \subset \mathcal{T}$, $f(t_1), \dots, f(t_k)$ have a multivariate Gaussian distribution (MacKay 2003). A GP is specified by a mean function $m(t)$ and a kernel function $k(t, t')$, where $t, t' \in \mathcal{T}$

$$f \sim \mathcal{GP}(m(t), k(t, t')),$$

where $m(t)$ defines the mean of $f(t)$ and is commonly assumed to be zero, and $k(t, t')$ specifies the covariance between $f(t)$ and $f(t')$. A commonly used kernel function is the squared exponential function defined as

$$k(t, t') = \lambda_1^2 \exp\left(-\frac{\|t - t'\|^2}{\lambda_2^2}\right), \quad (1)$$

where $\|\cdot\|$ denotes the Euclidean norm, λ_1^2 and λ_2^2 are scale and decay parameters, respectively. A main strength of GP is that the predictive posterior distribution at an unobserved point has a simple analytic form. The mean of predictive posterior distribution is simply a linear combination of observations. Suppose that $Y = \{y(t_i)\}_{i=1}^N$ is a set of observations, where $y(t_i) = f(t_i) + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$ is a random noise. Assuming that the observations are modeled by N -dimensional Gaussian distribution with the GP mean prior f and white noise variance σ^2 , the predicted posterior distribution $p(y(t^*)|t^*, \mathbf{t}, Y)$ at a new point t^* is given by

$$\begin{aligned} p(y(t^*)|\mathbf{t}, Y, t^*) &= \int p(y(t^*)|t^*, f)p(f|\mathbf{t}, Y)df \\ &= N(\mathbf{k}^*\mathbf{T}[K + \sigma^2 I]^{-1}Y, \sigma^2 + k(t^*, t^*) \\ &\quad - \mathbf{k}^*\mathbf{T}[K + \sigma^2 I]^{-1}\mathbf{k}^*), \end{aligned}$$

where I is an $N \times N$ identity matrix, $\mathbf{k}^* = \{k(t^*, t_i)\}_{i=1}^N$ is an N -dimensional vector, and K is an $N \times N$ matrix whose element at row i , column j is $k(t_i, t_j)$.

2.2. Dirichlet Process Gaussian Mixture Model

A DP is specified by a base measure H_0 and a concentration parameter α (Ferguson 1973). We write $G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \sim \text{DP}(\alpha H_0)$ if a distribution G is drawn from DP with the parameters α and H_0 , where δ_{θ_k} is an indicator function centered on θ_k that is drawn from H_0 , and $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{\infty}$ is an infinite sequence of mixture weights that are generated from the following stick-breaking process with $\alpha > 0$

$$\pi_k = B_k \prod_{h=1}^{k-1} (1 - B_h), \quad B_k \sim \text{Beta}(1, \alpha), \quad k = 1, \dots, \infty. \quad (2)$$

A set of mixture weights $\boldsymbol{\pi}$ generated as in Equation (2) is said to be distributed according to a Griffiths, Engen, and McCloskey (GEM) process with concentration parameter α , commonly denoted by $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ (Ewens 1990).

An important application of DP is in a Dirichlet process Gaussian mixture model (DPGMM), or also called an infinite Gaussian mixture model, which is an infinite extension of a finite Gaussian mixture model. The finite Gaussian mixture model with C components can be written as

$$p(x|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^C \pi_k \mathcal{N}(x|\theta_k),$$

where $\boldsymbol{\theta} = \{\theta_k\}_{k=1}^C = \{\mu_k, \Sigma_k^{-1}\}_{k=1}^C$, where μ_k and Σ_k^{-1} are the mean and inverse covariance matrix of the k th Gaussian distribution. The model can be extended to its Bayesian version by setting prior distributions for parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ (Görür and Rasmussen 2010). Suppose that $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^C$ is drawn from a C -dimensional symmetric Dirichlet distribution with parameter α/C and that each θ_k is given a normal-Wishart (\mathcal{NW}) distribution prior with hyperparameters $\{\mu, \zeta, \mathbf{V}, \varrho\}$, where μ is a mean parameter, ζ is a relative precision, \mathbf{V} is a positive definite scale matrix, and ϱ is the number of degrees of freedom. The DPGMM is derived by letting $C \rightarrow \infty$ and setting a DP prior for $\boldsymbol{\pi}$ instead of a Dirichlet distribution

$$p(x|\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(x|\theta_k), \quad (3)$$

$$\boldsymbol{\pi} \sim \text{GEM}(\alpha),$$

$$\theta_k \sim H_0,$$

where the base measure H_0 is equivalent to $\mathcal{NW}(\mu, \zeta, \mathbf{V}, \varrho)$. Compared to the finite Gaussian mixture model, DPGMM has a great advantage that it does not require choosing the number of mixture components in advance. This flexibility has made the DPGMM widely used for various problems such as density estimation and clustering (Kao et al. 2015).

3. Proposed Model

Let $\mathcal{D} = [Y, S] = [Y_i, S_i]_{i=1}^N$ be a collection of N observations, where $Y_i = \{y_i(t_m)\}_{m=1}^M$ denotes a time series of length M , where t_m is an observed time point, and S_i denotes the location of Y_i in a d -dimensional space. In this article, we simply take $d = 2$. The dataset \mathcal{D} is grouped into C clusters according

to their spatio-temporal similarity using the proposed Bayesian nonparametric joint mixture model. The model assumes that each spatially indexed time series is assigned to a specific group. We assume that for each group, a set of time series are generated by a GP model with spatial random effects and the corresponding locations are generated by a DPGMM. Therefore, the variables and parameters in the proposed model are specified according to a group index $c \in \{1, 2, \dots, C\}$. A brief description of the variables used in the proposed model is summarized in Table 1. Before presenting the proposed joint model, we describe the model structures for Y_i and S_i in Sections 3.1 and 3.2, respectively, assuming that $[Y_i, S_i]$ belongs to the c th group, $c \in \{1, 2, \dots, C\}$.

3.1. Time Series Modeling

We model $y_i(t_m)$, an observation at a time point t_m at a location S_i , as a realization of $f_c(t_m)$ with an additive noise

$$y_i(t_m) = f_c(t_m) + w_m(S_i) + \epsilon_{i,m}, \quad (4)$$

where $f_c(t_m)$ is the mean function of the c th group at a time point t_m , $w_m(S_i)$ is a spatial random effect at a time point t_m at a location S_i , and $\epsilon_{i,m}$ is a Gaussian distributed white noise with zero mean and variance σ_ϵ^2 . We assume that the

Table 1. Notations for variables used in this article.

Variable and parameter	Distribution	Description
$Y = \{Y_i\}_{i=1}^N$	Normal	Observed time series
σ_ϵ^2	–	Variance of white noise
$S = \{S_i\}_{i=1}^N$	DPGMM	Observed locations
$W = \{W_m\}_{m=1}^M$	Normal	Spatial random effects
Σ	–	Covariance matrix for W_m
σ_s^2, ρ_s^2	–	Variance and decay parameters for Σ
μ_m^W, Σ^W	–	Mean and covariance of variational distribution for W_m
$F = \{f_c\}_{c=1}^C$	\mathcal{GP}	Mean time series functions
κ_c	–	Hyperparameter set of kernel function for f_c
μ_c^F, Σ_c^F	–	Mean and covariance of variational distribution for f_c
$Z = \{Z_i\}_{i=1}^N$	Categorical	Cluster assignment variables for joint mixture model
ϕ_i^Z	–	Mixing weights parameter of variational distribution for Z_i
$H = \{H_{i,c}\}_{i,c=1}^{N,C}$	Categorical	Cluster assignment variables for DPGMM
$\phi_{i,c}^H$	–	Mixing weights parameter of variational distribution for $H_{i,c}$
ω	GEM	Prior for variables Z
α	–	Concentration parameter for ω
$\{b_c^\omega\}_{c=1}^C$	Beta	Stick construction variables for ω
a_c^ω, b_c^ω	–	Variational parameters for b_c^ω
$\Theta_v = \{v_c\}_{c=1}^C$	GEM	Prior sets for variables H_c
β	–	Concentration parameter for v_c
$\{b_{c,l}^v\}_{l=1}^{\infty}$	Beta	Stick construction variables for v_c
$a_{c,l}^v, b_{c,l}^v$	–	Variational parameters for $b_{c,l}^v$
$\Theta_{\mu,\Omega} = \{\mu_c, \Omega_c\}_{c=1}^C$	\mathcal{NW}	Prior sets for Gaussian components of DPGMM
$\mu_0, \tau, \Lambda, \psi$	–	Parameters for $\{\mu_c, \Omega_c\}$
$\mu_{c,l}^S, \tau_{c,l}^S, \Lambda_{c,l}^S, \psi_{c,l}^S$	–	Parameters of variational distribution for $\{\mu_c, \Omega_c\}$

spatial random effect $w_m(S_i)$ is independent across time t_m , $m = 1, \dots, M$, and the spatio-temporal covariance function is separable. We also assume that the covariance function is stationary in both space and time, that is, we assume a stationary covariance function (Finkenstadt, Held, and Isham 2006). As an alternative to parametric representation of f_c in the literature for clustering spatially correlated time series, we set a GP prior on the mean function: $f_c \sim \mathcal{GP}(0, k_c(t, t'))$. In our study, two kernel functions are considered for specifying the time series behavior: we basically use the squared exponential kernel function in Equation (1), but if the proposed model needs to capture the periodicity of time series (e.g., seasonality of annual temperature, fMRI signals for periodic stimulation), the following periodic kernel function (MacKay 1998) is added to the basic kernel function

$$\exp\left(-\frac{2 \sin^2(\pi|t - t'|/\rho)}{\tau^2}\right), \quad (5)$$

where ρ determines the length of periodicity and τ^2 controls the decay speed of correlation. The spatial random effect w_m in Equation (4) captures residual spatial association; for each time point t_m , a set of spatial random effects $W_m = \{w_m(S_i)\}_{i=1}^N$ follows an N -dimensional multivariate normal distribution with zero mean and an $N \times N$ covariance matrix Σ that consists of covariance $\text{cov}(w_m(S_i), w_m(S'_j))$. To calculate the covariance, we use the squared exponential kernel function with σ_s^2 of scale parameter and ρ_s of decay parameter. Considering all time points, we define $W = \{W_m\}_{m=1}^M$ as a variable set of spatial random effects.

Given the above description, the joint distribution and its factorized probabilities for the proposed time series model take the following form

$$p(Y_i, f_c, W, S) = p(Y_i | f_c, W, S) p(f_c) p(W | S), \quad (6)$$

where

$$\begin{aligned} p(Y_i | f_c, W, S) &= \prod_{m=1}^M \mathcal{N}(y_i(t_m) | f_c(t_m) + w_m(S_i), \sigma_\epsilon^2), \\ p(f_c) &= \mathcal{GP}(f_c | 0, k_c(t, t')), \\ p(W | S) &= \prod_{m=1}^M p(W_m | S) = \prod_{m=1}^M \mathcal{N}(W_m | 0, \Sigma). \end{aligned}$$

3.2. Spatial Distribution Modeling

Recall we assume that $[Y_i, S_i]$ belongs to the c th group, $c \in \{1, 2, \dots, C\}$. For the given c th group, we assume that a spatial coordinate S_i is generated from a bivariate Gaussian mixture model with the number of components L , mixture weights $\mathbf{v}_c = \{v_{c,l}\}_{l=1}^L$, mean parameters $\boldsymbol{\mu}_c = \{\mu_{c,l}\}_{l=1}^L$, and inverse covariance matrices $\boldsymbol{\Omega}_c = \{\Omega_{c,l}\}_{l=1}^L$. We introduce latent variables $H_c = \{H_{i,c}\}_{i=1}^N$ for indicating component assignment. Let each $H_{i,c} = \{H_{i,c,l}\}_{l=1}^L$ be the realization from an L -dimensional categorical distribution with mixture weights \mathbf{v}_c . Then, the probability mass function can be expressed as $p(H_{i,c} | \mathbf{v}_c) = \prod_{l=1}^L v_{c,l}^{H_{i,c,l}}$. If S_i is generated by the l th Gaussian density specified by parameters $\mu_{c,l}$ and $\Omega_{c,l}$, then $H_{i,c,l}$ is 1 and otherwise, $H_{i,c,l}$ is 0. We

place a normal-Wishart prior on the parameters of Gaussian components $\boldsymbol{\mu}_c$ and $\boldsymbol{\Omega}_c$. In the same way as described in Equation (3), we extend a finite mixture model to DPGMM by setting a number of mixture components L to infinity and placing the stick breaking prior with the concentration parameter β on the mixture weights \mathbf{v}_c . This allows for the model to determine the number of centroids of spatial distribution in a data-driven fashion. The joint distribution and its factorized probabilities for the proposed spatial model take the following form

$$\begin{aligned} p(S_i, H_{i,c}, \boldsymbol{\mu}_c, \boldsymbol{\Omega}_c, \mathbf{v}_c) \\ = p(S_i | H_{i,c}, \boldsymbol{\mu}_c, \boldsymbol{\Omega}_c) p(H_{i,c} | \mathbf{v}_c) p(\boldsymbol{\mu}_c, \boldsymbol{\Omega}_c) p(\mathbf{v}_c), \end{aligned} \quad (7)$$

where

$$\begin{aligned} p(S_i | H_{i,c}, \boldsymbol{\mu}_c, \boldsymbol{\Omega}_c) &= \prod_{l=1}^L \mathcal{N}(S_i | \mu_{c,l}, \Omega_{c,l}^{-1})^{H_{i,c,l}}, \\ p(H_{i,c} | \mathbf{v}_c) &= \text{categorical}(H_{i,c} | \mathbf{v}_c), \\ p(\boldsymbol{\mu}_c, \boldsymbol{\Omega}_c) &= \prod_{l=1}^L p(\mu_{c,l}, \Omega_{c,l}) \\ &= \prod_{l=1}^L \mathcal{NW}(\mu_{c,l}, \Omega_{c,l} | \mu_0, \tau, \Lambda, \psi), \\ p(\mathbf{v}_c) &= \text{GEM}(\beta), \quad v_{c,l} = B_{c,l}^\mathbf{v} \prod_{h=1}^{l-1} (1 - B_{c,h}^\mathbf{v}), \\ B_{c,l}^\mathbf{v} &\sim \text{Beta}(1, \beta), \quad l = 1, \dots, \infty, \end{aligned}$$

where μ_0 , τ , Λ , and ψ are the hyperparameters of a normal-Wishart prior.

3.3. Joint Mixture Model for Clustering Spatially Correlated Time Series

Let the i th spatial time series $\{Y_i, S_i\}$ belong to the c th group, $c \in \{1, 2, \dots, C\}$. The time series pattern and spatial distribution of $\{Y_i, S_i\}$ can be modeled as in Equations (6) and (7), respectively. We combine the two models as a joint mixture model for clustering spatially correlated time series. We introduce latent variables $Z = \{Z_i\}_{i=1}^N$, where $Z_i = \{Z_{i,c}\}_{c=1}^C$ follows C -dimensional categorical distribution parameterized by mixture weights $\boldsymbol{\omega} = \{\omega_c\}_{c=1}^C$ and $Z_{i,c} = 1$ if the i th spatial time series is assigned to the c th cluster. Similar to spatial distribution modeling in Section 3.2, C is assumed to be infinity and the prior for $\boldsymbol{\omega}$ is set to be $\text{GEM}(\alpha)$ with concentration parameter α for Bayesian nonparametric modeling.

Recall that we impose independent GP prior on each f_c with kernel function $k_c(t, t')$, because we intend to find cluster-specific characteristics of time series patterns. We denote κ_c as a set of hyperparameters for $k_c(t, t')$. For example, $\kappa_c = \{\lambda_{1,c}^2, \lambda_{2,c}^2\}$ if the squared exponential function in Equation (1) is used with $\lambda_1^2 = \lambda_{1,c}^2$ and $\lambda_2^2 = \lambda_{2,c}^2$; $\kappa_c = \{\lambda_{1,c}^2, \lambda_{2,c}^2, \xi_{1,c}^2, \xi_{2,c}^2\}$ if the periodic kernel function in Equation (5) is additionally used with $\rho = \xi_{1,c}^2$ and $\tau^2 = \xi_{2,c}^2$ for capturing periodicity of a time series pattern.

Let F, H, Θ_v , and $\Theta_{\mu, \Omega}$ denote $\{f_c\}_{c=1}^C$, $\{H_c\}_{c=1}^C$, $\{\mathbf{v}_c\}_{c=1}^C$, and $\{\boldsymbol{\mu}_c, \boldsymbol{\Omega}_c\}_{c=1}^C$, respectively. The joint distribution and its factorized probabilities of the proposed joint mixture model are given

by

$$\begin{aligned} p(Y, S, F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega}) \\ = p(Y|Z, F, W, S)p(F)p(W|S)p(S|Z, H, \Theta_{\mu, \Omega}) \quad (8) \\ \times p(H|Z, \Theta_v)p(\Theta_v)p(\Theta_{\mu, \Omega})p(Z|\boldsymbol{\omega})p(\boldsymbol{\omega}), \end{aligned}$$

where the probabilities regarding the time series model can be further factorized as

$$\begin{aligned} p(Y|Z, F, W, S) &= \prod_{i=1, c=1}^{N, C} p(Y_i|f_c, W, S)^{Z_{i,c}}, \\ p(F) &= \prod_{c=1}^C p(f_c), \\ p(W|S) &= \prod_{m=1}^M p(W_m|S), \end{aligned}$$

where $p(Y_i|f_c, W, S)$, $p(f_c)$, and $p(W_m|S)$ are defined in Equation (6), and the probabilities regarding the spatial distribution model can be further factorized as

$$\begin{aligned} p(S|Z, H, \Theta_{\mu, \Omega}) &= \prod_{i=1, c=1}^{N, C} p(S_i|H_{i,c}, \boldsymbol{\mu}_c, \boldsymbol{\Omega}_c)^{Z_{i,c}}, \\ p(H|Z, \Theta_v) &= \prod_{i=1, c=1}^{N, C} p(H_{i,c}|\mathbf{v}_c)^{Z_{i,c}}, \\ p(\Theta_v) &= \prod_{c=1}^C p(\mathbf{v}_c), \\ p(\Theta_{\mu, \Omega}) &= \prod_{c=1}^C p(\boldsymbol{\mu}_c, \boldsymbol{\Omega}_c), \end{aligned}$$

where $p(S_i|H_{i,c}, \boldsymbol{\mu}_c, \boldsymbol{\Omega}_c)$, $p(H_{i,c}|\mathbf{v}_c)$, $p(\mathbf{v}_c)$, and $p(\boldsymbol{\mu}_c, \boldsymbol{\Omega}_c)$ are defined in Equation (7), and the remaining components of the proposed model can be expressed as

$$\begin{aligned} p(Z|\boldsymbol{\omega}) &= \prod_{i=1}^N p(Z_i|\boldsymbol{\omega}) = \prod_{i=1}^N \text{categorical}(Z_i|\boldsymbol{\omega}), \\ p(\boldsymbol{\omega}) &= \text{GEM}(\alpha), \quad \omega_c = B_c^\omega \prod_{h=1}^{c-1} (1 - B_h^\omega), \quad (9) \\ B_c^\omega &\sim \text{Beta}(1, \alpha), \quad c = 1, \dots, \infty. \end{aligned}$$

In summary, the generative process for the proposed model is as follows

- For each data point, $i = 1, \dots, N$, choose a cluster index $c \in \{1, \dots, C\}$ according to $\text{categorical}(Z_i|\boldsymbol{\omega})$.
 1. Choose a spatial mixture component l according to $\text{categorical}(H_{i,c}|\mathbf{v}_c)$.
 - Sample location S_i from $\mathcal{N}(S_i|\mu_{c,l}, \Omega_{c,l}^{-1})$.
 2. Sample mean function f_c from $\mathcal{GP}(f_c|0, k_c(t, t'))$.
 3. For each time point, $m = 1, \dots, M$, sample spatial random effects W_m from $\mathcal{N}(W_m|0, \Sigma)$.
 4. For each time point, $m = 1, \dots, M$, sample temporal observation $y_i(t_m)$ from $\mathcal{N}(y_i(t_m)|f_c(t_m) + w_m(S_i), \sigma_\epsilon^2)$.

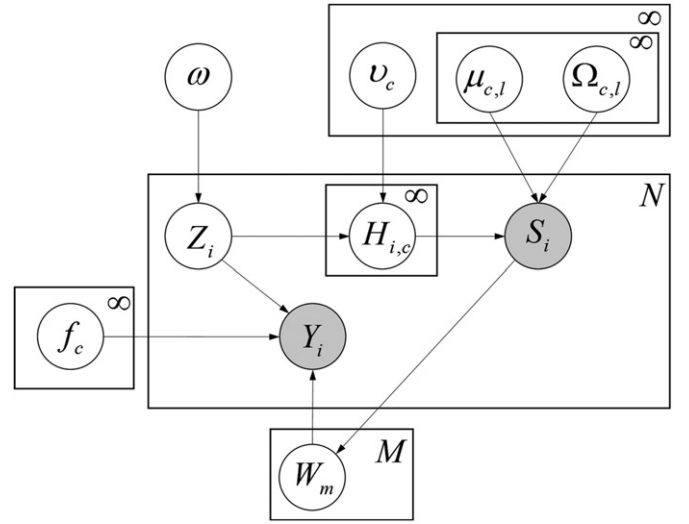


Figure 1. Graphical model representation of the proposed joint mixture model.

The graphical model representation of the proposed model is given in Figure 1. This representation shows the dependence structure of factorized probabilities in Equation (8) more clearly. Gray nodes Y_i and S_i indicate observed variables and the other transparent nodes indicate latent variables or priors. The dependence between the variables is expressed through directed arcs. Note that the probability being assigned to the same cluster for two samples is influenced by similarity in two aspects: time series pattern and spatial location as discussed in Section 1.

4. Model Estimation and Prediction

We address the variational inference for computing posterior distribution of the proposed model and derive the predictive posterior distribution of cluster-specific time series.

4.1. Posterior Computation

The posterior distribution of the latent variables conditional on data can be derived using the Bayes' theorem

$$\begin{aligned} p(F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega} | Y, S) \\ = \frac{p(Y, S, F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega})}{\int p(Y, S, F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega}) dF dW dZ dH d\boldsymbol{\omega} d\Theta_v d\Theta_{\mu, \Omega}}. \quad (10) \end{aligned}$$

We apply an approximate inference method because the analytical computation of the integral term in the denominator of Equation (10) is intractable. In this article, we use the variational inference (Jordan et al. 1999) as a deterministic alternative to Markov chain Monte Carlo (MCMC) that is widely used for approximating posterior distributions in Bayesian statistics (Andrieu et al. 2003). Variational inference has been used extensively in Bayesian machine learning. Although its theoretical properties have not been yet well studied compared with MCMC methods and it is known to underestimate posterior variance, variational inference tended to be faster than sampling based approaches in many applications (Paisley, Wang, and Blei 2011; Park, Kim, and Choi 2013; Foti et al. 2014).

Variational inference turns the inference problem for approximating the exact posterior into an optimization problem for minimizing the Kullback–Leibler (KL) divergence between a so-called variational distribution q and the true posterior p such that

$$q^* = \arg \min_{q \in \mathcal{Q}} \text{KL}(q||p),$$

where \mathcal{Q} is a parametrized family of variational distribution q that is indexed by a set of variational parameters. This optimization problem is not directly computable because the normalizing constant of the true posterior p is required, but we can use the fact that minimizing the KL divergence is equivalent to maximizing the lower bound, denoted by \mathcal{L}_{VB} , on the log marginal likelihood (Blei, Kucukelbir, and McAuliffe 2017)

$$\begin{aligned} \log p(Y, S) &= \log \int p(Y, S, F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega}) \\ &\quad \times dF dW dZ dH d\boldsymbol{\omega} d\Theta_v d\Theta_{\mu, \Omega} \\ &= \log \int \frac{p(Y, S, F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega})}{q(F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega})} \\ &\quad \times q(F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega}) \\ &\quad \times dF dW dZ dH d\boldsymbol{\omega} d\Theta_v d\Theta_{\mu, \Omega} \\ &= \log E_q \left[\frac{p(Y, S, F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega})}{q(F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega})} \right] \\ &\geq E_q[\log p(Y|Z, F, W, S)] + E_q[\log p(F)] + E_q[\log p(W|S)] \\ &\quad + E_q[\log p(S|Z, H, \Theta_{\mu, \Omega})] \\ &\quad + E_q[\log p(Z|\boldsymbol{\omega})] + E_q[\log p(H|Z, \Theta_v)] + E_q[\log p(\boldsymbol{\omega})] \\ &\quad + E_q[\log p(\Theta_v)] + E_q[\log p(\Theta_{\mu, \Omega})] \\ &\quad - E_q[\log q(F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega})] \\ &= \mathcal{L}_{VB}, \end{aligned} \quad (11)$$

where the inequality is derived using the Jensen's inequality and Equation (8). E_q means the expectation with respect to variational distribution q . We set the variational distribution to be in a parametrized distribution family that is tractable. A common choice is to set the variational distribution to be fully factorized over the latent variables, according to the mean field variational inference (Jordan et al. 1999), so that each expectation term in \mathcal{L}_{VB} can be easily calculated

$$\begin{aligned} q(F, W, Z, H, \boldsymbol{\omega}, \Theta_v, \Theta_{\mu, \Omega}) &= q(F)q(W)q(Z)q(H)q(\boldsymbol{\omega})q(\Theta_v)q(\Theta_{\mu, \Omega}) \\ &= \prod_{c=1}^{T_1} q(f_c) \prod_{m=1}^M q(W_m) \prod_{i=1}^N q(Z_i) \prod_{i=1, c=1}^{N, T_1} q(H_{i,c})q(\boldsymbol{\omega}) \prod_{c=1}^{T_1} q(\mathbf{v}_c) \\ &\quad \times \prod_{c=1, l=1}^{T_1, T_2} q(\mu_{c,l}, \Omega_{c,l}), \end{aligned} \quad (12)$$

where $q(\boldsymbol{\omega})$ and $q(\mathbf{v}_c)$ are constructed by a stick-breaking representation using $\prod_{c=1}^{T_1-1} q(B_c^\omega)$ and $\prod_{l=1}^{T_2-1} q(B_{c,l}^\nu)$, respectively, as described in Equations (9) and (7). Recall we assume that spatial random effects are independent and identically distributed

across all time points; this assumption makes it possible to construct a multivariate normal with an $N \times N$ covariance matrix for the variational distribution of spatial random effects, where N is the number of spatial locations. In Equation (12), the numbers of sticks (i.e., the number of mixture components) are truncated to be T_1 and T_2 , respectively. This truncation allows for a finite-dimensional variational distribution to approximate true posterior p that is an infinite-dimensional (Blei and Jordan 2006). To make sure that the mixture weights are equal to zero for clusters whose indices are greater than the truncation level, we impose $q(B_{T_1}^\omega = 1) = 1$ and $q(B_{c,T_2}^\nu = 1) = 1$ for all c . Each variational distribution of the latent variables in Equation (12) is set as

$$\begin{aligned} f_c &\sim \text{Normal}(\mu_c^F, \Sigma_c^F), \\ W_m &\sim \text{Normal}(\mu_m^W, \Sigma^W), \\ Z_i &\sim \text{Categorical}(\phi_i^Z), \\ H_{i,c} &\sim \text{Categorical}(\phi_{i,c}^H), \\ B_c^\omega &\sim \text{Beta}(a_c^\omega, b_c^\omega), \\ B_{c,l}^\nu &\sim \text{Beta}(a_{c,l}^\nu, b_{c,l}^\nu), \text{ and} \\ (\mu_{c,l}, \Omega_{c,l}) &\sim \mathcal{NW}(\mu_{c,l}^S, \tau_{c,l}^S, \Lambda_{c,l}^S, \psi_{c,l}^S), \end{aligned}$$

where $\mu_c^F = \{\mu_c^F(t_m)\}_{m=1}^M$, $\mu_m^W = \{\mu_m^W(S_i)\}_{i=1}^N$, $\phi_i^Z = \{\phi_{i,c}^Z\}_{c=1}^{T_1}$, and $\phi_{i,c}^H = \{\phi_{i,c,l}^H\}_{l=1}^{T_2}$. To avoid confusion with existing parameters, all variational parameters are superscripted.

Finding the variational distribution that maximizes \mathcal{L}_{VB} is implemented by updating variational parameters using a coordinate ascent algorithm. Because all priors in the proposed model take the conjugate setting, all update equations for variational parameters can be obtained in closed form. For example, the update equation for μ_m^W is obtained by taking a derivative of $\mathcal{L}_{VB}(\mu_m^W)$, the terms containing μ_m^W in \mathcal{L}_{VB} , with respect to μ_m^W and setting the derivative to zero. The update equations for all variational parameters are derived in Appendix A.

Algorithm 1 summarizes the update procedure of variational inference for the proposed model. The hyperparameters for GEM priors (i.e., α and β) and for normal-Wishart prior (i.e., μ_0 , τ , Λ , and ψ) are not estimated, but fixed. For our numerical studies in Sections 5 and 6, we conducted a sensitivity analysis

Algorithm 1 Variational inference for the proposed model

Initialize hyperparameters and variational parameters.

Specify the maximum number of iterations $MaxIter$ and the truncation levels T_1 and T_2 .

for 1: $MaxIter$ **do**

Step 1. Update the variational parameters for W :

$$\mu_m^W \text{ and } \Sigma^W$$

Step 2. Update the variational parameters for F, Z, B^ω :

$$\mu_c^F, \Sigma_c^F, \phi_i^Z, a_c^\omega, \text{ and } b_c^\omega.$$

Step 3. Update the variational parameters for $H, B^\nu, \Theta_{\mu, \Omega}$:

$$\phi_{i,c,l}^H, a_{c,l}^\nu, b_{c,l}^\nu, \mu_{c,l}^S, \tau_{c,l}^S, \Lambda_{c,l}^S, \text{ and } \psi_{c,l}^S.$$

Step 4. Update the hyperparameters using a gradient ascent method:

$$\sigma_\epsilon^2, \sigma_s^2, \rho_s^2, \text{ and } \kappa_c.$$

end for

to investigate the effects of these hyperparameters by varying their values, but they showed little or no effects on the results of clustering. Other hyperparameters σ_ϵ^2 , σ_s^2 , ρ_s^2 , κ_c are updated using a gradient ascent algorithm at the end of each iteration. Because μ_c^F and μ_m^W are closely related each other in our inference in the sense that if one has a similar value to the observation Y at the beginning of the algorithm, the other is predicted to be close to zero, we should set their initial values carefully to prevent the parameters from converging to bad local optimum. In our numerical studies in Section 5 and Section 6, we set μ_m^W to zero and applied only Step 2 during the first 3–5 iterations of Algorithm 1.

4.2. Posterior Predictive Distribution

Similar to the standard GP regression prediction presented in Section 2.1, the predictive distribution for $f_c(t^*)$ at an unobserved time point t^* can be approximated using the variational distribution $q(f_c)$

$$\begin{aligned} p(f_c(t^*)|\mathbf{t}, Y, t^*) &= \int p(f_c(t^*)|t^*, f_c) p(f_c|\mathbf{t}, Y) df_c \\ &\simeq \int p(f_c(t^*)|f_c) q(f_c) df_c \\ &\simeq N(\mathbf{k}_c^{*\top} K_c^{-1} \mu_c^F, k_c(t^*, t^*) \\ &\quad - \mathbf{k}_c^{*\top} K_c^{-1} \mathbf{k}_c^* + \mathbf{k}_c^{*\top} K_c^{-1} \Sigma_c^F K_c^{-1} \mathbf{k}_c^*), \end{aligned}$$

where $\mathbf{k}_c^* = \{k_c(t^*, t_i)\}_{i=1}^N$ is an N -dimensional vector and K_c is an $N \times N$ matrix whose element at row i , column j is $k_c(t_i, t_j)$.

5. Simulated Examples

To evaluate the performance of the proposed model, simulated examples were considered. A dataset of spatially correlated time series with three clusters was generated according to

$$y_{ij}(t_m) = f_j(t_m) + w_m(S_{ij}) + \epsilon_{i,m}, \quad i = 1, \dots, 60, \quad j = 1, 2, 3, \quad m = 1, \dots, 10, \quad (13)$$

where i is the time series index, j is the cluster membership index, m is the time point index, and $\epsilon_{i,m} \sim N(0, 0.4)$ is the

random error term. The mean functions for the three clusters were assumed as

$$\begin{aligned} f_1(t) &= \cos(10t)/\exp(t) + 1, \\ f_2(t) &= 2\cos(2t) + t^3, \text{ and} \\ f_3(t) &= 2\sin(\pi t), \end{aligned}$$

and they were depicted in Figure 2(a): f_1 , f_2 , and f_3 are indicated in colors of blue, red, and yellow, respectively. We considered three scenarios by differently generating spatial random effects $w_m(S_{ij})$ in Equation (13) by varying the value of the scale parameter of the squared exponential kernel function to be $\sigma_s^2 = 0.1$, 0.5, and 1. Each t_m is randomly chosen in the range from 0 to 1. To make the time series patterns have spatial similarity, the spatial location S_{ij} was generated using a Gaussian mixture distribution that has one or two modes for each cluster as shown in Figure 2(b).

As we explained in Section 4.1, some hyperparameters of the proposed model were fixed manually: the concentration parameters α and β of GEM were both set to 1, the hyperparameters μ_0 , τ , Λ , and ψ of normal-Wishart distribution were set to 0, 1, identity matrix, and 3, respectively. T_1 and T_2 of the truncation levels were set to 15 and 7, respectively. The parameters of the proposed model was estimated using the variational inference method as we discussed in Section 4. The maximum number of iterations for the variational inference method was set to 150: we checked that \mathcal{L}_{VB} tends to converge within 50 iterations as illustrated in Figure 3.

The performance of the proposed model was compared with that of the following eight competing methods:

1. Gaussian process mixture model (GPMM): This model considers temporal correlation and white noise, but does not consider location features and spatial random effects. It is a special case of the proposed model when spatial factors are not considered.
2. GPMM with spatial random effects (GPSRE): This model considers temporal correlation, white noise, and spatial random effects, but does not consider location features. It is a special case of the proposed model when location features are not considered.
3. GPMM with location features (GPLoc): This model considers temporal correlation, white noise, and location features,

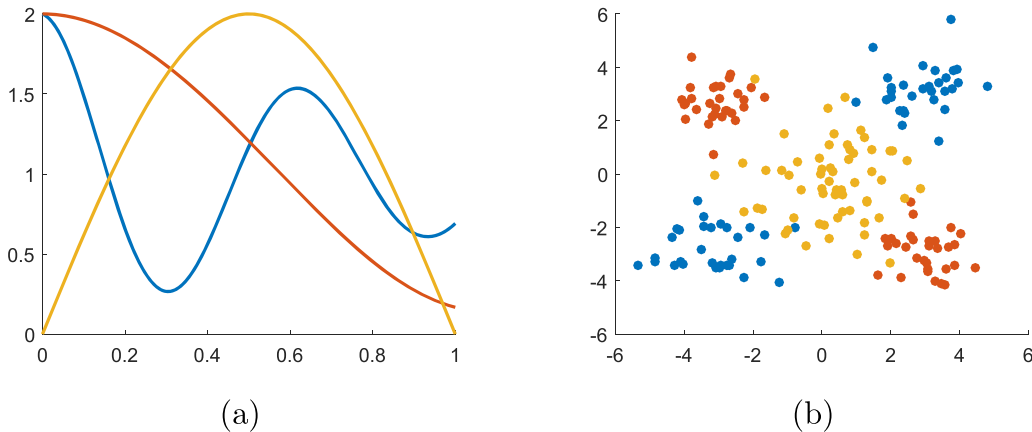


Figure 2. (a) Mean functions $f_1(t)$, $f_2(t)$, and $f_3(t)$ in colors of blue, red, and yellow, respectively; (b) spatial distribution of the synthetic data.

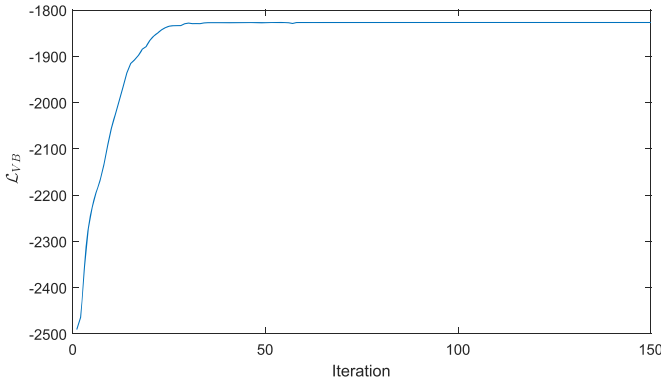


Figure 3. Mean \mathcal{L}_{VB} as a function of iteration of variational inference for the synthetic dataset.

but does not consider spatial random effects. It is a special case of the proposed model when spatial random effects are not considered.

4. Functional spatial clustering model (FSCM) proposed in Jiang and Serban (2012): This model assumes Markov dependency between latent component assignment variables and considers spatially correlated noise in time series.
5. Gaussian mixture model (GMM): This model is a basic parametric mixture model that considers neither temporal correlation nor spatially correlated noise. Each Gaussian distribution in GMM can have full, diagonal, or spherical covariance matrix. In our experiments, full covariance matrices were assumed and estimated.
6. K-means clustering: This method is widely used in cluster analysis because of the simplicity of the algorithm. It aims to divide observations into clusters by minimizing the distance between each observation and the center of the cluster it belongs to.
7. Hierarchical clustering (HC) (Johnson 1967): This method finds a hierarchical structure of cluster by sequentially merging the groups (bottom-up) or by sequentially dividing the groups (top-down), so that observations in the same group have high degree of similarity.
8. Hierarchical clustering with spatial constraints (ClustGeo) (Chavent et al. 2018): This is a discriminative clustering method, which uses a weighted sum of spatial and temporal similarities as a metric for hierarchical clustering. The weights between the spatial and temporal similarities are controlled by a parameter $\alpha \in [0, 1]$.

Except for the proposed model, GPMM, GPSRE, and GPLoc, the five other methods of FSCM, GMM, K-means, HC, and ClustGeo require the determination of the number of clusters in advance. For FSCM, GMM, and K-means, we evaluated the clustering performance by changing the number of clusters from 2 to 5, and selected the number that resulted in the smallest value of AIC. For HC and ClustGeo, the number of clusters was determined by cutting a dendrogram at a height where the similarity between groups changed abruptly (Kim, Lee, and Kim 2018). For the proposed model, GPMM, GPSRE, and GPLoc, the squared exponential kernel function was used. For FSCM, 10 b-spline basis functions were used and the neighborhood structure was constructed by k -nearest-neighbor method with

$k = 5$. For HC, the Euclidean distance was used as the similarity metric and average linkage was used as the linkage criteria. For ClustGeo, the value of α was set to 0.1, because it produced the best clustering performance in our evaluation with various values of α from 0 to 1.

We evaluated the clustering performance using the performance measures of the Rand index (RI) (Rand 1971) and normalized mutual information (NMI) (Strehl and Ghosh 2002). Assume that $X = \{X_1, X_2, \dots, X_k\}$ is the true partition of the dataset \mathcal{D} and $\hat{X} = \{\hat{X}_1, \hat{X}_2, \dots, \hat{X}_l\}$ is the estimated partition of the dataset \mathcal{D} . Then, RI and NMI are defined as

$$RI(X, \hat{X}) = \frac{a + b}{\binom{N}{2}},$$

and

$$NMI(X, \hat{X}) = \frac{-2 \sum_k \sum_l \frac{|X_k \cap \hat{X}_l|}{N} \log \frac{N|X_k \cap \hat{X}_l|}{|X_k||\hat{X}_l|}}{\sum_k \frac{|X_k|}{N} \log \frac{|X_k|}{N} + \sum_l \frac{|\hat{X}_l|}{N} \log \frac{|\hat{X}_l|}{N}},$$

respectively, where N is the number of elements in \mathcal{D} , a is the number of pairs in \mathcal{D} that belong to the same cluster in X and to the same cluster in \hat{X} , and b is the number of pairs in \mathcal{D} that belong to different clusters in X and to the different clusters in \hat{X} . A higher value of RI indicates higher clustering accuracy. The NMI is a normalized version of the mutual information that quantifies how much two partitions X and \hat{X} share the information. The NMI lies between 0 and 1, and it is 1 when the two partitions match completely.

For each of the three considered scenarios, we randomly regenerated samples 300 times, and computed the average of each performance measure over the 300 datasets using each competing method. The results are summarized in Table 2. The third column shows the average of the estimated number of clusters with standard errors in parentheses. The fourth and fifth columns show the average values of the RI and NMI, respectively, with standard errors in parentheses. For the three scenarios, we can see that the proposed model achieves the highest clustering accuracy in terms of both considered performance measures, followed by the three special cases of the proposed model (i.e., GPMM, GPSRE, and GPLoc). The GPSRE and GPLoc achieved better performances than the GPMM. This shows the advantage of incorporating spatial random effects or location features in the model for spatio-temporal clustering. The GPLoc showed a slightly better performance than the GPSRE, demonstrating a slightly more significant contribution of the location features than the spatial random effects. The proposed method showed a consistently good performance for all three scenarios, whereas the other methods exhibited a significantly degraded performance as the variance of the spatial random effects increased. We can also see that the models that consider the temporal correlation (i.e., the proposed model, GPMM, GPSRE, GPLoc, and FSCM) exhibited better performances than the other four methods.

The FSCM, ClustGeo, and K-means achieved the fifth, sixth, and seventh best results, respectively, followed by the GMM. In applying the GMM, the matrix inversion problem was frequently encountered in calculating the full covariance matrix,

Table 2. Clustering results for the simulated examples.

σ_s^2	Method	# of clusters	RI	NMI
0.1	Proposed	3.12 (0.3)	0.95 (0.02)	0.85 (0.06)
	GPMM	3.40 (0.4)	0.92 (0.02)	0.76 (0.06)
	GPSRE	3.10 (0.3)	0.94 (0.02)	0.82 (0.06)
	GPLoc	3.24 (0.4)	0.95 (0.02)	0.84 (0.08)
	FSCM	3.33 (0.4)	0.92 (0.03)	0.77 (0.08)
	GMM	2.00 (0.0)	0.67 (0.05)	0.48 (0.06)
	K-means	5.00 (0.0)	0.82 (0.02)	0.58 (0.05)
	HC	2.09 (1.0)	0.59 (0.19)	0.32 (0.24)
	ClustGeo	3.42 (1.9)	0.90 (0.08)	0.73 (0.07)
0.5	Proposed	3.13 (0.3)	0.94 (0.02)	0.81 (0.08)
	GPMM	3.47 (0.4)	0.90 (0.03)	0.73 (0.09)
	GPSRE	3.26 (0.4)	0.90 (0.05)	0.73 (0.09)
	GPLoc	3.32 (0.5)	0.93 (0.05)	0.77 (0.06)
	FSCM	2.82 (0.4)	0.85 (0.08)	0.65 (0.17)
	GMM	2.00 (0.0)	0.56 (0.06)	0.16 (0.10)
	K-means	5.00 (0.0)	0.74 (0.03)	0.41 (0.07)
	HC	2.59 (1.5)	0.55 (0.15)	0.26 (0.08)
	ClustGeo	2.82 (0.2)	0.87 (0.14)	0.70 (0.13)
1.0	Proposed	3.25 (0.4)	0.91 (0.06)	0.75 (0.07)
	GPMM	3.60 (0.5)	0.87 (0.04)	0.71 (0.08)
	GPSRE	3.20 (0.3)	0.89 (0.02)	0.73 (0.06)
	GPLoc	3.52 (0.6)	0.89 (0.07)	0.73 (0.06)
	FSCM	2.69 (0.7)	0.77 (0.07)	0.49 (0.11)
	GMM	2.00 (0.0)	0.55 (0.06)	0.15 (0.10)
	K-means	5.00 (0.0)	0.73 (0.03)	0.38 (0.08)
	HC	3.99 (2.4)	0.52 (0.17)	0.25 (0.09)
	ClustGeo	2.78 (0.4)	0.73 (0.14)	0.42 (0.11)

which may result in lower performance than the simpler K-means algorithm. HC exhibited the worst performance, maybe because it is sensitive to noise in data (Balcan, Liang, and Gupta 2014). For FSCM, the parameter controlling the local dependency of the cluster assignment variable was estimated to be large (0.5–0.8), but it exhibited a weaker performance than GPMM. In contrast, the proposed model achieved a better performance than GPMM by using location information as a feature in the model. This shows that incorporating location information explicitly in the model has a more direct impact on clustering results. Although the number of clusters was not set in advance, the proposed model estimated the number of clusters more accurately than the other methods. In particular, for GMM, K-means, and HC, the numbers of clusters were chosen to be too small or too large.

To highlight the advantage of the proposed method in incorporating spatial information, Figure 4 compares the clustering results of the proposed method and ClustGeo, a typical discriminative clustering method with a metric that incorporates spatial similarity. When $\alpha = 0$, the ClustGeo does not incorporate spatial information; clustering is performed based only on temporal similarity. The clustering results in this case are shown in Figure 4(a): several misclustered points are observed. With an increased value of $\alpha = 0.1$, some of the misclustered points were corrected by using spatial similarity as shown in Figure 4(b). However, if the influence of spatial information is even more increased with $\alpha = 0.3$, some clusters are further separated into multiple clusters as shown in Figure 4(c). The ClustGeo, like most discriminative clustering methods, uses a predefined distance metric for calculating spatial similarity, which prevents multiple subgroups spaced apart from being clustered together in a single group, even though they have similar temporal pat-

terns. In contrast, the proposed method can cluster multiple subgroups that have similar temporal patterns but are located far away from each other in a single group, as shown in Figure 4(d), by effectively incorporating spatial information. As a generative method, the proposed model can flexibly incorporate spatial information by using the estimated spatial density function that is possibly multimodal. The estimated mean functions for the three clusters detected by the proposed method are shown in Figure 5. Overall, the estimated functions are similar to the true functions, although they tend to be biased near the boundaries, because the randomly selected design points lie approximately between 0.1 and 0.8.

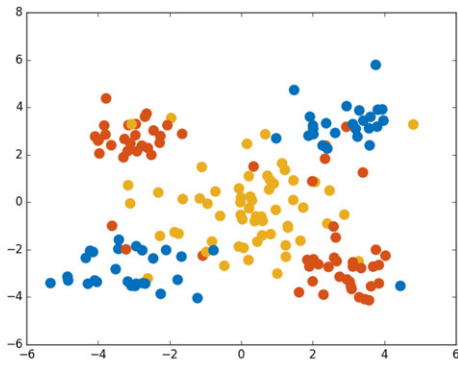
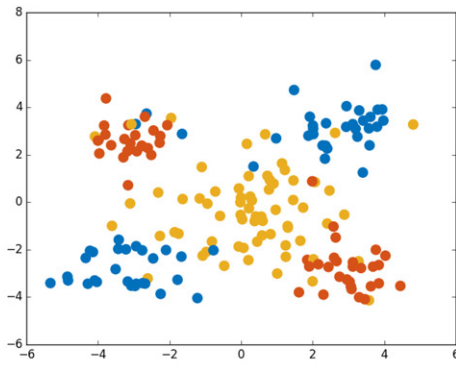
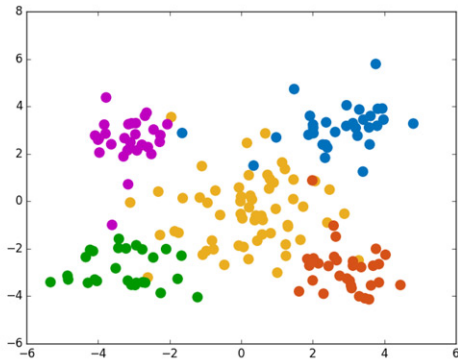
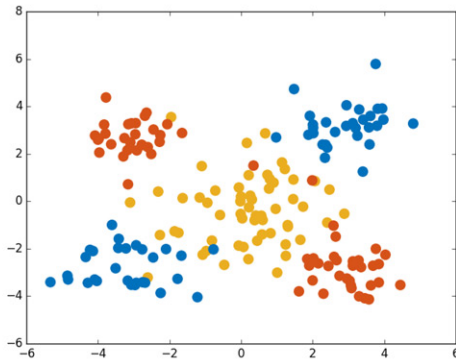
6. Real Data Examples

6.1. Application to Mobility Networks

In mobility network capacity planning, understanding the spatial and temporal characteristics of the mobility network traffic is an important issue for service providers to optimize the network capacity and service quality. For example, understanding the spatial distribution of traffic demand is essential for efficient network planning (Lee et al. 2014). Identifying the temporal traffic patterns is important to predict future traffic demand and establish long-term operational plans (Paul et al. 2011). Understanding the spatial and temporal characteristics of network traffic is also important for identifying events in the mobility network and assessing the event impacts (Au et al. 2010; Kim et al. 2019).

To characterize the mobility network traffic, we conducted a spatio-temporal clustering analysis for cell tower traffic data. We applied the proposed method to the data of request-response records extracted from hypertext transfer protocol (HTTP) traffic collected from an operational cellular network deployed in a popular city of China (Chen et al. 2015). The dataset consists of hourly traffic records from August 19 (Sunday) to 26, 2012 for a total of 13,000 cell towers. For our analysis, we used the data for a total of 5700 cell towers, discarding time series instances with missing rates larger than 30%. The missing values mostly occurred at dawn when normally close-to-zero traffic volumes were observed from other cell towers without missing values. For each time series, the missing values were filled using cubic spline interpolation; however, if negative values were resulted, the nearest neighbor approach was used instead. We normalized each time series by subtracting its mean and dividing by its standard deviation.

In applying the proposed method, we considered the periodicity of the time series by using the sum of the squared exponential function and the periodic kernel function as a kernel function for time series modeling. For the kernel function, the period parameter was set to 24, and the other parameters were set to be the same as in Section 5. Using the proposed method, two clusters were detected. The variance of the spatial random effects (σ_s^2) was estimated as 0.12. Figure 6 visualizes the spatial and temporal characteristics of the clustering results. Figure 6(a) shows the cluster membership for each cell tower by red and blue colors, and Figure 6(b) shows the mean functions of the two clusters. The cluster in red, mainly located in urban regions, shows that traffic in daytime during weekdays is higher than

(a) ClustGeo with $\alpha=0$ (RI=88.8)(b) ClustGeo with $\alpha=0.1$ (RI=90.1)(c) ClustGeo with $\alpha=0.3$ (RI=84.9)

(d) Proposed model (RI=97.1)

Figure 4. Graphical comparison of the clustering results of ClustGeo with various α values and the proposed model.

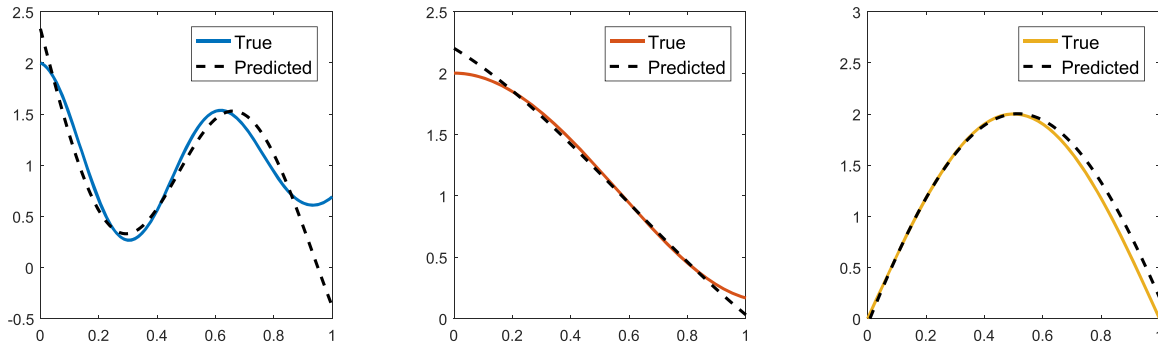


Figure 5. The estimated mean functions for the three clusters.

that of the cluster in blue, which mainly corresponds to rural regions. Moreover, the cluster in red shows a decrease in the overall amount of traffic during weekends. This may be because of pendular movements between home and workplace (Chen et al. 2015). The identified traffic variations between urban and rural regions should be appropriately considered for network capacity planning.

When the competing methods considered in Section 5 were applied, the three special cases of the proposed model (i.e., GPMM, GPSRE, and GPLoc) produced very similar results as ours. However, using FSCM, GMM, and K-means, we could not determine the optimal number of clusters, because the value of AIC decreased as the number of clusters increased. In contrast, HC and ClustGeo detected only a single cluster.

6.2. Application to Brain Imaging

Blood oxygen level-dependent (BOLD) contrast, based on fMRI, has emerged as a powerful tool for identifying activated regions of the brain in relation to cognitive processes (Simmonds, Pekar, and Mostofsky 2008; Zhang et al. 2016). Because the area of the brain responsible for a specific cognitive ability appears to spatially coalesce into several areas, the voxels corresponding to activated regions are likely to exist in several chunks, which indicates that location information on voxel coordinates is also an important factor for finding activated regions. Moreover, head movements, heartbeat, breathing, and other physiological factors create spatially correlated noise in BOLD signals, and this should be considered in the analysis.

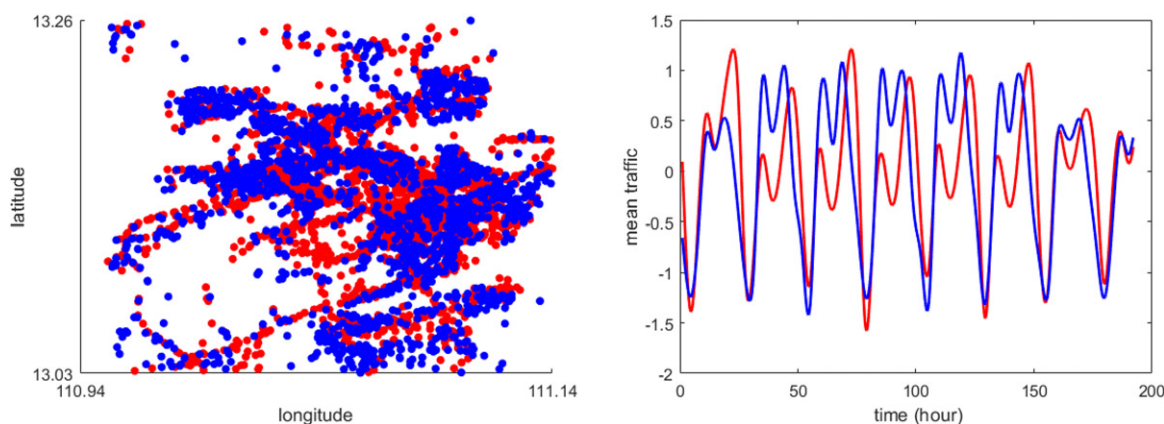


Figure 6. The clustering results of the proposed model for the traffic data.

These points make our model appealing in analyzing fMRI data, although the fMRI data locations are regularly spaced voxels, unlike we assume that spatial locations are random and follow a DP Gaussian mixture model.

We applied the proposed model to single subject epoch (block) auditory fMRI activation data, which is provided by the Wellcome Trust Centre for Neuroimaging at University College London. The dataset was acquired on a modified 2T Siemens MAGNETOM Vision system, and is composed of 96 scans of a human brain BOLD image with a 7 sec repetition time (Rees 1999). The fMRI experiment for this dataset was designed with successive blocks alternating between rest and auditory stimulation, beginning with rest. Auditory stimulation consisted of bisyllabic words presented binaurally at a rate of 60 per minute. Each block consists of six scans, each involving $64 \times 64 \times 64$ voxels. After preprocessing the data using SPM12 software, the volume was changed to size $53 \times 63 \times 46$ (Friston, Jezzard, and Turner 1994). In our experiments, we removed the first 12 scans to avoid T1 equilibration effects and used a 53×63 voxel matrix coronal plane at $Z = 25$. All BOLD signals were mean-centered, and a linear trend was removed. To consider the periodicity of the signals, we used the sum of the squared exponential function and the periodic kernel function as a kernel function for time series modeling. Instead of the squared exponential function, we also tried the exponential function, a less smooth kernel, for evaluating the sensitivity, but we obtained the same clustering results. For the periodic kernel function, the period parameter was set to 12, and the other parameters were set to be the same as in Section 5.

The performance of the proposed model was compared with the competing methods considered in Section 5. The parameters of the competing methods were specified in the same manner as in Section 5. To detect the activated regions, we calculated the correlation coefficient between the mean function of each cluster and the box-car function convolved with a hemodynamic response function (Woolrich et al. 2001), and selected the clusters with the correlation coefficient larger than 0.5 as the activated regions, for all competing methods except for ClustGeo; using this approach, we could not find the activated clusters for ClustGeo. The five clusters identified using the proposed method are depicted in Figure 7, together with the mean

function of each cluster. Among the five clusters, the cluster in Figure 7(a) was detected as the activated regions, because the corresponding correlation coefficient was more than 0.7; the correlation coefficients for other clusters were near zero.

Figure 8 compares the activated regions detected by the proposed model and the other competing methods. Besides the competing methods considered in the previous experiments, we considered three additional methods: statistical parametric mapping (SPM), independent component analysis (ICA), and self-organizing map (SOM), as these methods are commonly employed tools for fMRI time series analysis. SPM is a statistical method, based on the general linear model. Unlike all the other methods, SPM uses an experimental design matrix that contains the information of an experimental session (Friston, Jezzard, and Turner 1994). In our experiments using the SPM, we first fitted a general linear model. Then, the estimated model parameters were used to test for activations with p -value = 0.05. Figures 8(i) and 8(j) show the SPM results with and without the Bonferroni correction for multiple comparisons, respectively. ICA is a dimension reduction method, which is generally used for finding inherently non-Gaussian features in signal datasets (Cheung and Xu 2001). SOM is a type of neural network model that constructs a topologically ordered feature map, and it is used for visualizing low-dimensional features for high-dimensional data (Bernard et al. 2012).

From Figure 8, we can see that although the proposed model did not use the prior knowledge of the design matrix, its detected regions are similar to those from the SPM. Moreover, we find that the detected regions using the proposed model almost coincide with the regions of the auditory cortex. Furthermore, compared to the results of the other methods, the regions detected by the proposed model and GPLoc were not scattered. This may be because the proposed model and GPLoc tend to avoid clustering data points with low spatial similarities. Compared with the proposed method, the GPSRE additionally detected several tiny regions of activations. This may be related to a larger estimate of the variance of the spatial random effects for the GPSRE ($\hat{\sigma}_s^2 = 24.2$) compared with that for the proposed method ($\hat{\sigma}_s^2 = 17.4$). The lower variance of the spatial random effects for the proposed model may result from the additional incorporation of location features.

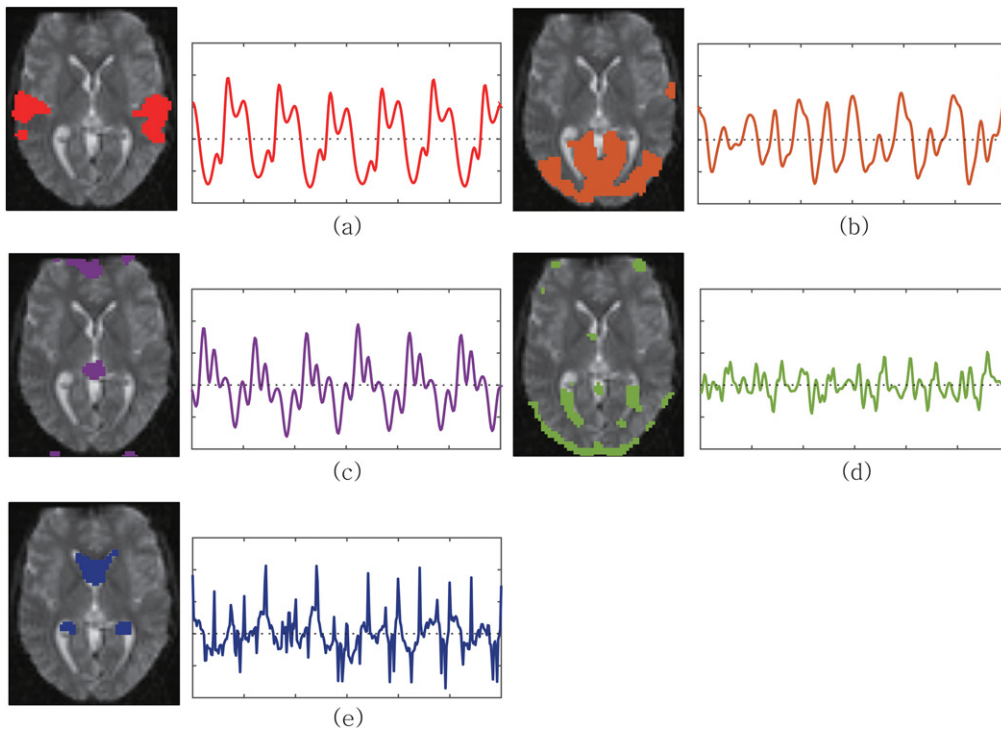


Figure 7. The clustering results of the proposed model for the fMRI data.

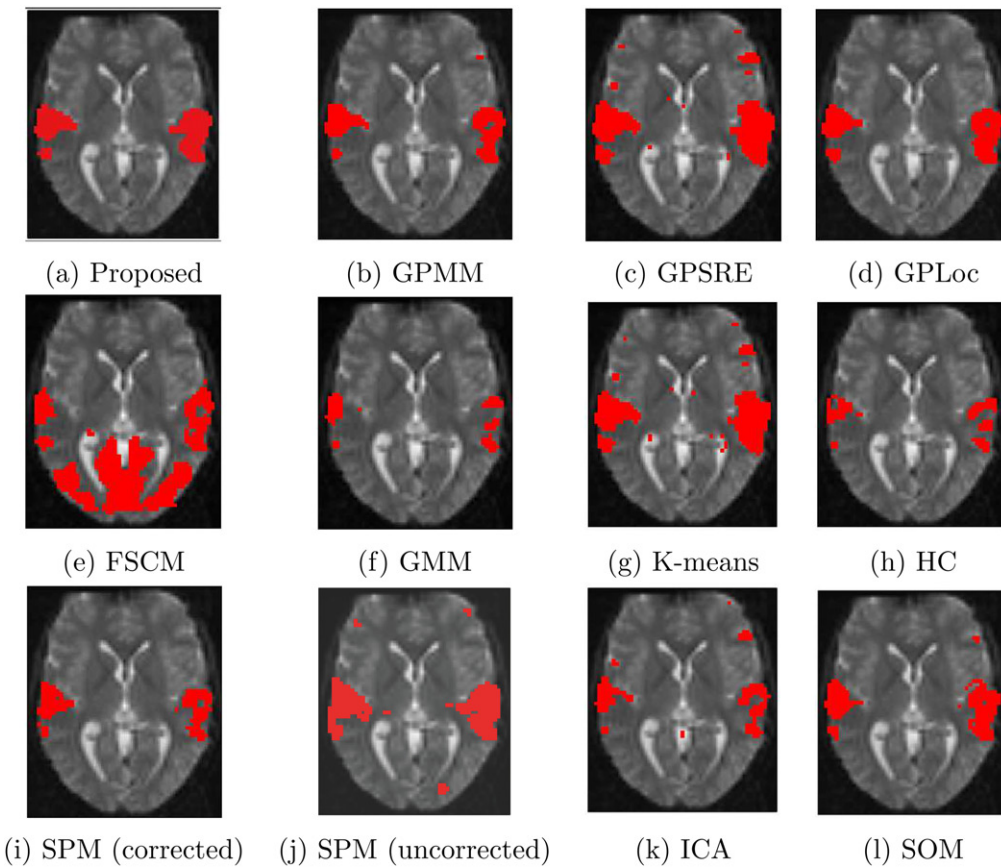


Figure 8. The estimated activated regions using several competing methods.

7. Conclusion

The challenge of finding meaningful patterns in spatio-temporal data has received considerable attention. In this article, we pro-

posed a Bayesian nonparametric joint mixture model, which incorporates spatial as well as temporal information as features for clustering spatially correlated time series. Using Bayesian nonparametric priors that help give the proposed model flexi-

bility and robustness, and to represent the spatial distribution of each cluster as a DP mixture, the proposed model enables the capture of unexposed complex spatial structures in cluster components.

Although the inference based on variational Bayes enables a faster implementation than a sampling based inference, some computational issues still remain. A large volume of locations or time points may result in a matrix inversion problem in updating variational parameters of mean time series or spatial random effects. Two possible strategies for remedying this problem can be considered. Titsias (2009) proposed a sparse approximation of GP using pseudo inputs for a variational method, and this approach allows the efficient computation of the sparse posterior GP. We could also introduce incomplete Cholesky decomposition (Bach and Jordan 2002) for the covariance matrix of spatial random effects, as in Jiang and Serban (2012). In this article, we focused on spatio-temporal data that are observed at the same time points for each time series. However, the proposed model could easily be extended to data containing different observed time points with different numbers of observations for each time series. Also, the proposed model can be extended to consider nonstationary spatial or temporal correlation (Gelfand, Kottas, and MacEachern 2005; Rodríguez and Dunson 2011).

Appendix A. Derivation of the Update Equations for Variational Parameters

A.1. The Lower Bound \mathcal{L}_{VB} in Equation (11)

Recall that in variational inference, we maximize the lower bound \mathcal{L}_{VB} in Equation (11)

$$\begin{aligned}\mathcal{L}_{VB} = & E_q[\log p(Y|Z, F, W, S)] + E_q[\log p(F)] + E_q[\log p(W|S)] \\ & + E_q[\log p(S|Z, H, \Theta_{\mu, \Omega})] + E_q[\log p(Z|\omega)] \\ & + E_q[\log p(H|Z, \Theta_v)] + E_q[\log p(\omega)] + E_q[\log p(\Theta_v)] \\ & + E_q[\log p(\Theta_{\mu, \Omega})] - E_q[\log q(F, W, Z, H, \omega, \Theta_v, \Theta_{\mu, \Omega})].\end{aligned}$$

Each expectation term in \mathcal{L}_{VB} can be expanded as follows.

A.1.1. $E_q[\log p(Y|Z, F, W, S)]$

$$\begin{aligned}E_q[\log p(Y|Z, F, W, S)] &= E_q \left[\log \prod_{i=1}^N \prod_{c=1}^C \prod_{m=1}^M N(Y_i(t_m) | f_c(t_m) + w_m(S_i), \sigma_\epsilon^2)^{Z_{i,c}} \right] \\ &= \sum_{i=1}^N \sum_{c=1}^{T_1} \phi_{i,c}^Z \sum_{m=1}^M E_q[\log N(Y_i(t_m) | f_c(t_m) + w_m(S_i), \sigma_\epsilon^2)] \\ &= \sum_{i=1}^N \sum_{c=1}^{T_1} \phi_{i,c}^Z \sum_{m=1}^M \left(-\frac{1}{2} \log 2\pi - \frac{1}{2} \log \sigma_\epsilon^2 \right. \\ &\quad \left. - \frac{(Y_i(t_m) - \mu_c^F(t_m) - \mu_m^W(S_i))^2 + |\Sigma_c^F|_{mm} + |\Sigma^W|_{ii}}{2\sigma_\epsilon^2} \right).\end{aligned}$$

A.1.2. $E_q[\log p(F)]$

$$E_q[\log p(F)] = \sum_{c=1}^C E_q[\log \mathcal{GP}(f_c | 0, k_c(t, t'))]$$

$$\begin{aligned}&= \sum_{c=1}^C E_q \left[-\frac{M}{2} \log 2\pi - \frac{1}{2} \log |K_c| - \frac{1}{2} f_c^T K_c^{-1} f_c \right] \\ &= \sum_{c=1}^C -\frac{M}{2} \log 2\pi - \frac{1}{2} \log |K_c| \\ &\quad - \frac{1}{2} \left(\text{tr}(K_c^{-1} \Sigma_c^F) + \mu_c^F{}^T K_c^{-1} \mu_c^F \right).\end{aligned}$$

A.1.3. $E_q[\log p(W|S)]$

$$\begin{aligned}E_q[\log p(W|S)] &= \sum_{m=1}^M E_q[\log N(W_m | 0, \Sigma)] \\ &= \sum_{m=1}^M E_q \left[-\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} W_m^T \Sigma^{-1} W_m \right] \\ &= \sum_{m=1}^M -\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| \\ &\quad - \frac{1}{2} \left(\text{tr}(\Sigma^{-1} \Sigma^W) + (\mu_m^W)^T (\Sigma^W)^{-1} \mu_m^W \right).\end{aligned}$$

A.1.4. $E_q[p(S|Z, H, \Theta_{\mu, \Omega})]$

$$\begin{aligned}E_q[p(S|Z, H, \Theta_{\mu, \Omega})] &= E_q \left[\prod_{i=1}^N \prod_{c=1}^C \prod_{l=1}^L N(S_i | \mu_{c,l}, \Omega_{c,l}^{-1})^{Z_{i,c} H_{i,c,l}} \right] \\ &= \sum_{i=1}^N \sum_{c=1}^T \phi_{i,c}^Z \sum_{l=1}^{T_2} \phi_{i,c,l}^H E_q \left[\log N(S_i | \mu_{c,l}, \Omega_{c,l}^{-1}) \right] \\ &= \sum_{i=1}^N \sum_{c=1}^{T_1} \phi_{i,c}^Z \sum_{l=1}^{T_2} \phi_{i,c,l}^H \left(-\frac{d}{2} \log 2\pi - \frac{d}{2\tau_{c,l}^S} \right. \\ &\quad \left. - \frac{\psi_{c,l}^S}{2} (S_i - \mu_{c,l}^S)^T \Lambda_{c,l}^S (S_i - \mu_{c,l}^S) - \frac{1}{2} E_q[\log |\Omega_{c,l}|] \right),\end{aligned}$$

where $E_q[\log |\Omega_{c,l}|] = \sum_{o=1}^d \Psi \left(\frac{\psi_{c,l}^S + 1 - o}{2} \right) + d \log 2 + \log |\Lambda_{c,l}^S|$ and $\Psi(\cdot)$ is a digamma function.

A.1.5. $E_q[\log p(Z|\omega)]$

$$\begin{aligned}E_q[p(Z|\omega)] &= \prod_{i=1}^N E_q \left[\log \prod_{i=1}^N q(Z_i | \omega) \right] \\ &= \sum_{i=1}^N E_q \left[\log \left(\prod_{c=1}^C (1 - B_c^\omega)^{\delta_{Z_i > c}} B_c^\omega \delta_{Z_i = c} \right) \right] \\ &= \sum_{i=1}^N \sum_{c=1}^{T_1} (q(Z_i > c) E_q[\log(1 - B_c^\omega)] + q(Z_i = c) E_q[\log B_c^\omega]) \\ &= \sum_{i=1}^N \sum_{c=1}^{T_1} \left(\left(\sum_{o=c+1}^{T_1} \phi_{i,o}^Z \right) (\Psi(b_c^{B^\omega}) - \Psi(a_c^{B^\omega} + b_c^{B^\omega})) \right. \\ &\quad \left. + \phi_{i,c}^Z (\Psi(a_c^{B^\omega}) - \Psi(a_c^{B^\omega} + b_c^{B^\omega})) \right).\end{aligned}$$

A.1.6. $E_q[\log p(H|Z, \Theta_v)]$

$$\begin{aligned}
E_q[\log p(H|Z, \Theta_v)] &= E_q \left[\log \prod_{i=1, c=1}^{N, C} q(H_{i,c} | \mathbf{v}_c)^{Z_{i,c}} \right] \\
&= \sum_{i=1}^N \sum_{c=1}^{T_1} \phi_{i,c}^Z E_q \left[\log \prod_{l=1}^L (1 - B_{c,l}^{\mathbf{v}})^{\delta_{H_{i,c} > l}} B_{c,l}^{\mathbf{v}}^{\delta_{H_{i,c} = l}} \right] \\
&= \sum_{i=1}^N \sum_{c=1}^{T_1} \phi_{i,c}^Z \sum_{l=1}^{T_2} \left(\left(\sum_{o=l+1}^{T_2} \phi_{i,k,o}^H \right) (\Psi(a_{c,l}^{\mathbf{v}}) - \Psi(a_{c,l}^{\mathbf{v}} + b_{c,l}^{\mathbf{v}})) \right. \\
&\quad \left. + \phi_{i,c,l}^H (\Psi(a_{c,l}^{\mathbf{v}}) - \Psi(a_{c,l}^{\mathbf{v}} + b_{c,l}^{\mathbf{v}})) \right).
\end{aligned}$$

A.1.7. $E_q[\log P(\omega)]$

$$\begin{aligned}
E_q[\log P(\omega)] &= E_q \left[\log \prod_{c=1}^C \text{Beta}(B_c^{\omega} | 1, \alpha) \right] \\
&= \sum_{c=1}^C E_q[\log \text{Beta}(B_c^{\omega} | 1, \alpha)] \\
&= \sum_{c=1}^{T_1-1} \left[\log \Gamma(1 + \alpha) - \log \Gamma(\alpha) + (\alpha - 1)(\Psi(b_c^{\omega}) \right. \\
&\quad \left. - \Psi(a_c^{\omega} + b_c^{\omega})) \right] + \text{const},
\end{aligned}$$

where $\Gamma(\cdot)$ is a gamma function.

A.1.8. $E_q[\log p(\Theta_v)]$ and $E_q[\log p(\Theta_{\mu, \Omega})]$

Recall that $p(\Theta_v) = \prod_{c=1}^C p(\mathbf{v}_c)$ and $p(\Theta_{\mu, \Omega}) = \prod_{c=1}^C p(\mu_c, \Omega_c)$. $E_q[\log p(\mathbf{v}_c)]$ and $E_q[\log p(\mu_c, \Omega_c)]$ can be expanded as follows

$$\begin{aligned}
E_q[\log p(\mathbf{v}_c)] &= E_q \left[\log \prod_{l=1}^{T_2-1} \text{Beta}(B_{c,l}^{\mathbf{v}} | 1, \beta) \right] \\
&= \sum_{l=1}^{T_2-1} E_q[\log \text{Beta}(B_{c,l}^{\mathbf{v}} | 1, \beta)] \\
&= \sum_{l=1}^{T_2-1} \left[\log \Gamma(1 + \beta) - \log \Gamma(\beta) + (\beta - 1)(\Psi(b_{c,l}^{\mathbf{v}}) \right. \\
&\quad \left. - \Psi(a_{c,l}^{\mathbf{v}} + b_{c,l}^{\mathbf{v}})) \right] + \text{const},
\end{aligned}$$

and

$$\begin{aligned}
E_q[\log p(\mu_c, \Omega_c)] &= E_q \left[\log \prod_{l=1}^L \mathcal{N}(\mu_{c,l}, \Omega_{c,l} | \mu_0, \tau, \Lambda, \psi) \right] \\
&= \sum_{l=1}^L E_q[\log \mathcal{N}(\mu_{c,l}, \Omega_{c,l} | \mu_0, \tau, \Lambda, \psi)] \\
&= \frac{1}{2} \sum_{l=1}^L \left(d \log \frac{\tau}{2\pi} + E_q[\log |\Omega_{c,l}|] - \frac{d\tau}{\tau_{c,l}^S} \right. \\
&\quad \left. - \tau \psi_{c,l}^S (\mu_{c,l}^S - \mu_0)^T \Lambda_{c,l}^S (\mu_{c,l}^S - \mu_0) \right. \\
&\quad \left. + \log B(\Lambda, \psi) + \frac{\psi - d - 1}{2} E_q[\log |\Omega_{c,l}|] - \frac{1}{2} \psi_{c,l}^S \text{Tr}(\Lambda^{-1} \Lambda_{c,l}^S) \right),
\end{aligned}$$

where $B(\Lambda, \psi) = |\Lambda|^{-\psi/2} \left(2^{\psi d/2} \pi^{d(d-1)/4} \prod_o^d \Gamma(\frac{\psi+1-o}{2}) \right)^{-1}$ and $\text{Tr}(A)$ denotes the trace of matrix A .

A.1.9. $E_q[\log q(F, W, Z, H, \omega, \Theta_v, \Theta_{\mu, \Omega})]$

Recall that $q(F, W, Z, H, \omega, \Theta) = q(F)q(W)q(Z)q(H)q(\omega)q(\Theta_v)$, $q(\Theta_v) = \prod_{c=1}^{T_1} q(\mathbf{v}_c)$ and $q(\Theta_{\mu, \Omega}) = \prod_{c=1}^{T_1} q(\mu_c, \Omega_c) = \prod_{c=1, l=1}^{T_1, T_2} q(\mu_{c,l}, \Omega_{c,l})$. Each log-expectation of factorized variational distribution can be expanded as follows

$$\begin{aligned}
E_q[\log q(F)] &= \sum_{c=1}^{T_1} E_q \left[\log q(f_c | \mu_c^F, \Sigma_c^F) \right] \\
&= \sum_{c=1}^{T_1} E_q \left[-\frac{M}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_c^F| \right. \\
&\quad \left. - \frac{1}{2} (f_c - \mu_c^F)^T (\Sigma_c^F)^{-1} (f_c - \mu_c^F) \right] \\
&= \sum_{c=1}^{T_1} \left(-\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_c^F| - \frac{M}{2} \right),
\end{aligned}$$

$$\begin{aligned}
E_q[\log q(W)] &= \sum_{m=1}^M E_q[\log q(W_m | \mu_m^W, \Sigma^W)] \\
&= \sum_{m=1}^M E_q \left[-\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\Sigma^W| \right. \\
&\quad \left. - \frac{1}{2} (W_m - \mu_m^W)^T (\Sigma^W)^{-1} (W_m - \mu_m^W) \right] \\
&= \sum_{m=1}^M \left(-\frac{N}{2} \log 2\pi - \frac{1}{2} \log |\Sigma^W| - \frac{N}{2} \right),
\end{aligned}$$

$$\begin{aligned}
E_q[\log q(Z)] &= E_q \left[\log \prod_{i=1}^N q(Z_i | \phi_i^Z) \right] = E_q \left[\log \prod_{i=1, c=1}^{N, T_1} (\phi_{i,c}^Z)^{Z_{i,c}} \right] \\
&= \sum_{i=1, c=1}^{N, T_1} E_q[\log (\phi_{i,c}^Z)^{Z_{i,c}}] = \sum_{i=1, c=1}^{N, T_1} \phi_{i,c}^Z \log \phi_{i,c}^Z,
\end{aligned}$$

$$\begin{aligned}
E_q[\log q(H)] &= E_q \left[\log \prod_{i=1, c=1}^{N, T_1} q(H_{i,c} | \phi_{i,c}^H) \right] \\
&= \sum_{i=1, c=1}^{N, T_1} E_q \left[\log \prod_{l=1}^{T_2} (\phi_{i,c,l}^H)^{H_{i,c,l}} \right] \\
&= \sum_{i=1, c=1, l=1}^{N, T_1, T_2} \phi_{i,c,l}^H \log \phi_{i,c,l}^H,
\end{aligned}$$

$$\begin{aligned}
E_q[\log q(\omega)] &= E_q \left[\log \prod_{c=1}^{T_1-1} q(B_c^{\omega} | a_c^{\omega}, b_c^{\omega}) \right] \\
&= \sum_{c=1}^{T_1-1} E_q[\log \text{Beta}(B_c^{\omega} | a_c^{\omega}, b_c^{\omega})] \\
&= \sum_{c=1}^{T_1-1} \left[\log \Gamma(a_c^{\omega} + b_c^{\omega}) - \log \Gamma(a_c^{\omega}) - \log \Gamma(b_c^{\omega}) \right. \\
&\quad \left. + (a_c^{\omega} - 1)(\Psi(a_c^{\omega}) - \Psi(a_c^{\omega} + b_c^{\omega})) \right. \\
&\quad \left. + (b_c^{\omega} - 1)(\Psi(b_c^{\omega}) - \Psi(a_c^{\omega} + b_c^{\omega})) \right],
\end{aligned}$$

Table A.1. A summary of update information for all variational parameters.

Variable	Parameters	Update information
f_c	μ_c^F, Σ_c^F	$\mu_c^F = [K_c^{-1} + \sum_{i=1}^N \frac{\phi_{i,c}^Z}{\sigma_\epsilon^2} I]^{-1} \sum_{i=1}^N \frac{\phi_{i,c}^Z}{\sigma_\epsilon^2} (Y_i - \{\mu_m^W(S_i)\}_{m=1}^M),$ $\Sigma_c^F = [K_c^{-1} + \sum_{i=1}^N \frac{\phi_{i,c}^Z}{\sigma_\epsilon^2} I]^{-1}.$
W_m	μ_m^W, Σ^W	$\mu_m^W = \frac{1}{\sigma_\epsilon^2} (\Sigma^{-1} + \frac{1}{\sigma_\epsilon^2} I)^{-1} (\{y_i(t_m)\}_{i=1}^N - \sum_{i=1}^N Q_i \{\mu_c^F(t_m)\}_{c=1}^{T_1}),$ $\Sigma^W = (\Sigma^{-1} + \frac{1}{\sigma_\epsilon^2} I)^{-1},$ where Q_i is a $T_1 \times T_1$ diagonal matrix with the i th diagonal element of $\phi_{i,c}^Z$.
Z_i	$\phi_{i,c}^Z$	$\phi_{i,c}^Z \propto \exp\{(\Psi(a_c^{\beta\omega}) - \Psi(a_c^{\beta\omega} + b_c^{\beta\omega}) + \sum_{v=1}^{c-1} \Psi(b_v^{\beta\omega}) - \Psi(a_v^{\beta\omega} + b_v^{\beta\omega}))$ $+ (-\frac{M}{2} \log 2\pi \sigma_\epsilon^2 - \frac{(Y_i - \mu_c^F - \mu_i^W)^T (Y_i - \mu_c^F - \mu_i^W) + \text{tr}[\Sigma_c^F] + M \Sigma^W _{ii})}{2\sigma_\epsilon^2})$ $+ \sum_{l=1}^{T_2} \phi_{i,c,l}^H ((\Psi(a_{c,l}^{\beta\nu}) - \Psi(a_{c,l}^{\beta\nu} + b_{c,l}^{\beta\nu})) + \sum_{o=1}^{l-1} (\Psi(a_{c,o}^{\beta\nu}) - \Psi(a_{c,o}^{\beta\nu} + b_{c,o}^{\beta\nu}))$ $- \log 2\pi + \frac{1}{2} \log E_q[\log \Omega_{c,l}] - \frac{d}{2\tau_{c,l}^S}$ $- \frac{\psi_{c,l}^S}{2} (S_i - \mu_{c,l}^S)^T \Lambda_{c,l}^S (S_i - \mu_{c,l}^S))\}.$
$H_{i,c}$	$\phi_{i,c,l}^H$	$\phi_{i,c,l}^H \propto \exp(\Psi(a_{c,l}^{\beta\nu}) - \Psi(a_{c,l}^{\beta\nu} + b_{c,l}^{\beta\nu}))$ $+ \sum_{o=1}^{l-1} (\Psi(a_{c,o}^{\beta\nu}) - \Psi(a_{c,o}^{\beta\nu} + b_{c,o}^{\beta\nu})) - \log 2\pi + \frac{1}{2} \log E_q[\log \Omega_{c,l}]$ $- \frac{d}{2\tau_{c,l}^S} - \frac{\psi_{c,l}^S}{2} (S_i - \mu_{c,l}^S)^T \Lambda_{c,l}^S (S_i - \mu_{c,l}^S).$
$B_c^{\beta\omega}$	$a_c^{\beta\omega}, b_c^{\beta\omega}$	$a_c^{\beta\omega} = \sum_{i=1}^N \phi_{i,c}^Z + 1,$ $b_c^{\beta\omega} = \sum_{i=1}^N \sum_{o=c+1}^{T_1} \phi_{i,o}^Z + \alpha.$
$B_{c,l}^{\beta\nu}$	$a_{c,l}^{\beta\nu}, b_{c,l}^{\beta\nu}$	$a_{c,l}^{\beta\nu} = \sum_{i=1}^N \phi_{i,c}^Z \phi_{i,c,l}^H + 1,$ $b_{c,l}^{\beta\nu} = \sum_{i=1}^N \phi_{i,c}^Z \sum_{o=c+1}^{T_2} \phi_{i,c,o}^H + \beta.$
$\mu_{c,l}, \Omega_{c,l}$	$\mu_{c,l}^S, \tau_{c,l}^S, \Lambda_{c,l}^S, \psi_{c,l}^S$	$\mu_{c,l}^S = \frac{\tau\mu_0 + \sum_{i=1}^N \phi_{i,c}^Z \phi_{i,c,l}^H S_i}{\tau + \Phi_{c,l}},$ $\tau_{c,l}^S = \tau + \bar{\Phi}_{c,l},$ $\Lambda_{c,l}^S = \Lambda^{-1} + \sum_{i=1}^N \phi_{i,c}^Z \phi_{i,c,l}^H (S_i - \bar{S}_{c,l})(S_i - \bar{S}_{c,l})^T$ $+ \frac{\tau\bar{\Phi}_{c,l}}{\tau + \bar{\Phi}_{c,l}} (\bar{S}_{c,l} - \mu_0)(\bar{S}_{c,l} - \mu_0)^T,$ $\psi_{c,l}^S = \psi + \bar{\Phi}_{c,l} + 1,$ where $\bar{\Phi}_{c,l} = \sum_{i=1}^N \phi_{i,c}^Z \phi_{i,c,l}^H$ and $\bar{S}_{c,l} = \frac{\sum_{i=1}^N \phi_{i,c}^Z \phi_{i,c,l}^H S_i}{\bar{\Phi}_{c,l}}.$

$$\begin{aligned}
E_q[\log q(\nu_c)] &= E_q \left[\log \prod_{l=1}^{T_2-1} q(B_{c,l}^{\beta\nu} | a_{c,l}^{\beta\nu}, b_{c,l}^{\beta\nu}) \right] \\
&= \sum_{l=1}^{T_2-1} E_q[\log \text{Beta}(B_{c,l}^{\beta\nu} | a_{c,l}^{\beta\nu}, b_{c,l}^{\beta\nu})] \\
&= \sum_{l=1}^{T_2-1} [\log \Gamma(a_{c,l}^{\beta\nu} + b_{c,l}^{\beta\nu}) - \log \Gamma(a_{c,l}^{\beta\nu}) - \log \Gamma(b_{c,l}^{\beta\nu}) \\
&\quad + (a_{c,l}^{\beta\nu} - 1)(\Psi(a_{c,l}^{\beta\nu}) - \Psi(a_{c,l}^{\beta\nu} + b_{c,l}^{\beta\nu})) \\
&\quad + (b_{c,l}^{\beta\nu} - 1)(\Psi(b_{c,l}^{\beta\nu}) - \Psi(a_{c,l}^{\beta\nu} + b_{c,l}^{\beta\nu}))],
\end{aligned}$$

and

$$\begin{aligned}
E_q[\log q(\mu_c, \Omega_c)] &= E_q \left[\log \prod_{l=1}^{T_2} \mathcal{NW}(\mu_{c,l}, \Omega_{c,l} | \mu_{c,l}^S, \tau_{c,l}^S, \Lambda_{c,l}^S, \psi_{c,l}^S) \right] \\
&= \sum_{l=1}^{T_2} E_q[\log \mathcal{NW}(\mu_{c,l}, \Omega_{c,l} | \mu_{c,l}^S, \tau_{c,l}^S, \Lambda_{c,l}^S, \psi_{c,l}^S)] \\
&= \sum_{l=1}^{T_2} \frac{1}{2} \left(E_q[\log |\Omega_{c,l}|] + d \log \frac{\tau_{c,l}^S}{2\pi} - d \right) - H[\Omega_{c,l}],
\end{aligned}$$

where $H[\Omega_{c,l}] = -B(\Lambda_{c,l}^S, \psi_{c,l}^S) - \frac{\psi_{c,l}^S - d - 1}{2} E_q[\log |\Omega_{c,l}|] + \frac{\psi_{c,l}^S d}{2}$ is the information entropy of $\Omega_{c,l}$ and $B(\Lambda_{c,l}^S, \psi_{c,l}^S) = |\Lambda_{c,l}^S|^{-\psi_{c,l}^S/2} \left(2^{\psi_{c,l}^S d/2} \pi^{d(d-1)/4} \prod_{o=1}^d \Gamma(\frac{\psi_{c,l}^S + 1 - o}{2}) \right)^{-1}.$

A.2. Update Information for Variational Parameters

As described in Section 4.1, update equations of all variational parameters are obtained in closed form as in Table A.1.

A.3. Update for Concentration Parameters

The simplest way to update the concentration parameters α and β for GEM prior is using the gradient method like the other hyperparameters. For α, β , we just need to derive the gradients $\frac{\partial \mathcal{L}_{VB}(\omega)}{\partial \alpha}$ and $\frac{\partial \sum_{c=1}^{T_2-1} \mathcal{L}_{VB}(\nu_c)}{\partial \alpha}$, respectively.

Alternatively, we can apply the fully Bayesian framework by setting the prior distribution for the concentration parameters. Let the prior distribution and variational distribution for α be Gamma(ι_1, ι_2) and Gamma($\iota_1^\alpha, \iota_2^\alpha$), respectively, where ι_1 and ι_1^α are shape parameters and ι_2 and ι_2^α are scale parameters. Then additional variational expectations

are

$$\begin{aligned}
E_q[\log P(\boldsymbol{\omega})] &= E_q \left[\log \prod_{c=1}^C \text{Beta}(B_c^{\boldsymbol{\omega}} | 1, \alpha) \right] \\
&= \sum_{c=1}^C E_q[\log \text{Beta}(B_c^{\boldsymbol{\omega}} | 1, \alpha)] \\
&= \int \left(\sum_{c=1}^{T_1-1} \left[\log \Gamma(1 + \alpha) - \log \Gamma(\alpha) + (\alpha - 1)(\Psi(b_c^{\boldsymbol{\omega}}) \right. \right. \\
&\quad \left. \left. - \Psi(a_c^{\boldsymbol{\omega}} + b_c^{\boldsymbol{\omega}})) \right] + \text{const} \right) dq(\alpha) \\
&= \sum_{c=1}^{T_1-1} [\Psi(\iota_1^\alpha) - \log \iota_2^\alpha + \frac{\iota_1^\alpha}{\iota_2^\alpha} (\Psi(b_c^{\boldsymbol{\omega}}) - \Psi(a_c^{\boldsymbol{\omega}} + b_c^{\boldsymbol{\omega}}))] + \text{const} \\
E_q[\log p(\alpha)] &= (\iota_1 - 1)(\Psi(\iota_1^\alpha) - \log \iota_2^\alpha) - \frac{\iota_1^\alpha}{\iota_2 \iota_2^\alpha} \\
E_q[\log q(\alpha)] &= -\log \Gamma(\iota_1^\alpha) + \iota_1^\alpha \log \iota_2^\alpha + (\iota_1^\alpha - 1)(\Psi(\iota_1^\alpha) - \log \iota_2^\alpha) - 1.
\end{aligned}$$

Denotes $\mathcal{L}_{\text{VB}}(\iota_1^\alpha, \iota_2^\alpha)$ to $E_q[\log P(\boldsymbol{\omega})] + E_q[\log p(\alpha)] - E_q[\log q(\alpha)]$. Then, the partial derivatives for ι_1^α and ι_2^α is as follows.

$$\begin{aligned}
\frac{\partial \mathcal{L}_{\text{VB}}(\iota_1^\alpha, \iota_2^\alpha)}{\partial \iota_1^\alpha} &= (\iota_1 + T_1 - 1 - \iota_1^\alpha) \Psi'(\iota_1^\alpha) \\
&\quad + \frac{\sum_{c=1}^{T_1-1} (\Psi(b_c^{\boldsymbol{\omega}}) - \Psi(a_c^{\boldsymbol{\omega}} + b_c^{\boldsymbol{\omega}})) - \iota_2}{\iota_2^\alpha} + 1 \\
\frac{\partial \mathcal{L}_{\text{VB}}(\iota_1^\alpha, \iota_2^\alpha)}{\partial \iota_2^\alpha} &= (\iota_1 + T_1 - 1) \iota_2^\alpha \\
&\quad - \iota_1^\alpha \left(\iota_2 - \sum_{c=1}^{T_1-1} (\Psi(b_c^{\boldsymbol{\omega}}) - \Psi(a_c^{\boldsymbol{\omega}} + b_c^{\boldsymbol{\omega}})) \right).
\end{aligned}$$

Letting both two partial derivatives be zero, the update equations for ι_1^α and ι_2^α are obtained as follows.

$$\begin{aligned}
\iota_1^\alpha &= \iota_1 + T_1 - 1, \\
\iota_2^\alpha &= \iota_2 - \sum_{c=1}^{T_1-1} (\Psi(b_c^{\boldsymbol{\omega}}) - \Psi(a_c^{\boldsymbol{\omega}} + b_c^{\boldsymbol{\omega}})).
\end{aligned}$$

The update equations for β can be easily derived in the same way.

Supplementary Materials

The codes and data for the simulated example can be downloaded online.

Acknowledgments

The authors thank the editor, associate editor, and referees for reviewing the article and providing valuable comments.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2018R1C1B6004511).

References

- Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723. [314]
- Andrieu, C., De Freitas, N., Doucet, A., and Jordan, M. I. (2003), "An Introduction to MCMC for Machine Learning," *Machine Learning*, 50, 5–43. [317]
- Assunção, R., and Krainski, E. (2009), "Neighborhood Dependence in Bayesian Spatial Models," *Biometrical Journal*, 51, 851–869. [314]
- Au, T. S., Duan, R., Kim, H., and Ma, G.-Q. (2010), "Spatiotemporal Event Detection in Mobility Network," in *2010 IEEE International Conference on Data Mining*, IEEE, pp. 28–37. [321]
- Bach, F. R., and Jordan, M. I. (2002), "Kernel Independent Component Analysis," *Journal of Machine Learning Research*, 3, 1–48. [325]
- Balcan, M.-F., Liang, Y., and Gupta, P. (2014), "Robust Hierarchical Clustering," *The Journal of Machine Learning Research*, 15, 3831–3871. [321]
- Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [313]
- Bernard, J. A., Seidler, R. D., Hassevoort, K. M., Benson, B. L., Welsh, R. C., Wiggins, J. L., Jaeggi, S. M., Buschkuhl, M., Monk, C. S., Jonides, J., and Peltier, S. J. (2012), "Resting State Cortico-Cerebellar Functional Connectivity Networks: A Comparison of Anatomical and Self-Organizing Map Approaches," *Frontiers in Neuroanatomy*, 6, 31. [323]
- Berger, V. J., Gelfand, A. E., and Holland, D. M. (2012), "Space-Time Data Fusion Under Error in Computer Model Output: An Application to Modeling Air Quality," *Biometrics*, 68, 837–848. [313]
- Blei, D. M., and Jordan, M. I. (2006), "Variational Inference for Dirichlet Process Mixtures," *Bayesian Analysis*, 1, 121–143. [318]
- Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [318]
- Blekas, K., Nikou, C., Galatsanos, N., and Tsekos, N. V. (2007), "Curve Clustering With Spatial Constraints for Analysis of Spatiotemporal Data," in *19th IEEE International Conference on Tools With Artificial Intelligence*, 2007. ICTAI 2007 (Vol. 1), IEEE, pp. 529–535. [313]
- Bouveyron, C., and Brunet-Saumard, C. (2014), "Model-Based Clustering of High-Dimensional Data: A Review," *Computational Statistics & Data Analysis*, 71, 52–78. [314]
- Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2018), "ClustGeo: An R Package for Hierarchical Clustering With Spatial Constraints," *Computational Statistics*, 33, 1799–1822. [320]
- Chen, X., Jin, Y., Qiang, S., Hu, W., and Jiang, K. (2015), "Analyzing and Modeling Spatio-Temporal Dependence of Cellular Traffic at City Scale," in *2015 IEEE International Conference on Communications (ICC)*, IEEE, pp. 3585–3591. [321,322]
- Cheung, Y.-m., and Xu, L. (2001), "Independent Component Ordering in ICA Time Series Analysis," *Neurocomputing*, 41, 145–152. [323]
- Coppi, R., D'Urso, P., and Giordani, P. (2010), "A Fuzzy Clustering Model for Multivariate Spatial Time Series," *Journal of Classification*, 27, 54–88. [313]
- Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., and Beard, J. (2007), "Evaluating the Effect of Neighbourhood Weight Matrices on Smoothing Properties of Conditional Autoregressive (CAR) Models," *International Journal of Health Geographics*, 6, 54–58. [314]
- Ewens, W. J. (1990), "Population Genetics Theory—The Past and the Future," in *Mathematical and Statistical Developments of Evolutionary Theory*, Dordrecht: Springer, pp. 177–227. [315]
- Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209–230. [315]
- Finkenstadt, B., Held, L., and Isham, V. (2006), *Statistical Methods for Spatio-Temporal Systems*, Boca Raton, FL: Chapman and Hall/CRC. [316]
- Foti, N., Xu, J., Laird, D., and Fox, E. (2014), "Stochastic Variational Inference for Hidden Markov Models," in *Advances in Neural Information Processing Systems*, pp. 3599–3607. [317]
- Friston, K. J., Jezzard, P., and Turner, R. (1994), "Analysis of Functional MRI Time-Series," *Human Brain Mapping*, 1, 153–171. [323]
- Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian Non-parametric Spatial Modeling With Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021–1035. [325]

- Giraldo, R., Delicado, P., and Mateu, J. (2012), "Hierarchical Clustering of Spatially Correlated Functional Data," *Statistica Neerlandica*, 66, 403–421. [313]
- Görür, D., and Rasmussen, C. E. (2010), "Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution," *Journal of Computer Science and Technology*, 25, 653–664. [315]
- Haggarty, R. A., Miller, C. A., and Scott, E. M. (2015), "Spatially Weighted Functional Clustering of River Network Data," *Journal of the Royal Statistical Society, Series C*, 64, 491–506. [313]
- Huang, B., Wu, B., and Barry, M. (2010), "Geographically and Temporally Weighted Regression for Modeling Spatio-Temporal Variation in House Prices," *International Journal of Geographical Information Science*, 24, 383–401. [313]
- Jiang, H., and Serban, N. (2012), "Clustering Random Curves Under Spatial Interdependence With Application to Service Accessibility," *Technometrics*, 54, 108–119. [313,320,325]
- Johnson, S. C. (1967), "Hierarchical Clustering Schemes," *Psychometrika*, 32, 241–254. [320]
- Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [314,317,318]
- Kao, Y., Reich, B., Storlie, C., and Anderson, B. (2015), "Malware Detection Using Nonparametric Bayesian Clustering and Classification Techniques," *Technometrics*, 57, 535–546. [315]
- Kim, H., Duan, R., Kim, S., Lee, J., and Ma, G.-Q. (2019), "Spatial Cluster Detection in Mobility Networks: A Copula Approach," *Journal of the Royal Statistical Society, Series C*, 68, 99–120. [321]
- Kim, J., Lee, Y., and Kim, H. (2018), "Detection and Clustering of Mixed-Type Defect Patterns in Wafer Bin Maps," *IIEE Transactions*, 50, 99–111. [320]
- Lee, D., Zhou, S., Zhong, X., Niu, Z., Zhou, X., and Zhang, H. (2014), "Spatial Modeling of the Traffic Density in Cellular Networks," *IEEE Wireless Communications*, 21, 80–88. [321]
- 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. [313]
- Liao, W., Chen, H., Yang, Q., and Lei, X. (2008), "Analysis of fMRI Data Using Improved Self-Organizing Mapping and Spatio-Temporal Metric Hierarchical Clustering," *IEEE Transactions on Medical Imaging*, 27, 1472–1483. [313]
- MacKay, D. J. (1998), "Introduction to Gaussian Processes," *NATO ASI Series F Computer and Systems Sciences*, 168, 133–166. [316]
- (2003), *Information Theory, Inference and Learning Algorithms*, Cambridge: Cambridge University Press. [314]
- Paisley, J., Wang, C., and Blei, D. (2011), "The Discrete Infinite Logistic Normal Distribution for Mixed-Membership Modeling," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 74–82. [317]
- Park, S., Kim, Y.-D., and Choi, S. (2013), "Hierarchical Bayesian Matrix Factorization With Side Information," in *IJCAI*, pp. 1593–1599. [317]
- Paul, U., Subramanian, A. P., Buddhikot, M. M., and Das, S. R. (2011), "Understanding Traffic Dynamics in Cellular Data Networks," in *2011 Proceedings IEEE INFOCOM*, IEEE, pp. 882–890. [321]
- Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850. [320]
- Rees, G. (1999), "Single Subject Epoch (Block) Auditory fMRI Activation Data," available at <http://www.fil.ion.ucl.ac.uk/spm/data/auditory/>. [323]
- Rodríguez, A., and Dunson, D. B. (2011), "Nonparametric Bayesian Models Through Probit Stick-Breaking Processes," *Bayesian Analysis*, 6, 145–177. [325]
- Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007), "High-Resolution Space-Time Ozone Modeling for Assessing Trends," *Journal of the American Statistical Association*, 102, 1221–1234. [313]
- Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464. [314]
- Simmonds, D. J., Pekar, J. J., and Mostofsky, S. H. (2008), "Meta-Analysis of Go/No-Go Tasks Demonstrating That fMRI Activation Associated With Response Inhibition Is Task-Dependent," *Neuropsychologia*, 46, 224–232. [322]
- Strehl, A., and Ghosh, J. (2002), "Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal of Machine Learning Research*, 3, 583–617. [320]
- Titsias, M. K. (2009), "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *AISTATS* (Vol. 12), pp. 567–574. [325]
- Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001), "Temporal Autocorrelation in Univariate Linear Modeling of fMRI Data," *Neuroimage*, 14, 1370–1386. [323]
- Wu, X., Zurita-Milla, R., and Kraak, M.-J. (2015), "Co-Clustering Geo-Referenced Time Series: Exploring Spatio-Temporal Patterns in Dutch Temperature Data," *International Journal of Geographical Information Science*, 29, 624–642. [313]
- Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., and Vannucci, M. (2016), "A Spatiotemporal Nonparametric Bayesian Model of Multi-Subject fMRI Data," *The Annals of Applied Statistics*, 10, 638–666. [313,322]
- Zhang, L., Guindani, M., Versace, F., and Vannucci, M. (2014), "A Spatio-Temporal Nonparametric Bayesian Variable Selection Model of fMRI Data for Clustering Correlated Time Courses," *NeuroImage*, 95, 162–175. [313]
- Zhou, Z., Matteson, D. S., Woodard, D. B., Henderson, S. G., and Micheas, A. C. (2015), "A Spatio-Temporal Point Process Model for Ambulance Demand," *Journal of the American Statistical Association*, 110, 6–15. [313]