Bayesian Analysis (2021) **TBA**, Number TBA, pp. 1–30

Bayesian Non-Parametric Factor Analysis for Longitudinal Spatial Surfaces

Samuel I. Berchuck*, Mark Janko†, Felipe A. Medeiros‡, William Pan§, and Sayan Mukherjee¶

Abstract. We introduce a Bayesian non-parametric spatial factor analysis model with spatial dependency induced through a prior on factor loadings. For each column of the loadings matrix, spatial dependency is encoded using a probit stick-breaking process (PSBP) and a multiplicative gamma process shrinkage prior is used across columns to adaptively determine the number of latent factors. By encoding spatial information into the loadings matrix, meaningful factors are learned that respect the observed neighborhood dependencies, making them useful for assessing rates over space. Furthermore, the spatial PSBP prior can be used for clustering temporal trends, allowing users to identify regions within the spatial domain with similar temporal trajectories, an important task in many applied settings. In the manuscript, we illustrate the model's performance in simulated data, but also in two real-world examples: longitudinal monitoring of glaucoma and malaria surveillance across the Peruvian Amazon. The R package `spBFA`, available on `CRAN`, implements the method.

Keywords: Bayesian non-parametrics, probit stick-breaking process, factor analysis, dimension reduction, spatiotemporal clustering.

MSC2020 subject classifications: Primary 62G08, 62F15; secondary 62H25.

1 Introduction

The covariance for the standard Bayesian factor model, $\boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma}$, is a matrix decomposition, constructed to learn a latent representation for some potentially high-dimensional data object $\mathbf{Y}_t = \{Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_m)\}^\top$. We use notation from the spatial statistics literature to indicate the dimension of $\mathbf{Y}_t$, however this is only for consistency throughout the remainder of the paper. In fact, the data object $\mathbf{Y}_t$ is often not spatial in nature, but a vector that contains a large number of highly collinear variables. As such, throughout this paper, we refer to this dimension as the "variable dimension" of the data. The subscript t describes observed repetitions of the data object and can be inherently independent, spatial, or temporal in nature; we refer to this data dimension as the "replication dimension".

*Duke University, Department of Statistical Science and Forge, sib2@duke.edu
†University of Washington – Seattle, Institute of Health Metrics and Evaluation, mjanko@uw.edu
‡Duke University, Department of Ophthalmology, felipe.medeiros@duke.edu
§Duke University, Global Health Institute, william.pan@duke.edu
¶Duke University, Department of Statistical Science, Mathematics, Computer Science, and Bioinformatics & Biostatistics, sayan@stat.duke.edu

 https://doi.org/10.1214/20-BA1253

In this manuscript, we deal with the data setting where the vector $\mathbf{Y}_t$ represents a spatial surface and is observed longitudinally across time, t. Our ultimate goal is to obtain a low-dimensional representation of $\mathbf{Y}_t$, at each time t, that is learned from a process that accounts for the spatial structure of the observed data. By incorporating these spatial dependencies, the hope is that meaningful latent factors are learned that aid in understanding rates of change across the spatial surface and provide a framework for clustering spatial locations based on comparable temporal trajectories. To accomplish this, we generalize the standard factor analysis, to allow for non-linear relationships (Equation (2.1)) and introduce a novel spatial Bayesian non-parametric (BNP) prior on the columns of the factor loadings matrix, $\mathbf{\Lambda}$ (Equation (2.3)). We begin by reviewing existing factor analysis methods for spatial data.

Factor analysis is characterized by dimension reduction along the variable dimension of the observed data and is accomplished by projecting the data into a lower dimensional space, defined by a set of k factors, $\boldsymbol{\eta}_t = (\eta_{t1}, \ldots, \eta_{tk})^\top$. In practice the number of factors is small compared to the dimension of the data object ($k \ll m$). By definition, the factors have lower variability than the original data and are more manageable for inferential purposes due to their low dimension. In a standard Bayesian factor analysis, the latent vector $\boldsymbol{\eta}_t$ is often modeled as a standard Gaussian (Murray et al., 2013).

Typically, much of the innovation in factor analysis involves the prior for the $m \times k$ dimensional factor loadings matrix, $\mathbf{\Lambda}$. The naive approach assumes independent Gaussian priors for each element of $\mathbf{\Lambda}$, which has obvious computational issues when m and k are large. Furthermore, it may lead to poor inference due to the weakness of the prior specification and is not identifiable without further restrictions. In general, the specification of $\mathbf{\Psi}$ is not unique, as there are infinitely many possible factor loading matrices that satisfy the form. This can be seen by noting that any matrix of the form $\mathbf{\Lambda P}$ satisfies the condition, for any orthogonal matrix $\mathbf{P}$ (i.e., $\mathbf{P}\mathbf{P}^\top = \mathbf{I}_k$).

To remedy this, $\mathbf{\Lambda}$ is often a lower diagonal matrix with the loadings on the diagonal forced to be positive. This has been made computationally more efficient in recent years through parameter expansion of the loadings using basis elements (Ghosh and Dunson, 2009). Although these methods can be useful for identifiability, they remain computationally burdensome. Furthermore, it has been noted that from a Bayesian perspective, one does not require identifiability for many applications, including prediction, covariance estimation, and clustering (Bhattacharya and Dunson, 2011).

Adaptations of factor analysis to the spatial setting are plentiful and predominately focus on spatial dependence in the replication dimension of the data. A typical application of spatial factor analysis involves learning a latent representation of some high-dimensional data object that is observed across a geography, whether point-referenced or areal. Here, the foundational assumption of spatial statistics, that dependence between observations weakens as the distance between locations increases, is applied to the factors, so that a latent factor at a location should be similar to factors at nearby locations.

Christensen and Amemiya (2002) used this assumption to fit a shift-factor analysis method to model multivariate spatial data with temporal behavior modeled by autoregressive (AR) components. This method entertained several forms of spatial dependence

through a single factor, a standard construction in the literature (Hogan and Tchernis, 2004). There have been many extensions to multiple factors, most often using Gaussian likelihoods (Ren and Banerjee, 2013; Nethery et al., 2015), but also generalizing to Poisson (Tzala and Best, 2008) and binary (Wall and Liu, 2009). There are also extensions to informative missingness (Reich and Bandyopadhyay, 2010), and spatial mis-alignment (Nethery et al., 2018). Finally, Guhaniyogi et al. (2013) combined a low rank predictive process approach with the non-stationary linear model of coregionalization for computationally feasible modeling with large spatially-referenced datasets.

In all of these methods, the latent factors are responsible for encoding spatial dependency for the purpose of reducing the observed data at each location. In this paper, however, we will focus on an alternative form of spatial factor analysis that instead introduces spatial structure along the variable dimension of the data. Thus, instead of dimension reduction for some high-dimensional response across locations, the response is now univariate, and the dimension reduction is performed across spatial units. This approach is advantageous when the modeling goal is to identify spatial clusters whose temporal behavior is similar.

This approach was introduced by Lopes et al. (2008) through a spatial dynamic factor model. The key to this approach is a spatial prior on the columns of the factor loadings matrix, that allows for dimension reduction to be informed by spatial proximities. In Lopes et al. (2008), space was modeled using a distance-based Gaussian random field, while a more recent version used a Gaussian Markov random field for sparsity purposes (Strickland et al., 2011). This method has been extended to the generalized likelihood setting (Lopes et al., 2011). These methods use a lower diagonal specification for the loadings matrix for identifiability purposes and the number of factors is learned through reversible jump Markov chain Monte Carlo (MCMC). While these methods are useful for learning factors across a spatial surface, they rely on complicated identifiability constraints and lack clustering properties.

In this manuscript, we introduce a spatial factor analysis that collapses spatial locations into meaningful latent factors using a spatial BNP prior for the fully specified factor loadings matrix. In particular, we will model spatial dependency within the columns of the factor loadings using the probit stick-breaking process (PSBP), which is a scalable extension of standard spatial processes that allows for clustering. The BNP world has a rich literature involving spatial processes, mainly involving extensions of the Dirichlet Process (DP). The DP is the work-horse of BNPs and, when considering spatial dependencies, is best represented using a stick-breaking construction, such that $G \sim DP(\alpha, G_0)$ if and only if $G(\cdot) = \sum_{l=1}^{\infty} w_l \delta_{\theta_l}(\cdot)$, where $\theta_l \overset{\text{iid}}{\sim} G_0$ and $w_l = u_l \prod_{r=1}^{l-1}(1-u_r), l = 2, 3, \ldots$, with $u_r \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha)$ and δ_{θ_l} is a Dirac distribution with point mass at θ_l. Since the introduction of the dependent DP by MacEachern (1999), which modeled dependency through covariate information in the atoms (θ_l) and the weights (w_l), many methods have extended the DP to the spatial setting.

A popular spatial DP extension is Gelfand et al. (2005), which places a univariate stationary Gaussian process on the atoms to yield a random spatial process that is neither Gaussian nor stationary. The process has been extended to the generalized

framework (Duan et al., 2007). Modeling spatial dependency through the weights of the stick-breaking representation has also been popular, however until recent years has been computationally inefficient. In the more general stick-breaking construction, Rodriguez and Dunson (2011) introduced the PSBP, which replaces the characteristic Beta distribution prior with probit transformations of normal random variables. With the introduction of the PSBP, incorporating spatial dependency in BNP priors has become computationally straightforward, and mainstream (Chung and Dunson, 2009; Pati et al., 2013; Pati and Dunson, 2014).

Using the PSBP prior to introduce spatial structure into the factor loadings matrix yields a non-separable and non-stationary spatiotemporal (ST) process, with temporal dependence introduced through the factors. We show that the PSBP prior offers benefits for scalability and is useful for clustering spatial locations into regions across space with similar risk trajectories. A computationally efficient MCMC sampler is introduced that uses slice sampling to allow for an infinite mixture model. Furthermore, a multiplicative gamma process shrinkage prior is used to adaptively determine the number of latent factors, avoiding the computationally intensive reversible jump technique.

This paper is outlined as follows. In Section 2, we introduce a general factor analysis modeling framework and detail our novel spatial BNP prior for the columns of the factor loadings matrix. Through simulation in Section 3, we assess the utility of the novel prior in ST data and for clustering temporal trends across a spatial surface. Then, in Section 4, we apply the model to two real-world data applications: glaucoma disease progression and malaria risk surveillance. We conclude in Section 5 with a discussion.

2 Methodology

We begin by introducing a generalized modeling framework for factor analysis that allows for non-linearity and detail the temporal process for the latent factors. We then introduce the spatial PSBP prior for the factor loadings matrix and describe the multiplicative gamma process shrinkage prior for adaptively learning the appropriate number of latent factors. We conclude the section by working out a computationally efficient MCMC sampler for the infinite mixture model, describing the clustering properties of the introduced prior, and detailing prediction theory.

2.1 A General Modelling Framework

A generalized factor analysis model can be written as follows,

$$Y_t(\mathbf{s}_{i,o})|\vartheta_t(\mathbf{s}_{i,o}), \boldsymbol{\zeta}_t(\mathbf{s}_{i,o}) \overset{\text{ind}}{\sim} f\left(Y_t(\mathbf{s}_{i,o}); g^{-1}\left(\vartheta_t(\mathbf{s}_{i,o})\right), \boldsymbol{\zeta}_t(\mathbf{s}_{i,o})\right), \tag{2.1}$$

$$g\left(\vartheta_t(\mathbf{s}_{i,o})\right) = \mathbf{x}_t(\mathbf{s}_{i,o})\boldsymbol{\beta} + \sum_{j=1}^{k} \lambda_j(\mathbf{s}_{i,o})\eta_{tj}.$$

Here, we formally define our observed data as $Y_t(\mathbf{s}_{i,o})$ for temporal visit t, $(t = 1, \ldots, T)$ and spatial realization $\mathbf{s}_{i,o}$, for location i, $(i = 1, \ldots, m)$, and observation type o, $(o =$

$1, \ldots, O$). This is a general specification, so that at each time t, the spatial object can be multi-layered (i.e., color channels or multiple disease outcomes per location) with O layers. We define vectorized versions of the observed data as follows, $\mathbf{Y} = (\mathbf{Y}_1^\top, \ldots, \mathbf{Y}_T^\top)^\top$, where $\mathbf{Y}_t = (\mathbf{Y}_{t1}^\top, \ldots, \mathbf{Y}_{tO}^\top)^\top$ and $\mathbf{Y}_{to} = \{Y_t(\mathbf{s}_{1,o}), \ldots, Y_t(\mathbf{s}_{m,o})\}^\top$.

In our specification the factor loadings matrix is fully specified, with loadings $\lambda_j(\mathbf{s}_{i,o})$, corresponding to the stacking of the observed data. So the j^{th} column is given by $\boldsymbol{\lambda}_j = (\boldsymbol{\lambda}_{j1}^\top, \ldots, \boldsymbol{\lambda}_{jO}^\top)^\top$, with $\boldsymbol{\lambda}_{jo} = \{\lambda_j(\mathbf{s}_{1,o}), \ldots, \lambda_j(\mathbf{s}_{m,o})\}^\top$. A full specification allows for a direct application of spatial structure to $\boldsymbol{\lambda}_j$, $j = 1, \ldots, k$, as it has the same dimension as the underlying process, mO. While a full specification limits the interpretability of the factors themself, as mentioned before, one does not require identifiability for many applications, including covariance estimation and prediction.

While standard Bayesian factor analysis is performed using a Gaussian likelihood, (2.1) is a generalized form. The Gaussian specification can be recovered if we choose f to be Gaussian with mean, $\mu_t(\mathbf{s}_{i,o}) = g^{-1}(\vartheta_t(\mathbf{s}_{i,o}))$, nuisance or variance, $\boldsymbol{\zeta}_t(\mathbf{s}_{i,o}) = \sigma^2(\mathbf{s}_{i,o})$ and g the identity link. This is equivalent to the following vectorized model specification, $\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \boldsymbol{\Lambda}\boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t$, with $\boldsymbol{\epsilon}_t \sim \mathrm{N}_{mO}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{Diag}(\boldsymbol{\sigma}_1^2, \ldots, \boldsymbol{\sigma}_O^2)$, and $\boldsymbol{\sigma}_o^2 = (\sigma^2(\mathbf{s}_{1,o}), \ldots, \sigma^2(\mathbf{s}_{m,o}))^\top$. The component, $\mathbf{X}_t\boldsymbol{\beta}$, allows for covariates to adjust the factor analysis. The design matrix, $\mathbf{X}_t$, has rows, $\mathbf{x}_t(\mathbf{s}_{i,o})$, which is p dimensional.

The purpose of writing the model in this general form is that it is more flexible, allowing for various likelihoods. For example, when we study malaria in Section 4.2 we will use a Bernoulli distribution, by specifying f as Bernoulli, with probability $\pi_t(\mathbf{s}_{i,o}) = g^{-1}(\vartheta_t(\mathbf{s}_{i,o}))$, and the logit link ($\boldsymbol{\zeta}_t(\mathbf{s}_{i,o})$ is null). Full details for deriving the Bernoulli likelihood in this context can be found in Section 5 of the Supplementary Materials online (Berchuck et al., 2020).

We conclude this section by specifying a temporal structure for the latent factors. Again, we specify a general framework, $\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \mathbf{H}(\psi) \otimes \boldsymbol{\Upsilon})$, where $\boldsymbol{\eta} = \{\boldsymbol{\eta}_1^\top, \ldots, \boldsymbol{\eta}_T^\top\}^\top$. This form is flexible, allowing for many common time series models, including the AR(1) and exponential processes To obtain the AR(1), choose $\mathbf{H}(\psi)$, such that $[\mathbf{H}(\psi)]_{tt'} = \psi^{|x_t - x_{t'}|}$, which results in $\boldsymbol{\eta}_t = \psi\boldsymbol{\eta}_{t-1} + \boldsymbol{\upsilon}_t, \boldsymbol{\upsilon}_t \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Upsilon})$, if time is uniform. The exponential correlation structure can be obtained with $[\mathbf{H}(\psi)]_{tt'} = \exp\{\psi|x_t - x_{t'}|\}$, where x_t is follow-up time t.

2.2 Spatial Bayesian Non-Parametric Factor Loadings

The scalar form of the likelihood in (2.1) motivates spatial dependency in the factor loadings. In particular, due to the fully specified factor loadings matrix, the following linear relationship between the transformed mean process and the latent factors exists,

$$g\left(\vartheta_t(\mathbf{s}_{i,o})\right) = \mathbf{x}_t(\mathbf{s}_{i,o})\boldsymbol{\beta} + \sum_{j=1}^{k} \lambda_j(\mathbf{s}_{i,o})\eta_{tj} = \mathbf{x}_t(\mathbf{s}_{i,o})\boldsymbol{\beta} + \lambda_1(\mathbf{s}_{i,o})\eta_{t1} + \cdots + \lambda_k(\mathbf{s}_{i,o})\eta_{tk}. \quad (2.2)$$

This illuminates that a factor loading $\lambda_j(\mathbf{s}_{i,o})$ represents the amount that observation $Y_t(\mathbf{s}_{i,o})$ is explained through the latent factor j at time t, η_{tj}. Therefore, for two obser-

vations, $Y_t(\mathbf{s}_{i,o})$ and $Y_t(\mathbf{s}_{i',o})$, that are spatially correlated, we would assume that their relationships to the latent factor, η_{tj} would be similar, $\lambda_j(\mathbf{s}_{i,o}) \approx \lambda_j(\mathbf{s}_{i',o})$.

To induce the desired spatial dependency, as motivated by (2.2), into the columns of the factor loadings matrix, we use a PSBP for each column,

$$\begin{aligned} \lambda_j(\mathbf{s}_{i,o})|G_j^{i,o} &\overset{\text{ind}}{\sim} G_j^{i,o}, \quad i = 1,\ldots,m, \quad o = 1,\ldots,O, \quad j = 1,\ldots,k, \\ G_j^{i,o}(\cdot) &= \sum_{l=1}^{L} w_{jl}(\mathbf{s}_{i,o})\delta_{\theta_{jl}}(\cdot), \\ w_{jl}(\mathbf{s}_{i,o}) &= \Phi(\alpha_{jl}(\mathbf{s}_{i,o}))\prod_{r<l}[1-\Phi(\alpha_{jr}(\mathbf{s}_{i,o}))], \end{aligned} \tag{2.3}$$

where $\{\alpha_{jl}(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}_{l=1}^{L-1}$ for $j = 1,\ldots,k$ has Gaussian marginals, with $\mathcal{D}$ some multivariate spatial surface, and $\{\theta_{jl}\}_{l=1}^{L}$ for $j = 1,\ldots,k$ are independent and identically distributed for each j. The form of (2.3) closely mirrors the stick-breaking construction of the DP, however the weights are now constructed using the standard Gaussian cumulative distribution function, Φ. As is shown in Rodriguez and Dunson (2011), this is a proper construction, because for finite L, it ensures that $\sum_{l=1}^{L} w_{jl}(\mathbf{s}_{i,o}) = 1$, for all j. When $L = \infty$, it is easy to verify that $\sum_{l=1}^{\infty} \mathbb{E}[\log\{1-\Phi(\alpha_{jl}(\mathbf{s}_i))\}] = -\infty$ and therefore $\sum_{l=1}^{\infty} w_{jl}(\mathbf{s}_{i,o}) = 1$, for all j. This property clearly transfers to our new prior across the columns of the factor loadings. A more detailed discussion of this is given in Section 6.4 of the Supplementary Materials online.

The parameters that dictate the weights, $\alpha_{jl}(\mathbf{s}_{i,o})$, have a joint distribution that induces spatial dependency. Define the joint parameter, $\boldsymbol{\alpha}_{jlo} = \{\alpha_{jl}(\mathbf{s}_{1,o}),\ldots,\alpha_{jl}(\mathbf{s}_{m,o})\}^\top$ and $\boldsymbol{\alpha}_{jl} = \{\boldsymbol{\alpha}_{jl1}^\top,\ldots,\boldsymbol{\alpha}_{jlO}^\top\}^\top$. We specify a simple, but flexible form using a separable specification, $\boldsymbol{\alpha}_{jl} \sim \text{N}_{Om}(\mathbf{0}, \boldsymbol{\kappa} \otimes \mathbf{F}(\rho))$. This separable Gaussian specification allows for conjugacy of $\boldsymbol{\alpha}_{jl}$ and $\boldsymbol{\kappa}$, thus maintaining computational feasibility. Notice, that while we treat space using a separable process, the resulting marginal process will be non-separable. The $m \times m$ matrix $\mathbf{F}(\rho)$ dictates the spatial neighborhood structure, for example a Gaussian process with exponential correlation, $\mathbf{F}(\rho) = \exp\{-\rho\mathbf{D}\}$, for a continuous spatial domain, or a Gaussian Markov random field for discrete spatial data, $\mathbf{F}(\rho)^{-1} = \mathbf{D}_w - \rho\mathbf{W}$; we assume a proper conditional autoregressive (CAR) prior. Here $\mathbf{D}$ is a distance matrix (typically Euclidean) and $\mathbf{W}$ is an adjacency matrix, with adjacencies $\{w_{ii'}\}$ that indicate the level of spatial correlation between locations i and i' and do not change over time ($\mathbf{D}_w$ is a diagonal matrix that weights the number of neighbors of each locations i). The parameter ρ indicates the level of spatial correlation.

Finally, in prior attempts to model the factor loadings matrix the number of latent factors (i.e., number of columns of $\boldsymbol{\Lambda}$) was determined using the reversible jump MCMC. This decision requires a preliminary run for each choice of the number of factors and is very computationally intensive. As such, we decide to model the atoms, θ_{jl}, using a multiplicative gamma process shrinkage prior (Bhattacharya and Dunson, 2011). This prior conveniently shrinks the magnitude of possible entries, where the degree of shrinkage increases with the column index. In particular, $\theta_{jl} \overset{\text{ind}}{\sim} \text{N}(0, \tau_j^{-1})$, where the precision is forced to increase over the column index, $\tau_j = \prod_{h=1}^{j} \delta_h$, with $\delta_1 \sim \text{Ga}(a_1, 1)$, and

$\delta_h \sim \text{Ga}(a_2, 1)$, for $h \geq 2$. This allows us to specify a value of k that is larger than the number of supposed factors, with the prior reducing the factors to a set of meaningful ones. Holding a_2 constant, increasing a_1 yields smaller column variances, and holding a_1 constant, increasing a_2 yields faster shrinkage of the column variances as j increases.

2.3 Computational Considerations

In order to facilitate Bayesian inference, the likelihood can be written in terms of the underlying atoms, a standard practice in mixture models,

$$g\left(\vartheta_t(\mathbf{s}_{i,o})\right) = x_t(\mathbf{s}_{i,o})\boldsymbol{\beta} + \sum_{j=1}^{k} \theta_{j\xi_j(\mathbf{s}_{i,o})}\eta_{tj}. \tag{2.4}$$

This is a simple replacement of the factor loadings, $\lambda_j(\mathbf{s}_{i,o})$, with their corresponding atom, θ_{jl}, which is determined by a clustering indicator $\xi_j(\mathbf{s}_{i,o}) = l$. The representation in (2.4) reminds us of the discrete nature of the PSBP as the categorical parameter, $\xi_j(\mathbf{s}_{i,o})$, indicates the cluster of $\lambda_j(\mathbf{s}_{i,o})$, and has the following distribution, $\text{Multinomial}(w_{j1}(\mathbf{s}_{i,o}), \ldots, w_{jL}(\mathbf{s}_{i,o}))$, so that $P(\xi_j(\mathbf{s}_{i,o}) = l) = w_{jl}(\mathbf{s}_{i,o})$. This construction helps to illuminate the importance of the spatial dependency introduced in the PSBP, as the value of $\xi_j(\mathbf{s}_{i,o})$ (i.e., cluster of $\lambda_j(\mathbf{s}_{i,o})$) is sampled from a multinomial distribution with weights that have been spatially smoothed to be similar to nearby locations. This is desirable, because it encourages close locations to belong to the same cluster, and thus constructs underlying factors that relate to regions of the spatial domain.

To facilitate efficient computations, we introduce a latent variable, $z_{jl}(\mathbf{s}_{i,o}) \overset{\text{ind}}{\sim} \text{N}(\alpha_{jl}(\mathbf{s}_{i,o}), 1)$, to facilitate conjugacy in the $\alpha_{jl}(\mathbf{s}_{i,o})$. Conjugacy follows from the following equivalency,

$$\begin{aligned} P(\xi_j(\mathbf{s}_{i,o}) = l | z_{jl}(\mathbf{s}_{i,o})) &= P(z_{jl}(\mathbf{s}_{i,o}) > 0, z_{jr}(\mathbf{s}_{i,o}) < 0, \forall r < l) \\ &= \Phi(\alpha_{jl}(\mathbf{s}_{i,o})) \prod_{r<l} [1 - \Phi(\alpha_{jr}(\mathbf{s}_{i,o}))] \\ &= w_{jl}(\mathbf{s}_{i,o}). \end{aligned}$$

This permits conjugacy in the $\alpha_{jl}(\mathbf{s}_{i,o})$ by noting the following conditional independence, $\xi_j(\mathbf{s}_{i,o}) \perp\!\!\!\perp \alpha_{jl}(\mathbf{s}_{i,o}) | z_{jl}(\mathbf{s}_{i,o})$. Furthermore, the data augmentation parameters, $z_{jl}(\mathbf{s}_{i,o})$, have conjugate form.

The theory described above was for finite L, however it can be easily extended to an infinite mixture model using a slice sampling technique (Walker, 2007). Slice sampling makes the infinite mixture model computationally feasible by introducing an upper bound for the number of clusters, thus reducing the process to a finite mixture model. In particular, all parameters that depend on the number of clusters (θ_{jl}, $\xi_j(\mathbf{s}_{i,o})$, $z_{jl}(\mathbf{s}_{i,o})$, $\alpha_{jl}(\mathbf{s}_{i,o})$, $\boldsymbol{\kappa}$, ρ, δ_h) will be augmented using the slice sampling truncation. The idea is to introduce a latent variable, $u_j(\mathbf{s}_{i,o})$, with uniform density so that conditional on $u_j(\mathbf{s}_{i,o}), j = 1, \ldots, k$, $o = 1, \ldots, O$ and $i = 1, \ldots, m$, the conditional mixture distribution

becomes finite. When dealing with full conditionals, this truncation corresponds to reducing the number of mixture components for each column of the factor loadings matrix to L_j^*, so that $l_j = 1, \ldots, L_j^* = \max\{L_j^{i,o}; i = 1, \ldots, m, o = 1, \ldots, O\}$, where $L_j^{i,o}$ is the minimum integer satisfying $\sum_{l_j=1}^{L_j^{i,o}} w_{jl_j}(\mathbf{s}_{i,o}) > 1 - u_j^* = \min\{u_j(\mathbf{s}_{i,o})\}$, for $i = 1, \ldots, m$ and $o = 1, \ldots, O$. Throughout the simulations and data illustrations in Sections 3 and 4, we use the infinite mixture model. Full details pertaining to the implementation of the model, including full conditionals, and theory related to slice sampling are contained in Sections 3 and 4 of the Supplementary Materials online.

2.4 Model Properties

We focus on the class of spatial PSBP models, where each column of the factor loadings matrix has the following form, $\mathcal{M}_j = \{G_j^{i,o} : \mathbf{s}_{i,o} \in \mathcal{D}\}$, where each column progressively shrinks due to the gamma process shrinkage prior on the atoms. For each column, the process $G_j^{i,o}$ marginally follows a PSBP for each $\mathbf{s} \in \mathcal{D}$. Therefore, for any set $B \in \mathcal{B}$, we can obtain the moments of the process. In this section, we describe the moments of the PSBP process, originally derived in Rodriguez and Dunson (2011), plus new results that describe the conditional and marginal moments of the introduced spatial factor analysis.

The process moments are as follows, beginning with the mean, $\mathbb{E}[G_j^{i,o}(B)] = G_{0j}(B)$. The variance, $\mathbb{V}(G_j^{i,o}(B))$, and covariance, $\mathbb{C}(G_j^{i,o}(B), G_j^{i',o'}(B))$, are as follows,

$$G_{0j}(B)\{1 - G_{0j}(B)\}\left[\beta_2(\mathbf{s}_{i,o})\left(\frac{1 - \{1 - 2\beta_1(\mathbf{s}_{i,o}) + \beta_2(\mathbf{s}_{i,o})\}^L}{2\beta_1(\mathbf{s}_{i,o}) - \beta_2(\mathbf{s}_{i,o})}\right)\right], \tag{2.5}$$

$$G_{0j}(B)\{1 - G_{0j}(B)\}\beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})\left[\frac{1 - \{1 - \beta_1(\mathbf{s}_{i,o}) - \beta_1(\mathbf{s}_{i',o'}) + \beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})\}^L}{\beta_1(\mathbf{s}_{i,o}) + \beta_1(\mathbf{s}_{i',o'}) - \beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})}\right].$$

Finally, it is easy to see that the covariance across columns is zero, meaning the shrinkage process does not introduce dependency at the PSBP level, $\mathbb{C}[G_j^{i,o}(B), G_{j'}^{i,o}(B)] = G_{0j}(B)G_{0j'}(B)\{1 - (1 - \beta_1(\mathbf{s}_{i,o}))^L\}^2 - G_{0j}(B)G_{0j'}(B) \overset{L\to\infty}{=} 0$. Here, we use the specification that, $u_{jl}(\mathbf{s}_{i,o}) = \Phi(\alpha_{jl}(\mathbf{s}_{i,o}))$, $\beta_p(\mathbf{s}_{i,o}) = \mathbb{E}[u_{jl}(\mathbf{s}_{i,o})^p]$ and the base distribution for atom, θ_{jl}, is given by G_{0j}. Higher moments are also defined as such, $\beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'}) = \mathbb{E}[u_{jl}(\mathbf{s}_{i,o})u_{jl}(\mathbf{s}_{i',o'})]$.

As described in Rodriguez and Dunson (2011), these stick-breaking expectations, which show up in the model properties, have closed forms as long as the underlying spatial process has a marginal distribution, $\alpha_{jl}(\mathbf{s}_{i,o}) \sim \mathrm{N}(\mu, \sigma^2)$. Then, using a change of variables, $t_1 = \alpha_{jl}(\mathbf{s}_{i,o}) - x$ and $t_2 = \alpha_{jl}(\mathbf{s}_{i,o})$, we see that $\beta_1(\mathbf{s}_{i,o}) = P(T_1 > 0)$, where $T_1 \sim \mathrm{N}(\mu, 1 + \sigma^2)$. Generally, the p-th moment is given by, $\beta_p(\mathbf{s}_{i,o}) = \mathbb{E}[u_{jl}^p(\mathbf{s}_{i,o})] = P(T_1 > 0, \ldots, T_p > 0)$, where $(T_1, \ldots, T_p)^\top$ is multivariate normal with $\mathbb{E}[T_i] = \mu$, $\mathbb{V}(T_i) = 1 + \sigma^2$, and $\mathbb{C}(T_i, T_j) = \sigma^2$. Finally, $\beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'}) = P(T_1 > 0, T_2 > 0)$, where $(T_1, T_2)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \mathbf{I})$, where the moments $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ come from the marginal distribution, $(\alpha_{jl}(\mathbf{s}_{i,o}), \alpha_{jl}(\mathbf{s}_{i',o'}))^\top \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The marginal moments indicate the importance of the base distribution, G_{0j}, as a centering measure, with the marginal moments of the underlying Gaussian parameter, $\alpha_{jl}(\mathbf{s}_{i,o})$ controlling the variance and covariance of the sampled distributions around G_{0j}. Furthermore, it was shown that as $\mathbf{s}_{i,o} \to \mathbf{s}_{i',o}$, that $\mathbb{C}(G_j^{i,o}(B), G_j^{i',o}(B)) \to \mathbb{V}(G_j^{i,o}(B))$, which can be explained by the fact that $\beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o}) \to \beta_1(\mathbf{s}_{i,o})\beta_1(\mathbf{s}_{i',o})$.

Having established the marginal moments of the process, we now turn our attention to deriving the moments of the observed data model. In order to derive these moments, we need to understand the induced spatial factor analysis model. The induced conditional likelihood for our model, $f\left(Y_t(\mathbf{s}_{i,o})|G_j^{i,o}, \boldsymbol{\eta}_t, \boldsymbol{\zeta}_t(\mathbf{s}_{i,o})\right)$, can be written in the following two forms,

$$\int \cdots \int f\left(Y_t(\mathbf{s}_{i,o}); g^{-1}\left(\vartheta_t(\mathbf{s}_{i,o})\right), \boldsymbol{\zeta}_t(\mathbf{s}_{i,o})\right) G_1^{i,o}(d\lambda_1(\mathbf{s}_{i,o})) \cdots G_k^{i,o}(d\lambda_k(\mathbf{s}_{i,o})),$$
$$\sum_{l_1=1}^{L} \cdots \sum_{l_k=1}^{L} w_{1l_1}(\mathbf{s}_{i,o}) \cdots w_{kl_k}(\mathbf{s}_{i,o}) f\left(Y_t(\mathbf{s}_{i,o}); g^{-1}\left(\vartheta_t(\mathbf{s}_{i,o})\right), \boldsymbol{\zeta}_t(\mathbf{s}_{i,o})\right).$$

These two equivalent forms of the induced model demonstrate the mixing, which averages over the factor loadings according to the PSBP (Equation (2.3)). These representations are critical for determining the marginal and conditional moments of the observed data model. In particular, we derive moments assuming a Gaussian likelihood, as the derivations become untenable without the identity link assumption. Therefore, $f(Y_t(\mathbf{s}_{i,o})|\boldsymbol{\beta}, \lambda_j(\mathbf{s}_{i,o}), \boldsymbol{\eta}_t, \sigma^2(\mathbf{s}_{i,o})) = \mathrm{N}(x_t(\mathbf{s}_{i,o})\boldsymbol{\beta} + \sum_{j=1}^{k} \lambda_j(\mathbf{s}_{i,o})\eta_{tj}, \sigma^2(\mathbf{s}_{i,o}))$. The conditional mean and variance are given as, $\mathbb{E}[Y_t(\mathbf{s}_{i,o})|G_i^{i,o}, \boldsymbol{\eta}_t, \sigma^2(\mathbf{s}_{i,o})] = \mathbf{x}_t(\mathbf{s}_{i,o})\boldsymbol{\beta} + \sum_{j=1}^{k}(\sum_{l_j=1}^{L} w_{jl_j}(\mathbf{s}_{i,o})\theta_{jl_j})\eta_{tj}$, and then $\mathbb{V}(Y_t(\mathbf{s}_{i,o})|G_j^{i,o}, \boldsymbol{\eta}_t, \sigma^2(\mathbf{s}_{i,o})) = \sigma^2(\mathbf{s}_{i,o})$. The mean process is elegant, as it takes the form of the original mean process, but replaces the loadings with a mixture over the underlying atoms, weighted according to the PSBP. The spatial covariance, $\mathbb{C}(Y_t(\mathbf{s}_{i,o}), Y_t(\mathbf{s}_{i',o'})|G_j^{i,o}, G_j^{i',o'}, \boldsymbol{\eta}_t, \sigma^2(\mathbf{s}_{i,o}), \sigma^2(\mathbf{s}_{i',o'}))$, is conditionally zero. The form of the conditional moments are reminiscent of a standard spatial analysis that uses a spatially varying intercept with a Gaussian process, where conditional on the spatial intercepts, the process is independent across space.

Finally, we turn our attention to the marginal moments of the process. The mean process is as follows, $\mathbb{E}[Y_t(\mathbf{s}_{i,o})] = 0$, while the variance, $\mathbb{V}(Y_t(\mathbf{s}_{i,o}))$, and covariance, $\mathbb{C}(Y_t(\mathbf{s}_{i,o}), Y_t(\mathbf{s}_{i',o'}))$, are given respectively as,

$$\mathbb{E}\left[\sigma^2(\mathbf{s}_{i,o})\right] + \left[\frac{\beta_2(\mathbf{s}_{i,o})\left(1 - \{1 - 2\beta_1(\mathbf{s}_{i,o}) + \beta_2(\mathbf{s}_{i,o})\}^L\right)}{2\beta_1(\mathbf{s}_{i,o}) - \beta_2(\mathbf{s}_{i,o})}\right]\left(\sum_{j=1}^{k} \tau_j^{-1}\mathbb{E}\left[\eta_{tj}^2\right]\right), \text{ and} \tag{2.6}$$
$$\left[\frac{\beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})\left(1 - \{1 - \beta_1(\mathbf{s}_{i,o}) - \beta_1(\mathbf{s}_{i',o'}) + \beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})\}^L\right)}{\beta_1(\mathbf{s}_{i,o}) + \beta_1(\mathbf{s}_{i',o'}) - \beta_2(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})}\right]\left(\sum_{j=1}^{k} \tau_j^{-1}\mathbb{E}\left[\eta_{tj}^2\right]\right).$$

Both the variance and covariance take the same form as the moments from the PSBP process in (2.5), however they are now being scaled by a summation, that is a function

of the atom variances for each column and the second moment of the latent factors. In particular, as we see increases in both the number of factors (k) and the variability in the underlying atoms, the variance and covariance become inflated. For a full interpretation of the marginal moments, however, we need to place priors on the hyperparameters. For a detailed derivation of these model properties, see Section 6 of the Supplementary Materials online.

2.5 Prior Specification

We finalize the model specification by introducing priors for the remaining parameters; spatial parameters, $\boldsymbol{\kappa}$ and ρ, temporal parameters, $\boldsymbol{\Upsilon}$ and ψ, along with any nuisance parameters, for example the variance in the Gaussian likelihood, $\sigma^2(\mathbf{s}_{i,o})$. For each of these parameters, we choose standard priors to promote conjugacy in the full conditionals.

The spatial covariance, $\boldsymbol{\kappa}$, is an $O \times O$ matrix over the multiple levels of the image and has the conjugate inverse-Wishart (IW) prior, $\boldsymbol{\kappa} \sim \text{IW}(\upsilon, \boldsymbol{\Theta})$. When $O = 1$ this prior reduces to an inverse-Gamma (IG) distribution, $\text{IG}(\upsilon/2, \boldsymbol{\Theta}/2)$. For degrees of freedom, we specify $\upsilon = O+1$ and the scale matrix we use $\boldsymbol{\Theta} = \mathbf{I}_O$. This prior is appealing since it induces marginally uniform priors on the correlations of $\boldsymbol{\kappa}$ and allows for the diagonals to be weakly informative (Gelman et al., 2013). For the spatial tuning parameter, the prior is dependent on the type of spatial data, areal or point-referenced. For areal data, the prior for ρ is typically fixed at 0.99 to promote spatial smoothing, or given a uniform prior between zero and one. For point-referenced data, the prior is often uniform with bounds informed based on the expected range of the spatial variability, as in Berchuck et al. (2016).

The hyperparameters for the temporal process are assigned priors in the same vein as the spatial parameters. For the $k \times k$ covariance of the latent factors, $\boldsymbol{\Upsilon} \sim \text{IW}(\zeta, \boldsymbol{\Omega})$, with $\zeta = k + 1$ and $\boldsymbol{\Omega} = \mathbf{I}_k$. The prior for the temporal tuning parameter ψ depends on the temporal correlation structure. For an AR(1) process the temporal tuning parameter ψ has a transformed Beta distribution, $\psi \propto (1+\psi)^{\gamma-1}(1-\psi)^{\beta-1}$. While in the case of an exponential process, a uniform prior is more appropriate. Finally, in the case of a Gaussian likelihood, a weakly informative prior is used for the variances, $\sigma^2(\mathbf{s}_{i,o}) \sim \text{IG}(a, b)$.

2.6 Clustering Temporal Trends

An important characteristic of our introduced methodology is its ability to cluster regions across a spatial surface dependent on rates of temporal change. In BNP, clustering is determined based on the posterior probability that two locations belong to the same underlying cluster, $g_j(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'}) = P(\xi_j(\mathbf{s}_{i,o}) = \xi_j(\mathbf{s}_{i',o'}))$, which alleviates the label switching issue. Unfortunately, due to the full specification of the factor loadings matrix, none of the columns are themselves identifiable, and we can not individually cluster on the k columns. Instead, we focus our clustering efforts on the factor loadings from all of the columns.

We begin by defining the loading probabilities for factor j at a particular location: $\mathbf{w}_j(\mathbf{s}_{i,o}) = \{w_{11}(\mathbf{s}_{i,o}), \ldots, w_{1L}(\mathbf{s}_{i,o})\}$. The full set across all the latent factors is given

by, $\mathbf{w}(\mathbf{s}_{i,o}) = \{\mathbf{w}_1(\mathbf{s}_{i,o}), \ldots, \mathbf{w}_{k^*}(\mathbf{s}_{i,o})\}$. Note that we will limit to the first k^* factors, which is designed to only include factors that exhibit variability across locations. To find a good value of k^* we propose the following criteria, $\min_{\{i,o,i',o'\}}\{g_j(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})\} < \pi_k$ and $\max_{\{i,o,i',o'\}}\{g_j(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})\} > (1 - \pi_k)$, for $j = 1, \ldots, k^*$. This criteria attempts to exclude any factors that are non-informative for clustering, by only keeping factors with $g_j(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})$ near zero or one. The tuning parameter, π_k, is continuous and ranges between 0 and 0.5. A value of $\pi_k = 0$ will remove all the factors, while $\pi_k = 0.5$ will not remove any factors, resulting in the original k factors. The final object, $\mathbf{w} = \{\mathbf{w}(\mathbf{s}_{1,1})^\top, \ldots, \mathbf{w}(\mathbf{s}_{m,1})^\top, \ldots, \mathbf{w}(\mathbf{s}_{1,O})^\top, \ldots, \mathbf{w}(\mathbf{s}_{m,O})^\top\}^\top$ has dimension $mO \times Lk^*$. More specifically, however, when using slice sampling each factor is truncated to L_j^*, and thus in theory $\mathbf{w}$ has a much smaller number of columns.

When clustering temporal trends across the spatial surface, we will use the factor loading probability matrix, $\mathbf{w}$. In particular, we apply simple k-means to $\mathbf{w}$ and use the gap-statistic to determine the proper number clusters (Tibshirani et al., 2001). While this process may not seem immediately intuitive, since $\mathbf{w}$ is potentially larger than the original data, we found that clustering $\mathbf{w}$ produced improved results over the raw data $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_T\}$.

2.7 Bayesian Non-Parametric Prediction

Once posterior samples have been obtained, prediction is often a priority. In particular, obtaining samples from the posterior predictive distribution (PPD) is of interest, for both new spatial and temporal instances. We begin by detailing how future instances of the spatial surface can be obtained, by defining the PPD as $f(\mathbf{Y}_{T+1}|\mathbf{Y})$. We express the PPD as an integral $\int_\Omega f(\mathbf{Y}_{T+1}|\Omega, \mathbf{Y}) f(\Omega|\mathbf{Y}) d\Omega$ and then further partition the integral,

$$\int_\Omega \underbrace{f\left(\mathbf{Y}_{T+1}|g^{-1}\left(\boldsymbol{\vartheta}_{T+1}\right), \boldsymbol{\zeta}_{T+1}\right)}_{1} \underbrace{f\left(\boldsymbol{\eta}_{T+1}|\boldsymbol{\eta}, \boldsymbol{\Upsilon}, \psi\right)}_{2} \underbrace{f\left(\boldsymbol{\Lambda}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\Upsilon}, \psi|\mathbf{Y}\right)}_{3} d\Omega, \tag{2.7}$$

where $\Omega = (\boldsymbol{\vartheta}_{T+1}, \boldsymbol{\zeta}_{T+1}, \boldsymbol{\Lambda}, \boldsymbol{\eta}, \boldsymbol{\zeta}, \boldsymbol{\Upsilon}, \psi)$. The convenient form of (2.7) is a function of three known densities that are defined as a consequence of the methodology introduced in Section 2.2. As such, the PPD can be obtained by composition sampling.

Density (2.7).1 represents the observed likelihood function written in vector form and is problem specific (in scalar form: $\prod_{o=1}^{O}\prod_{i=1}^{m} f(Y_{T+1}(\mathbf{s}_{i,o})|g^{-1}(\vartheta_{T+1}(\mathbf{s}_{i,o})), \zeta_{T+1}(\mathbf{s}_{i,o})))$. Density (2.7).2 depends on properties of the conditional multivariate normal (MVN) density, yielding $f(\boldsymbol{\eta}_{T+1}|\boldsymbol{\eta}, \boldsymbol{\Upsilon}, \psi) \sim \text{MVN}(\mathbb{E}_{\boldsymbol{\eta}_{T+1}}, \mathbb{C}_{\boldsymbol{\eta}_{T+1}})$. The moments are $\mathbb{E}_{\boldsymbol{\eta}_{T+1}} = (\mathbf{H}^+ \otimes \mathbf{I})\boldsymbol{\eta}$ and $\mathbb{C}_{\boldsymbol{\eta}_{T+1}} = \mathbf{H}^* \otimes \boldsymbol{\Upsilon}$ with $\mathbf{H}^+ = [\mathbf{H}(\psi)]_{T+1,1:T}[\mathbf{H}(\psi)]_{1:T,1:T}^{-1}$ and finally $\mathbf{H}^* = [\mathbf{H}(\psi)]_{T+1,T+1} - \mathbf{H}^+[\mathbf{H}(\psi)]_{1:T,T+1}$. Here $\mathbf{H}(\psi)$ represents the temporal correlation matrix including the new time point $T + 1$, so that $[\mathbf{H}(\psi)]_{T+1,1:T}$ is a subset including the row $T + 1$ and columns 1 up to T. Finally, density (2.7).3 is the posterior distribution obtained in the MCMC sampler from the original model fit. Full details of the prediction theory and an extension to predicting at new spatial locations are given in Section 7 of the Supplementary Materials online.

3 Simulation Experiments

3.1 Justifying the Spatial PSBP for Factor Analysis

In our first simulation experiment, we aimed to illuminate the importance of the spatial PSBP and multiplicative gamma process shrinkage priors in the presence of spatial variability. We simulated data using the full model across various settings, including a different number of latent factors, $k = 1, 3, 6$, and the presence or absence of spatial correlation. To simulate data from a spatial process the spatial covariance, $\mathbf{F}(\rho)$, was set to a proper CAR prior, with $\rho = 0.99$. To simulate data with no spatial dependence, the spatial covariance was fixed at the identity matrix, (i.e., $\mathbf{F}(\rho) = \mathbf{I}$). Finally, for an additional simulation setting, we simulated data from a model that removed the PSBP mechanism, assigning Gaussian processes directly to the columns of the factor loadings matrix. The purpose of this final setting was to assess our proposed model under the data generating mechanism of existing methods that rely on Gaussian process models, for example, Lopes et al. (2008). This yielded twelve simulation settings.

The simulation aimed to imitate the monitoring of glaucoma progression, so we fixed the number of spatial locations, $(m = 52, O = 1)$, to the number on a visual field (details of this setting follow in Section 4.1). The visits are uniformly spaced between zero and one with total number of visits set to the average in our visual field data ($T = 10$). Furthermore, we used the adjacency matrix from the visual field, where two locations i and i' are considered neighbors if they share an edge or corner, $w_{ii'} = 1(i \sim i')$. Finally, we set $\boldsymbol{\kappa} = 1$, $\tau_j = 1$, $\psi = 0.3$, and $\boldsymbol{\Upsilon}$ was sampled from its prior distribution and was dependent on k. For each setting, 100 datasets were simulated, where every dataset is generated from one simulated instance of $\boldsymbol{\alpha}$ to ensure that the results are not affected by a particular realization. In the simulation settings that removed the PSBP, the factor loadings matrix was simulated directly from $\boldsymbol{\alpha}$.

We now describe specific details of the model implementation, which, unless otherwise noted, apply to all subsequent modeling examples. For the spatial process, we used a proper CAR with $\rho = 0.99$ to encourage spatial dependency, similar to how the data was simulated. An exponential correlation structure was used for the temporal process, so that $\psi \sim \text{Uniform}(a_\psi, b_\psi)$. The bounds for ψ cannot be specified arbitrarily since it is important to account for temporal range. We specified the following conditions for finding the bounds, $[a_\psi : [\mathbf{H}(a_\psi)]_{t,t'} = 0.95, |x_t - x_{t'}| = x_{\max}]$ and $[b_\psi : [\mathbf{H}(b_\psi)]_{t,t'} = 0.01, |x_t - x_{t'}| = x_{\min}]$, where $x_{\min}$ and $x_{\max}$ are the minimum and maximum temporal differences between visits. The remaining priors come directly from Section 2.5, however, because there was only one spatial observation type, we specified the following prior, $\boldsymbol{\kappa} \sim \text{IG}(0.001, 0.001)$. Finally, to promote shrinkage from the gamma process prior on the columns of the factor loadings matrix, we specified, $a_1 = 1$ and $a_2 = 20$. Inference proceeds using the MCMC sampler described in Section 2.3, with non-convergence evaluated primarily through examination of traceplots, but also the Geweke statistic (Geweke, 1992). For each simulated dataset, the MCMC sampler is run for 100,000 iterations after a 10,000 burn-in and thinned to a final sample size of 10,000. We call this method Model 1.

			WAIC					CRPS				
	Space	k	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
PSBP	Y	1	-1215	-997	-1212	-1203	-1058	0.408	0.465	0.409	0.413	0.416
		3	-1185	-972	-1182	-1006	-875	1.126	1.152	1.136	1.168	1.180
		6	-1073	-890	-1076	-922	-804	1.622	1.645	1.610	1.639	1.648
	N	1	-1219	-1013	-1219	-1195	-1073	0.409	0.463	0.409	0.412	0.412
		3	-1178	-966	-1178	-1044	-884	1.087	1.112	1.094	1.126	1.134
		6	-1050	-842	-1061	-907	-789	1.645	1.675	1.671	1.682	1.686
GP	Y	1	-1219	-1011	-1217	-1195	-1073	0.410	0.463	0.409	0.412	0.412
		3	-1178	-966	-1178	-1044	-884	1.087	1.112	1.094	1.126	1.134
		6	-1050	-842	-1061	-907	-789	1.645	1.675	1.671	1.682	1.686
	N	1	-1199	-982	-1199	-1181	-1056	0.426	0.474	0.425	0.427	0.428
		3	-1079	-881	-1063	-933	-826	1.253	1.231	1.254	1.295	1.311
		6	-977	-838	-985	-824	-759	1.519	1.508	1.517	1.517	1.524

Table 1: Assessing the performance of the spatial stick-breaking process and multiplicative gamma process shrinkage prior. Simulation settings are defined by the number of true underlying factors ($k = 1, 3, 6$), whether spatial dependency is present (Y: $\mathbf{F}(0.99)$, N: $\mathbf{I}$), and whether the PSBP or Gaussian process (GP) was used. Each of the simulated datasets has 10 uniform time points, which is the average in our visual field dataset (i.e., $T = 10$). Model fit was assessed using widely applicable information criterion (WAIC) and prediction performance was defined as accuracy of predicting the 13^{th} time point, and is determined by the continuous ranked probability score (CRPS). Smaller values are preferred for both. Each summary is based on 100 simulated datasets.

In order to compare our introduced methodology, we compared it to various simplifications. Model 2 removed the spatial component, setting $\mathbf{F}(\rho) = \mathbf{I}$. Model 3 removed the gamma shrinkage prior, instead using independent priors, $\delta_h \sim \text{Ga}(a_1, a_2)$. Models 4 and 5 replaced the PSBP prior with a standard multivariate CAR prior for each column of the factor loadings matrix, comparable to the model of Lopes et al. (2008). Furthermore, Models 4 and 5 removed the multiplicative gamma process shrinkage prior, instead using the same criteria as Model 3. Finally, Model 5 removed spatial dependency, using the identity matrix. All models were fit assuming six underlying latent factors (i.e., $k = 6$).

We compared the five models using both a model fit and prediction summary. Model fit was assessed using widely applicable information criterion (WAIC) and prediction performance was defined as accuracy of predicting a 13^{th} simulated time point, and was measured by the continuous ranked probability score (CRPS) (Hersbach, 2000; Vehtari et al., 2017). Smaller values are preferred for both.

Results are found in Table 1. We begin by studying the results for model fit. The most clear conclusion is that Models 1 and 3, the models that have the spatial PSBP (Model 3 loses the multiplicative gamma process shrinkage prior), perform the best across all settings. The difference between Models 1 and 3 is minimal, but inclusion of the gamma shrinkage prior (Model 1) does normally fit better. The only time Model 3 outperforms Model 1 is when the true number of latent factors is the same as the simulated data (i.e., $k = 6$), indicating that the gamma shrinkage prior is useful when the true number of fac-

tors is less than the number specified in the model. Another valuable comparison is with Models 4 and 5, which do not use the stick-breaking construction or the gamma shrinkage prior (Model 5 also does not include space), and have worse performance. Clearly, in the presence of space the spatial PSBP is crucial for model fit. The prediction results mirror the same trend as the model fit, with Models 1 and 3 being superior. Finally, we can compare the results of the simulation across settings dependent on whether the PSBP or Gaussian process were used. The setting where a Gaussian process was used is reflective of existing models such as the one from Lopes et al. (2008). In general, we see little changes across this simulation setting, indicating that the PSBP and Gaussian process data generating mechanisms may be similar. This seems to indicate that regardless of the data generating mechanism, the PSBP has superior performance, and therefore has utility over models with only a Gaussian process specification.

Of course gains in performance need to be considered within the context of computation time. In this simulation study, the average (SD) runtime in minutes for models 1–5, respectively, was 273 (58), 260 (54), 268 (54), 279 (60), and 234 (52), when using the `spBFA` R package. Therefore, as expected the models with spatial variability took the longest to run, while Model 5, with only Gaussian processes and no spatial variability was shortest. This seems to indicate that the increased model performance from Model 1 has utility as it has a runtime comparable to the other models. However, we should note that the MCMC sampler was optimized for Model 1 and the computation time for Models 4 and 5 can likely be lowered substantially.

3.2 Clustering Using the Spatial PSBP

In our second simulation study, we aimed to demonstrate the use of the spatial PSBP to cluster temporal changes across space. In order to do this we used a similar data generating process as the simulation in Section 3.1, however we made some key changes. In particular, we simulated data based on two true clusters, where the first cluster represented the region of the visual field called the inferior nasal, which includes eight spatial locations. The second cluster consisted of the remaining locations on the visual field.

Data was generated from point-wise logistic regression models where the intercepts and slopes were drawn jointly from a spatial process, $\mathrm{N}_{2m}((\boldsymbol{\beta}_0^\top, \boldsymbol{\beta}_1^\top)^\top, \boldsymbol{\kappa} \otimes \mathbf{F}(\rho))$. Here, $\boldsymbol{\kappa}$ was a 2×2 dimensional covariance, with entries, $[\boldsymbol{\kappa}]_{11} = 4$, $[\boldsymbol{\kappa}]_{21} = -0.5$, and $[\boldsymbol{\kappa}]_{22} = 2$. Both the intercept ($\boldsymbol{\beta}_0$) and slope ($\boldsymbol{\beta}_1$) were piece-wise constant, with the components corresponding to the second cluster equal to β_0 and β_1, respectively, and the first cluster, $\beta_0 + \delta_{\beta_0}$ and $\beta_1 + \delta_{\beta_1}$. Once the intercepts and slopes had been generated across the visual field they were used in point-wise regressions, where the mean squared error (MSE) was set to σ^2 in the second cluster, and $\sigma^{2*} = \sigma^2 + \delta_{\sigma^2}$ in the first cluster. Our simulation settings looked at varying magnitudes of δ_{β_0}, δ_{β_1} and δ_{σ^2}, which dictated variability across clusters. We set $\beta_0 = -8$, $\beta_1 = -4$, and $\sigma^2 = 3$, and then looked at $\delta_{\beta_0} = 0, 6$, $\delta_{\beta_1} = 0, 3, 6$ and $\delta_{\sigma^2} = 0, 2$. We allowed for both spatial (Y: $\rho = 0.99$) and independent (N: $\rho = 0$) processes.

We compared the clustering technique introduced in Section 2.6 to a simplified version that performed k-means on the raw data. For comparison, we used the ratio of between sum of squares (BSS), $\sum_{o=1}^{O} \sum_{i=1}^{m} (\hat{\mathbf{w}}(\mathbf{s}_{i,o}) - \bar{\mathbf{w}})^2$, over total SS (TSS),

$\sum_{o=1}^{O}\sum_{i=1}^{m}(\mathbf{w}(\mathbf{s}_{i,o})-\bar{\mathbf{w}})^2$, and $\hat{\mathbf{w}}(\mathbf{s}_{i,o})$ represented a fitted cluster mean, with $\hat{\mathbf{w}}(\mathbf{s}_{i,o}) = \sum_{o=1}^{O}\sum_{i=1}^{m}\mathbf{w}(\mathbf{s}_{i,o})1\{\mathbf{w}(\mathbf{s}_{i,o}) \in c\}/\sum_{o=1}^{O}\sum_{i=1}^{m}1\{\mathbf{w}(\mathbf{s}_{i,o}) \in c\}$ if $\mathbf{w}(\mathbf{s}_{i,o})$ belonged to cluster c. The quantity $\bar{\mathbf{w}}$ was the overall mean, $\sum_{o,i}\mathbf{w}(\mathbf{s}_{i,o})/(Om)$. For adequate clusters, we would have expected this ratio to be close to one, because a large BSS indicates high variability between clusters (and accordingly, small variability within clusters). We present the SS ratio (BSS / TSS) for the raw clustering method, and also for the PSBP technique with $\pi_k = 0.2, 0.3, 0.4, 0.5$. We presented results only for Models 1–3, since they are the only models with clustering capabilities.

The results of the simulation can be found in Figure 1. We can interpret the clustering performance of the PSBP across values of π_k in absolute and relative terms (i.e., compared to the raw clustering). In general, we can see that, in relative terms, the PSBP has improved clustering performance. In the only settings where the raw clustering technique has better relative performance ($\delta_{\beta_0} = 0$, $\delta_{\beta_1} = 0$ or 3 and $\delta_{\sigma^2} = 2$), the results are negligible, because the SS are close to zero, meaning the settings were overly difficult. Overall, it appears that the biggest boost in performance for the PSBP is when the true underlying intercepts are different between groups. There is also evidence that the PSBP process is capable of detecting clusters based on differences in the underlying true slope (i.e., δ_{β_1}), which is particularly impactful for clustering spatial locations based on temporal trajectories.

We see that the results of the simulations are relatively robust to the choice of π_k. Interestingly, the greatest performance corresponds to $\pi_k = 0.2$, which indicates the utility of the multiplicative gamma process shrinkage prior, as this setting limits the clustering to only the most informative latent factors. This is confirmed when looking at the average (SD) number of latent factors used for clustering across values of π_k, 2.11 (1.02), 4.17 (1.35), 6.00 (0.00), 6.00 (0.00), for $\pi_k = 0.2, 0.3, 0.4, 0.5$, respectively. This indicates that $\pi_k = 0.2$ is likely preferred, as the true number of clusters in the simulated data was two.

Finally, the average (SD) runtime for Models 1–3, respectively, was 225 (19), 221 (18), 221 (18) minutes, using the `spBFA` R package. Thus, the model that included both the gamma shrinkage prior and the spatial variability was 4 minutes longer on average. As with the first simulation, for each simulated dataset, the MCMC sampler was run for 100,000 iterations after a 10,000 burn-in and thinned to a final sample size of 10,000.

4 Data Illustrations

4.1 Glaucoma Progression Using Visual Fields

In our first case study using real data, we used the spatial PSBP to determine glaucoma progression from longitudinal visual fields. Glaucoma, an optic neuropathy, is the leading cause of irreversible vision loss worldwide. Although glaucomatous damage is irreversible, early treatment can usually prevent or slow down progression to functional damage and visual impairment. Estimation of rates of functional deterioration by visual fields is essential for determining patient prognosis and aggressiveness of therapy (Weinreb et al., 2014).

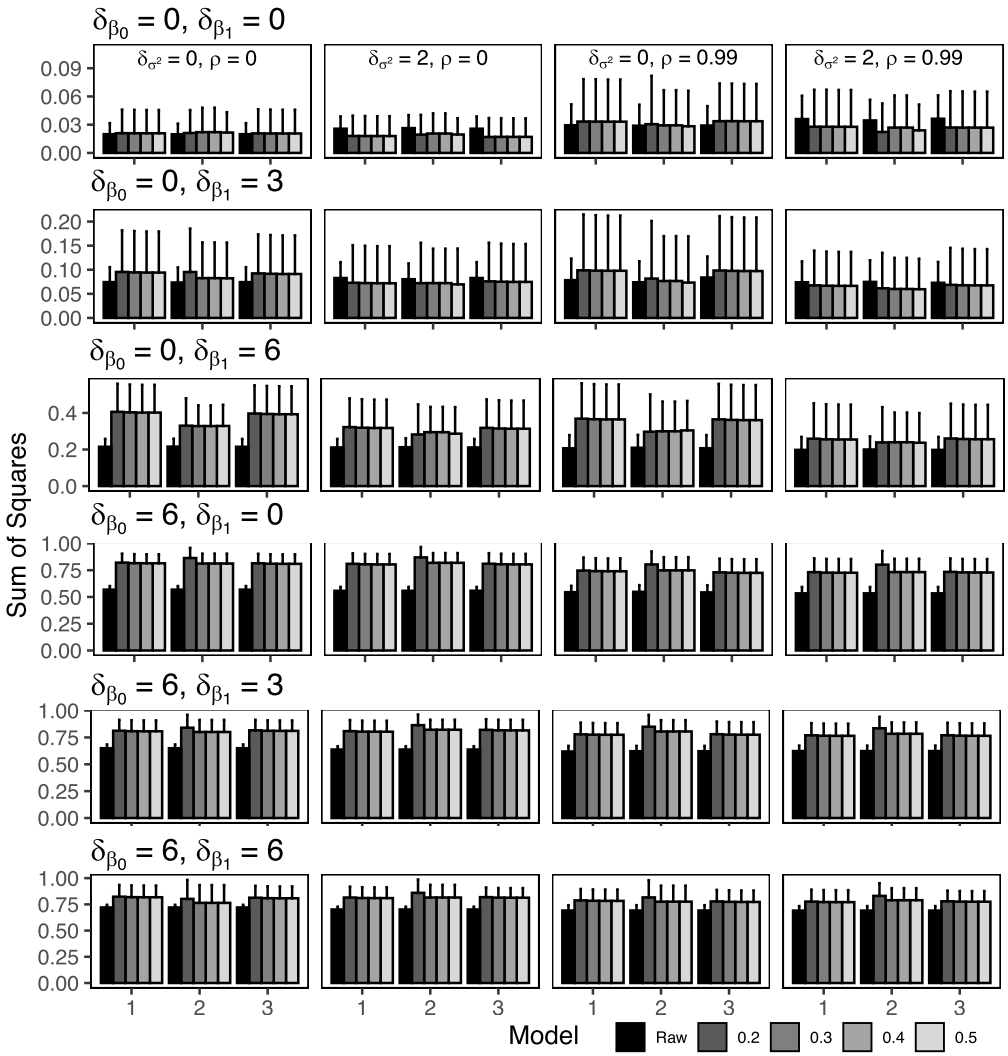

Figure 1: Assessing the clustering performance of the spatial PSBP and multiplicative gamma process shrinkage prior. Simulations are based on a true setting with two clusters. The first cluster is generated from a linear regression model with mean values of intercept, slope, and variance, $\beta_0 = -8$, $\beta_1 = -4$, and $\sigma^2 = 3$. The second cluster is simulated from the following mean parameters, $\beta_0^* = \beta_0 + \delta_{\beta_0}$, $\beta_1^* = \beta_1 + \delta_{\beta_1}$, and $\sigma^{2*} = \sigma^2 + \delta_{\sigma^2}$, where δ_{β_0}, δ_{β_1} and δ_{σ^2} dictate the variability across clusters. Furthermore, model parameters were simulated either from an independent ($\rho = 0$) or spatial ($\rho = 0.99$) process. The PSBP clustering from Section 2.6 with $\pi_k = 0.2, 0.3, 0.4, 0.5$ was compared with k-means performed on the raw data. We present the ratio of between sum of squares (SS) over total SS. Greater values of SS are preferred.

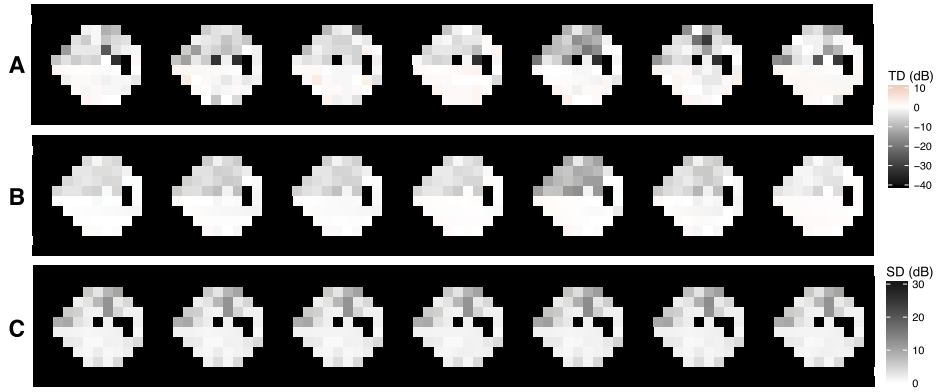

Figure 2: Example longitudinal series of visual fields, presented in total deviation (TD), a measure of age-adjusted loss. Negative values of TD indicate poorer vision. Clinicians are tasked with determining whether the rate of progression is clinically significant for intervention.

Visual fields are a psychophysical procedure that assesses a patient's field of vision, with standard automated perimetry (SAP) being the default method. In this study, we analyzed fields generated from the Humphrey Field Analyzer-II (HFA-II; Carl Zeiss Meditec Inc., Dublin, CA). The HFA-II is an interactive technology that assesses a patient's reaction as light is systematically introduced at gridded locations across their visual field. In this study, we represented functional loss using total deviation (TD) values, an age-adjusted measure of sensitivity loss, measured in decibels (dB). TD is a continuous measure, with large negative values indicating functional loss. An example longitudinal series of visual fields can be found in Figure 2A. Our data included 79 patients (110 eyes) diagnosed with glaucoma at baseline, with an average of 10 clinic visits and 4 years of follow-up. Of the 110 glaucomatous eyes, 51 (46%) were defined as progressing and the remaining 59 (54%) were stable. More details about the glaucoma population are given in Section 1 of the online Supplementary Materials. See Berchuck et al. (2019b) for a more in depth introduction to visual fields for statisticians.

Rates of Glaucoma Progression

The spatial factor analysis was fit to each of the 110 eyes in our study. For this case study, the MCMC sampler used a burn-in of 50,000 samples, 300,000 subsequent iterations, and then thinned to a final sample of 10,000. Across the 110 eyes, the MCMC chains ran for 554 (152) minutes on average (SD) using the `spBFA` R package. For the example longitudinal series of visual fields, presented in Figure 2, we present posterior mean and standard deviation fits at each clinic visit. From this visualization, we see that the method is properly spatially smoothing the observed data to better reveal patterns across time.

We are interested in using the spatial factor analysis model to improve clinicians' ability to quantify rates of change across time. Our introduced methodology is appro-

priate for assessing longitudinal changes on the visual field, because instead of analyzing each location, it models temporal changes of underlying regions. This is much closer to how a clinician interprets change on the visual field, as glaucoma has characteristic patterns. To this end, we performed independent linear regressions of the posterior mean estimates of the latent factors across time. The two-sided p-values from these regressions were used as predictors of progression. In particular, we present five variations that include, all six of the factors (i), the first three (ii), and only the first (iii), second (iv), and third (v) factors. To assess the diagnostic ability of these methods to discriminate progression status, we performed logistic regressions of the resulting p-values of the latent factors across time, using the predicted probabilities of progression as a diagnostic.

We compared the probabilities from the latent factors to established methods of determining progression on SAP fields, mean deviation (MD) and pattern standard deviation (PSD). Both MD and PSD are age-adjusted measures of vision loss on a visual field, with MD representing a global loss and PSD indicating the level of localized loss. In practice MD (and PSD) is used to assess rates of progression using ordinary least squares (OLS) regression across time, with a lower (and upper) p-value less than 0.05 indicating progression. Again, p-values were regressed against disease status and predicted progression probabilities were obtained.

We compared diagnostics using area under the receiver operating characteristic (ROC) curve (AUC) and partial AUC (pAUC). Larger values of AUC and pAUC indicate superior discriminatory ability. Based on the precedent of a previous study, we limited the pAUC to regions of clinically relevant specificity, 85–100% (Berchuck et al., 2019a).

Of the two established methods, PSD had better performance with an AUC of 0.65 and pAUC of 0.17 (Figure 3). When all six latent factors were included in the analysis, the AUC was improved slightly, however the pAUC decreased. This is problematic, because the pAUC is clinically more meaningful than overall AUC. Through inspection of the posterior factors, we determined that for the majority of eyes only three factors contained meaningful data (a result of the shrinkage prior on the loadings matrix). When we limited the metric to only contain the first three factors, the pAUC nearly doubled with a maximum pAUC of 0.28. When further exploring each of the first three factors independently, it became clear that the meaningful information had been encoded in the second factor, as the corresponding ROC curve has a much steeper trajectory from 100% specificity. These results indicate that the rates of change learned from the posterior latent factors have clinical utility in determining underlying structural progression from visual fields.

Sensitivity to Choice of Hyperparameters

Since our proposed model has multiple layers with numerous hyperparameters, we present a thorough sensitivity analysis of the results in the previous section. In particular, we assess sensitivity of our results to four modeling assumptions: i) the number of latent factors, k, ii) the base parameter in the multiplicative gamma process shrinkage prior, a_1, iii) along with the multiplicative parameter, a_2, and iv) the criterion for the bounds of ψ. For each setting, we replicated the results from Figure 3, which looked at

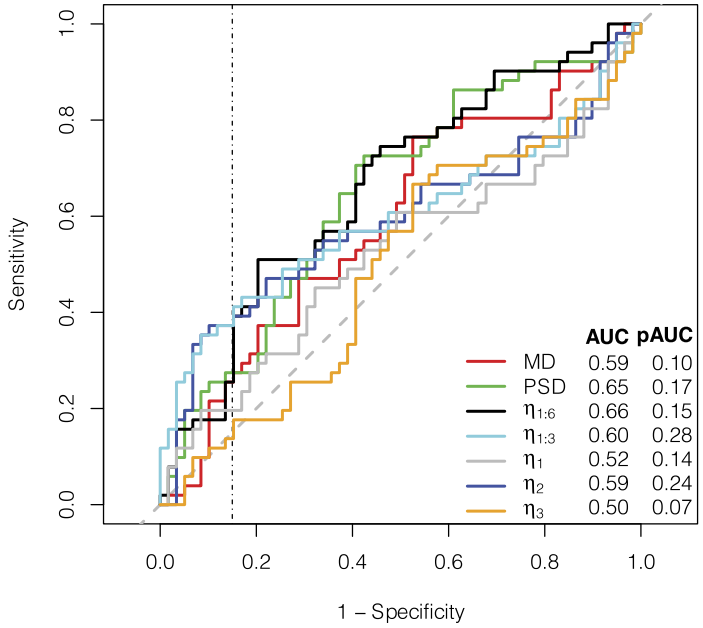

Figure 3: Receiver operating characteristic (ROC) curves for metrics diagnosing structural glaucomatous progression. Established metrics mean deviation (MD) and pattern standard deviation (PSD) are compared with metrics derived from regressing the posterior latent factors across time. Also, presented are area under the ROC curve (AUC) and partial AUC (pAUC), limited to the region of specificity greater than 85%.

diagnostic performance of the posterior factors for discriminating progression status. In this sensitivity analysis, we look at version (i), that included all latent factors, since it is invariant across our sensitivity settings when changing the number of latent factors. The results of the sensitivity analysis, which display AUC and pAUC at 85% specificity, are displayed in Table 2.

In the original analysis, the number of latent factors, k, is set to 6, while values of 2, and 10 are used for comparison in the sensitivity analysis. To demonstrate sensitivity to the choice of prior for the multiplicative gamma process shrinkage prior, we allowed the hyperparameters, a_1 and a_2 to vary from the following values we used in the original analysis: $a_1 = 1$, and $a_2 = 20$. A value of $a_1 = 3$, will yield smaller variances compared to a value of 1, holding a_2 constant, while changing a_2 to a value of 10, results in the shrinkage process becoming slower, compared to a value of 20, and holding a_1 constant. Finally, we compare the original results to a version where the upper bound of ψ was based on the following criterion, $[b_\psi : [\mathbf{H}(b_\psi)]_{t,t'} = 0.05, |x_t - x_{t'}| = x_{\min}]$, changed from the original 0.01. This criterion is motivated to give narrower and symmetric bounds.

The results from the sensitivity analysis show an interesting relationship between the choice of hyperparameters for the multiplicative gamma process shrinkage prior and the

			AUC			pAUC		
a_1	a_2	b_ψ	$k=2$	$k=6$	$k=10$	$k=2$	$k=6$	$k=10$
1	10	0.05	0.60	0.69	0.71	0.12	0.29	0.31
1	10	0.01	0.54	0.62	0.71	0.11	0.09	0.24
1	20	0.05	0.60	0.60	0.68	0.14	0.14	0.23
1	20	0.01	0.56	**0.66**	0.61	0.10	**0.15**	0.27
3	10	0.05	0.55	0.62	0.57	0.16	0.16	0.18
3	10	0.01	0.56	0.64	0.75	0.12	0.25	0.20
3	20	0.05	0.58	0.65	0.69	0.16	0.20	0.24
3	20	0.01	0.52	0.70	0.69	0.10	0.21	0.12

Table 2: Sensitivity analysis to various hyperparameter assumptions in the analysis of visual field data and glaucoma progression risk. For each set of assumptions, we present area under the receiver operating characteristic curve (AUC) and partial AUC (pAUC) for discriminating progression status. Assumptions include i) the number of latent factors, k, ii) the base parameter in the multiplicative gamma process shrinkage prior, a_1, iii) along with the multiplicative parameter, a_2, and iv) the criterion for the bounds of ψ. The results from the original analysis are bolded.

number of latent factors, as related to the diagnostic performance in predicting glaucoma progression status. While at first it appears the performance increases as the number of latent factors increases, it actually appears that this increase in performance depends on the hyperparameters of the shrinkage prior. When you look at the scenario where the shrinkage prior has the highest variance and slowest shrinkage ($a_1 = 1, a_2 = 10$), you see that the performance generally increases as the number of latent factors increases. However, when the shrinkage prior has a smaller initial variance and faster shrinkage ($a_1 = 3, a_2 = 20$), the difference in performance is not as clear, and in fact the setting with 6 latent factors performs very well. This seems to indicate the importance of choosing the number of latent factors and hyperparameters for the shrinkage prior with care and while considering the form of the data. In our data setting, we were analyzing visual fields, and therefore chose the number of latent factors to be equal to the number of meaningful anatomical regions of the retina. Then, our hyperparameters for the shrinkage prior allow for some shrinkage, as it is unlikely that each anatomical region has a unique rate of progression.

Clustering Temporal Trends

We close this data illustration by demonstrating the clustering ability of the spatial PSBP detailed in Section 2.6. In Figure 4A, for each latent factor, the posterior probabilities of belonging to the same cluster, $g_j(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})$, are presented. For this eye, only the first three factors contain meaningful information. This is further reinforced in Figure 4B, where the stacked posterior probabilities $\mathbf{w}(\mathbf{s}_{i,o})$ are presented across the spatial surface and observation type ($\mathbf{w}$). On the left of Figure 4B, the probabilities are presented in their original order (i.e., un-ordered), while the right frame represents the ordered factors (according to k-means with two groups, determined using the gap statistic). The ordered version removed the three non-informative factors based on the

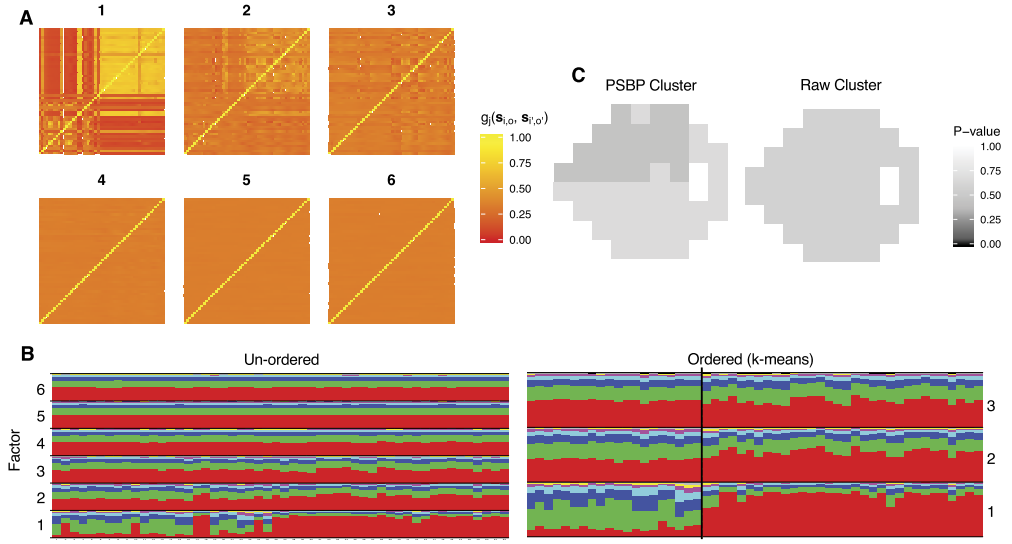

Figure 4: Demonstrating the clustering ability of the spatial PSBP for the example patient in Figure 2. Frame A contains the posterior probabilities of belonging to the same cluster, $g_j(\mathbf{s}_{i,o}, \mathbf{s}_{i',o'})$, for each factor. On the left of frame B, the stacked posterior probabilities $\mathbf{w}(\mathbf{s}_{i,o})$ are presented in their original order, while on the right the ordered version is presented, with only the meaningful factors included. Finally, in frame C, clusters are presented with p-values, which are the average lower p-values from point-wise logistic regressions of the PPD across all locations.

criteria from Section 2.6, with $\pi_k = 0.2$. Due to the identifiability issue, the right of Figure 4B is required for clustering.

We presented the clusters obtained from the spatial PSBP (left) and clustering of the raw data using k-means with one cluster, based on the gap statistic (right). The clusters, presented in Figure 4C, show p-values, which are the average lower p-values from point-wise regressions of the PPD across all locations within each cluster. From this presentation, we can see that the spatial PSBP was capable of producing a map with regions of varying temporal trends that are straightforward for clinicians to interpret. While clustering the raw data was non-informative, the PSBP illustrated a region with faster temporal trajectories, that may require intervention, and corresponds with true progression (Figure 2A).

4.2 Malaria Incidence in Peru

In our second case study using real data, we investigated rates of malaria across Loreto, Peru's northernmost region that is located in the Amazon rainforest. This case study is a nice complement to the glaucoma example, as it introduces additional complexities,

including dealing with non-Gaussian data, having multiple spatial observation types (i.e., malaria types), and introducing covariates into the mean process.

To model malaria counts we used conventions based on the Peruvian Ministry of Health, which defines significant levels of malaria based on observed counts from previous years. In particular, define the counts of malaria $c_t(\mathbf{s}_{i,o})$ at time t for district i and malaria type o. For malaria data, the temporal index is actually composed of two dimensions that represent year and epidemiological week, $t = \{y, w\}$, for $y = 2012, \ldots, 2018$, and $w = 1, \ldots, 52$. Furthermore, the Loreto region has 51 districts and there are two types of malaria monitored, *P. falciparum* ($o = 1$) and *P. vivax* ($o = 2$).

In order to model malaria counts, we defined the proportion of malaria $p_{yw}(\mathbf{s}_{i,o}) = c_{yw}(\mathbf{s}_{i,o})/n_{yw}(\mathbf{s}_{i,o})$, where $n_{yw}(\mathbf{s}_{i,o})$ represents the population. Then, we defined our outcome as $Y_{yw}(\mathbf{s}_{i,o}) = 1\{p_{yw}(\mathbf{s}_{i,o}) > \mathbf{p}_{yw}(\mathbf{s}_{i,o})\}$, where $\mathbf{p}_{yw}(\mathbf{s}_{i,o}) = \sum_{x=(y-5)}^{y-1} p_{xw}(\mathbf{s}_{i,o})/5$, is the average proportion of cases for a malaria type at a location over the past five years. This binary definition, allowed us to model the probability of exceeding the number of cases seen in the past five years, $\pi_{yw}(\mathbf{s}_{i,o})$, an important indicator for specifying interventions.

To best model the mean process, we incorporated the following covariates through $\mathbf{x}_t(\mathbf{s}_{i,o})$, rainfall (millimeters), temperature (Celcius), and an indicator of being in the rainy season, which is approximately from February-July, but defined using weeks, $1(w \in \{6, \ldots, 31\})$, in addition to an intercept, so that $p = 4$ (Vittor et al., 2009). To demonstrate the spatial PSBP methodology, we used data from the entire 2017 year and the first five weeks of 2018, in order to predict malaria severity into the rainy season of 2018. This yielded the following set of temporal observations, $\{2017, 1\}, \ldots, \{2017, 52\}$, $\{2018, 1\}, \ldots, \{2018, 5\}$, resulting in $t = 1, \ldots, T$, with $T = 57$.

To model the malaria indicator we used a Bernoulli likelihood, specified as follows, $Y_t(\mathbf{s}_{i,o}) \sim \text{Bernoulli}(\pi_t(\mathbf{s}_{i,o}))$, where $\pi_t(\mathbf{s}_{i,o}) = g^{-1}(\vartheta_t(\mathbf{s}_{i,o}))$. The likelihood is then as follows,

$$\prod_{t=1}^{T}\prod_{o=1}^{O}\prod_{i=1}^{m}\frac{\exp\{\vartheta_t(\mathbf{s}_{i,o})Y_t(\mathbf{s}_{i,o})\}}{1+\exp\{\vartheta_t(\mathbf{s}_{i,o})\}}.$$

While inference can proceed using this likelihood, it is computationally intensive due to a loss of conjugacy across the majority of parameters. Computation can be made feasible through data augmentation using Pólya–Gamma (PG) latent variables (Polson et al., 2013).

In particular, based on Polson et al. (2013), we chose to model the observed data through a joint likelihood, $f(Y_t(\mathbf{s}_{i,o}), \omega_t(\mathbf{s}_{i,o})|\vartheta_o(\mathbf{s}_{i,o}))$, that consists of an augmented parameter with distribution, $\omega_t(\mathbf{s}_{i,o}) \sim \text{PG}(1, 0)$. The reason for this is that the conditional distribution $f(\omega_t(\mathbf{s}_{i,o})|\vartheta_t(\mathbf{s}_{i,o}))$ is a tilted version of the PG with the following density,

$$\left[\frac{(1+\exp\{\vartheta_t(\mathbf{s}_{i,o})\})}{(2\exp\{\vartheta_t(\mathbf{s}_{i,o})/2\})}\right]\exp\left\{-0.5\vartheta_t^2(\mathbf{s}_{i,o})\omega_t(\mathbf{s}_{i,o})\right\}f(\omega_t(\mathbf{s}_{i,o})),$$

using the fact that $\cosh\{x\} = (1+\exp\{2x\})/(2\exp\{x\})$.

This is useful, because the likelihood can be expressed in the form of a Gaussian kernel, $\prod_{t=1}^{T}\prod_{o=1}^{O}\prod_{i=1}^{m}\exp\{-0.5\omega_t(\mathbf{s}_{i,o})(Y_t^*(\mathbf{s}_{i,o})-\vartheta_t(\mathbf{s}_{i,o}))^2\}$, where $\chi_t(\mathbf{s}_{i,o}) = Y_t(\mathbf{s}_{i,o}) - 1/2$ and $Y_t^*(\mathbf{s}_{i,o}) = \chi_t(\mathbf{s}_{i,o})/\omega_t(\mathbf{s}_{i,o})$. This can be further expressed in vector form, $\prod_{t=1}^{T}\exp\{-\frac{1}{2}(\mathbf{Y}_t^*-\mathbf{X}_t\boldsymbol{\beta}-\boldsymbol{\Lambda}\boldsymbol{\eta}_t)^T\boldsymbol{\Delta}_t(\mathbf{Y}_t^*-\mathbf{X}_t\boldsymbol{\beta}-\boldsymbol{\Lambda}\boldsymbol{\eta}_t)\}$, with $\boldsymbol{\Delta}_t = \text{diag}(\boldsymbol{\omega}_t)$ and $\mathbf{Y}_t^*$, $\boldsymbol{\vartheta}_t$ and $\boldsymbol{\omega}_t$ defined as vectors stacked the same as the original $\mathbf{Y}_t$. Because this is the kernel of a Gaussian, we maintain conjugacy and will only need to add an additional sampling step for $\omega_t(\mathbf{s}_{i,o})$. More details of this derivation are given in Section 5 of the Supplementary Materials online.

For this case study, we used a burn-in of 50,000 samples, 500,000 subsequent iterations, and then thinned to a final sample of 10,000. This MCMC sampler took 2,563 minutes to run using `spBFA`. To demonstrate the spatial PSBP process using the malaria data, we present predictions for the rainy season in 2018 using the prediction theory from Section 2.7. Results are presented from predictions of the fifth week of the rainy season (i.e., t = 62 $\iff$ $\{y = 2018, w = 10\}$). Figures 5A and B show the proportion, $p_t(\mathbf{s}_{i,o})$, and indicator outcomes, $Y_t(\mathbf{s}_{i,o})$, across Loreto for both types of malaria. Then, in frames C and D of Figure 5, we presented the posterior mean predicted probabilities of exceeding the average amount of malaria and their standard deviations (SD), respectively. While these posterior predictions are useful, they can be difficult for government agencies to draw actionable decisions.

Fortunately, the clustering method of the spatial PSBP, introduced in Section 2.6, can be used to provide more actionable information about where to focus intervention efforts. In Figure 6A, the resulting clusters are presented, where it can be seen that four groups were found. Recall, that clustering is based on temporal changes, so that within groups, locations across both malaria types are presumed to have similar rates of change. The clusters themselves are only informative about groupings and do not differentiate on speeds of change, so, similar to the glaucoma visual fields in Figure 4, we presented p-values over the clusters in Figure 6B. The p-values are the average upper p-values from point-wise logistic regressions of the PPD across all locations and malaria types within each cluster.

The representation of malaria incidence in Figure 6B is useful, because it allows for the Ministry of Health to focus interventions on areas across Loreto that have similarly increasing rates of malaria. For example, in Figure 5C, we can see that for *P. vivax* the most northern district has a posterior mean probability of close to one. While interventions should likely be performed in the district, based on Figure 6B, it appears that districts centered near (5°S, 74°W) all belong to a cluster with a p-value around 0.2 (the smallest of all the clusters). This information, not available from standard predictions, can alert the Ministry of Health to allocate resources to these districts.

5 Summary

We have provided a factor analysis that can be used in the setting of spatial correlation. The spatial dependencies are introduced through a spatial BNP prior on the columns of the factor loadings matrix. The prior incorporates a multiplicative gamma

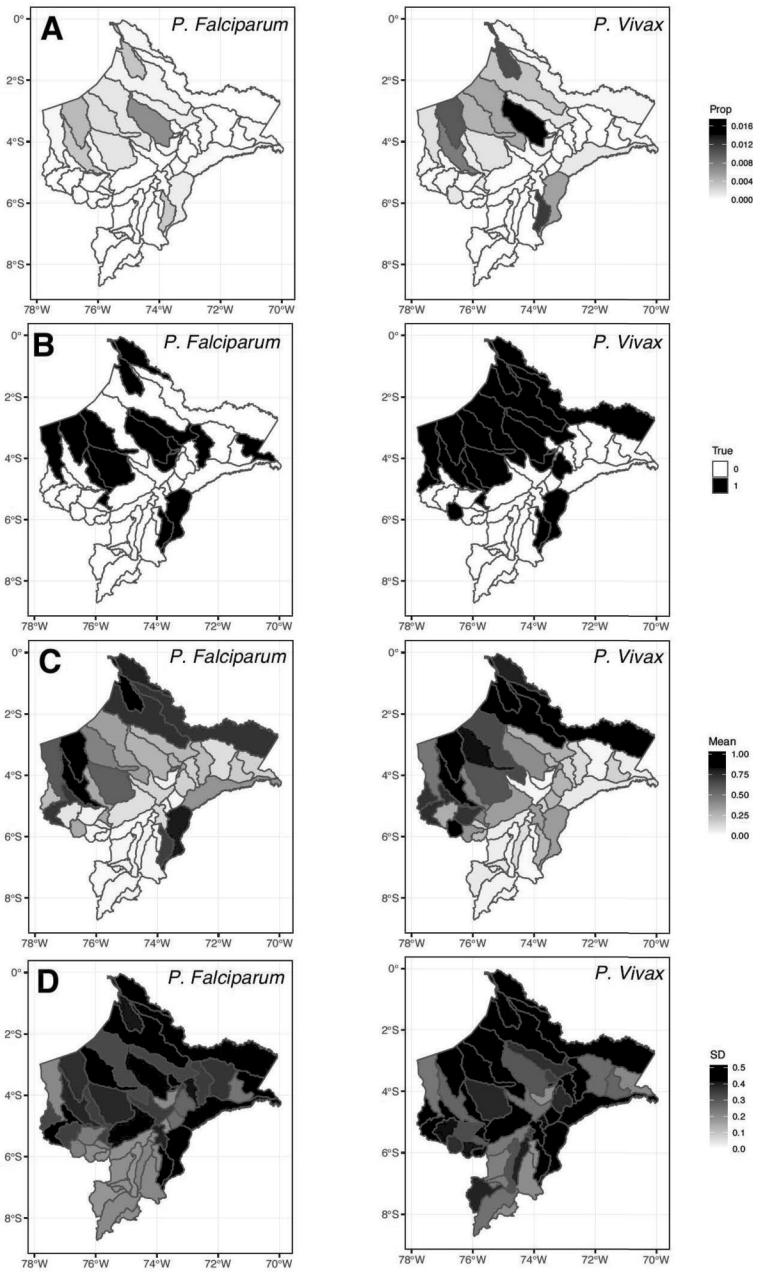

Figure 5: Presenting malaria predictions for the fifth week of the 2018 rainy season. A and B show the true proportion, $p_t(\mathbf{s}_{i,o})$, and indicator outcomes, $Y_t(\mathbf{s}_{i,o})$, across Loreto for both types of malaria. In frames C and D, we present the posterior mean predicted probabilities of exceeding the average amount of malaria and their standard deviations (SD), respectively.

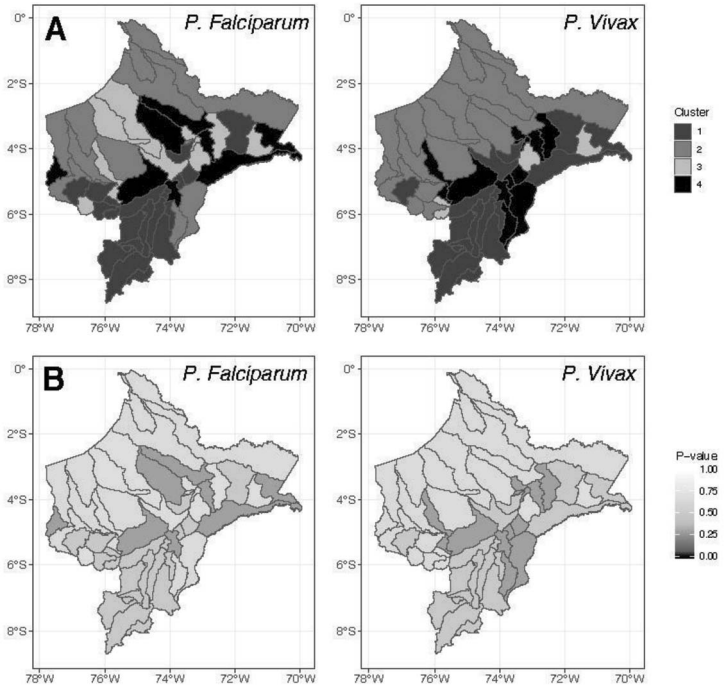

Figure 6: Presenting results from applying the PSBP clustering to the rates of malaria across the Loreto region. In frame A, four clusters are presented that have been identified to have similar temporal rates of change. Frame B presents the same clusters with p-values, which are the average upper p-values from point-wise logistic regressions of the PPD across all locations and malaria types within each cluster.

process shrinkage prior to adaptively learn the proper number of latent factors. We have described an efficient MCMC sampler to accompany the model, which has been published as an `R` package, `spBFA`. We have illustrated the model's performance in both simulated and real data examples. In particular, we showed that by encoding spatial information into the loadings matrix, meaningful factors were learned that respect the observed neighborhood dependencies, making them useful for assessing rates of change across space.

While much of the factor analysis literature performs dimension reduction on the mean process for Gaussian data, we place no such restriction, allowing for a general likelihood and non-linear relationship with the latent factors. We illustrated this through our modeling of malaria counts with a Bernoulli likelihood. Although our specification of both space and time are separable and stationary Gaussian processes, we have noted that the resulting model is neither separable, stationary nor Gaussian.

Importantly, our proposed model relies on the fully-specified factor loadings matrix from Bhattacharya and Dunson (2011). One of the important drawbacks of

Bhattacharya and Dunson (2011) is that the factors are not identifiable, hence inference can't be obtained directly on the factors. This also applies to the factors in our proposed model. However, we overcame this limitation by proposing a method for clustering temporal trajectories across a spatial surface that relies on the entire collection of factor loadings, which in their entirety are identifiable (Section 2.6). Through simulation in Section 3.2, we showed this technique was capable of identify true underlying spatial clusters of temporal change. We then confirmed this in two real-world data applications in Section 4. Finally, as noted in Bhattacharya and Dunson (2011), there are numerous cases where this fully-specified form of the factor loadings matrix is useful, including covariance estimation, and prediction. We confirmed this by using the posterior distribution of the factors to diagnose glaucoma disease progression in Figure 3.

Using our spatial PSBP prior we were able to find regions within the spatial domain with similar temporal trajectories, an important task in many applied settings. Through simulation, we showed that our clustering technique is preferred over a standard clustering routine. In real data, we showed that aggregating rates of change at the cluster level produced patterns that aided in making actionable decisions. The task of clustering trajectories has been addressed sparsely in the literature, with the majority of methods clustering trends with no shape constraint, like our temporal process. A recent method by Napier et al. (2018), instead, clusters trajectories on pre-specified parametric forms (e.g., linear). While this technique can be limiting, in applied settings with known trajectories, it has utility and therefore would be a valuable extension to our model.

Finally, this work opens up numerous avenues for future statistical research. If applicable, a seasonal component could be added to the temporal model of the latent factors, $\boldsymbol{\eta}_t$. Currently, we only allow for constant temporal smoothing, however this could easily be generalized to incorporate any type of seasonal effect, for example a Fourier dynamic linear model (West et al., 1985). Any temporal effect that has Gaussian errors will maintain conjugacy for the MCMC sampling. Other natural extensions to this approach include the treatment of sparse data and remedy to deal with a large number of m locations. Although, we did not treat these problems in the current manuscript, existing methods could be nicely incorporated in our modeling framework. For example, the highly scalable nearest-neighbor Gaussian process model, that allows for inference in settings of a large number of spatial locations, could be applied in a straightforward fashion (Datta et al., 2016).

Supplementary Material

Supplementary material to 'Bayesian Non-Parametric Factor Analysis for Longitudinal Spatial Surfaces' (DOI: 10.1214/20-BA1253SUPP; .pdf). The supplement contains details related to the model, including full conditionals, moment derivations, prediction theory, and also specifics of the glaucoma population.

References

Berchuck, S. I., Janko, M., Medeiros, F. A., Pan, W., and Mukherjee, S. (2020). "Supplementary material to 'Bayesian Non-Parametric Factor Analysis for Longitudinal Spatial Surfaces'." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1253SUPP. 5

Berchuck, S. I., Mwanza, J.-C., Tanna, A. P., Budenz, D. L., and Warren, J. L. (2019a). "Improved Detection of Visual Field Progression Using a Spatiotemporal Boundary Detection Method." *Scientific Reports*, 9(1): 4642. 18

Berchuck, S. I., Mwanza, J.-C., and Warren, J. L. (2019b). "Diagnosing glaucoma progression with visual field data using a spatiotemporal boundary detection method." *Journal of the American Statistical Association*, 114: 1063–1074. MR4011758. doi: https://doi.org/10.1080/01621459.2018.1537911. 17

Berchuck, S. I., Warren, J. L., Herring, A. H., Evenson, K. R., Moore, K. A., Ranchod, Y. K., and Diez-Roux, A. V. (2016). "Spatially modelling the association between access to recreational facilities and exercise: the 'Multi-ethnic study of atherosclerosis'." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(1): 293–310. MR3461578. doi: https://doi.org/10.1111/rssa.12119. 10

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291–306. MR2806429. doi: https://doi.org/10.1093/biomet/asr013. 2, 6, 25, 26

Christensen, W. F. and Amemiya, Y. (2002). "Latent variable analysis of multivariate spatial data." *Journal of the American Statistical Association*, 97(457): 302–317. MR1947288. doi: https://doi.org/10.1198/016214502753479437. 2

Chung, Y. and Dunson, D. B. (2009). "Nonparametric Bayes conditional distribution modeling with variable selection." *Journal of the American Statistical Association*, 104(488): 1646–1660. MR2750582. doi: https://doi.org/10.1198/jasa.2009.tm08302. 4

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). "Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets." *Journal of the American Statistical Association*, 111(514): 800–812. MR3538706. doi: https://doi.org/10.1080/01621459.2015.1044091. 26

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). "Generalized spatial Dirichlet process models." *Biometrika*, 94(4): 809–825. MR2416794. doi: https://doi.org/10.1093/biomet/asm071. 4

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005). "Bayesian nonparametric spatial modeling with Dirichlet process mixing." *Journal of the American Statistical Association*, 100(471): 1021–1035. MR2201028. doi: https://doi.org/10.1198/016214504000002078. 3

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC. MR3235677. 10

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to calculating posterior moments." In Bernardo, J. M., Berger, J., Dawid, A. P., and Smith, J. F. M. (eds.), *Bayesian Statistics 4*, 169–193. Oxford: Oxford University Press. MR1380276. 12

Ghosh, J. and Dunson, D. B. (2009). "Default prior distributions and efficient posterior computation in Bayesian factor analysis." *Journal of Computational and Graphical Statistics*, 18(2): 306–320. MR2749834. doi: https://doi.org/10.1198/jcgs.2009.07145. 2

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Kobe, R. K. (2013). "Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data." *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3): 274–298. MR3110894. doi: https://doi.org/10.1007/s13253-013-0140-3. 3

Hersbach, H. (2000). "Decomposition of the continuous ranked probability score for ensemble prediction systems." *Weather and Forecasting*, 15(5): 559–570. 13

Hogan, J. W. and Tchernis, R. (2004). "Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data." *Journal of the American Statistical Association*, 99(466): 314–324. MR2109313. doi: https://doi.org/10.1198/016214504000000296. 3

Lopes, H. F., Gamerman, D., and Salazar, E. (2011). "Generalized spatial dynamic factor models." *Computational Statistics & Data Analysis*, 55(3): 1319–1330. MR2741417. doi: https://doi.org/10.1016/j.csda.2010.09.020. 3

Lopes, H. F., Salazar, E., Gamerman, D., et al. (2008). "Spatial dynamic factor analysis." *Bayesian Analysis*, 3(4): 759–792. MR2469799. doi: https://doi.org/10.1214/08-BA329. 3, 12, 13, 14

MacEachern, S. N. (1999). "Dependent nonparametric processes." In *ASA Proceedings of the Section on Bayesian Statistical Science*, volume 1, 50–55. Alexandria, Virginia. Virginia: American Statistical Association; 1999. 3

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). "Bayesian Gaussian copula factor models for mixed data." *Journal of the American Statistical Association*, 108(502): 656–665. MR3174649. doi: https://doi.org/10.1080/01621459.2012.762328. 2

Napier, G., Lee, D., Robertson, C., and Lawson, A. (2018). "A Bayesian space-time model for clustering areal units based on their disease trends." *Biostatistics*, 00(0): 00–00. MR4019725. doi: https://doi.org/10.1093/biostatistics/kxy024. 26

Nethery, R. C., Sandler, D. P., Zhao, S., Engel, L. S., and Kwok, R. K. (2018). "A joint spatial factor analysis model to accommodate data from misaligned areal units with application to Louisiana social vulnerability." *Biostatistics*, 20(3): 468–484. MR3973121. doi: https://doi.org/10.1093/biostatistics/kxy016. 3

Nethery, R. C., Warren, J. L., Herring, A. H., Moore, K. A., Evenson, K. R., and Diez-Roux, A. V. (2015). "A common spatial factor analysis model for mea-

sured neighborhood-level characteristics: The Multi-Ethnic Study of Atherosclerosis." *Health & place*, 36: 35–46. 3

Pati, D. and Dunson, D. B. (2014). "Bayesian nonparametric regression with varying residual density." *Annals of the Institute of Statistical Mathematics*, 66(1): 1–31. MR3147543. doi: https://doi.org/10.1007/s10463-013-0415-z. 4

Pati, D., Dunson, D. B., and Tokdar, S. T. (2013). "Posterior consistency in conditional distribution estimation." *Journal of Multivariate Analysis*, 116: 456–472. MR3049916. doi: https://doi.org/10.1016/j.jmva.2013.01.011. 4

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American Statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 22

Reich, B. J. and Bandyopadhyay, D. (2010). "A latent factor model for spatial data with informative missingness." *The Annals of Applied Statistics*, 4(1): 439. MR2758179. doi: https://doi.org/10.1214/09-AOAS278. 3

Ren, Q. and Banerjee, S. (2013). "Hierarchical factor models for large spatially misaligned data: a low-rank predictive process approach." *Biometrics*, 69(1): 19–30. MR3058048. doi: https://doi.org/10.1111/j.1541-0420.2012.01832.x. 3

Rodriguez, A. and Dunson, D. B. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian Analysis*, 6(1): 145–177. MR2781811. doi: https://doi.org/10.1214/11-BA605. 4, 6, 8

Strickland, C., Simpson, D., Turner, I., Denham, R., and Mengersen, K. (2011). "Fast Bayesian analysis of spatial dynamic factor models for multitemporal remotely sensed imagery." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(1): 109–124. MR2758572. doi: https://doi.org/10.1111/j.1467-9876.2010.00739.x. 3

Tibshirani, R., Walther, G., and Hastie, T. (2001). "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2): 411–423. MR1841503. doi: https://doi.org/10.1111/1467-9868.00293. 11

Tzala, E. and Best, N. (2008). "Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality." *Statistical Methods in Medical Research*, 17(1): 97–118. MR2420192. doi: https://doi.org/10.1177/0962280207081243. 3

Vehtari, A., Gelman, A., and Gabry, J. (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. MR3647105. doi: https://doi.org/10.1007/s11222-016-9696-4. 13

Vittor, A. Y., Pan, W., Gilman, R. H., Tielsch, J., Glass, G., Shields, T., Sánchez-Lozano, W., Pinedo, V. V., Salas-Cobos, E., Flores, S., et al. (2009). "Linking deforestation to malaria in the Amazon: characterization of the breeding habitat of the principal malaria vector, Anopheles darlingi." *The American Journal of Tropical Medicine and Hygiene*, 81(1): 5. 22

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics—Simulation and Computation®*, 36(1): 45–54. MR2370888. doi: https://doi.org/10.1080/03610910601096262. 7

Wall, M. M. and Liu, X. (2009). "Spatial latent class analysis model for spatially distributed multivariate binary data." *Computational Statistics & Data Analysis*, 53(8): 3057–3069. MR2667610. doi: https://doi.org/10.1016/j.csda.2008.07.037. 3

Weinreb, R. N., Aung, T., and Medeiros, F. A. (2014). "The pathophysiology and treatment of glaucoma: a review." *JAMA*, 311(18): 1901–1911. 15

West, M., Harrison, P. J., and Migon, H. S. (1985). "Dynamic generalized linear models and Bayesian forecasting." *Journal of the American Statistical Association*, 80(389): 73–83. MR0786598. 26

Acknowledgments

This work was partially supported by the National Aeronautics and Space Administration (MJ and WP; NNX15AP74G S005), the National Institutes of Health/National Eye Institute (FAM; EY029885, EY027651, and EY021818), the Human Frontier Science Program (SM; RGP005), and the National Science Foundation (SM; DMS 17-13012, DBI 1661386, and DEB 1840223) as well as high-performance computing partially supported by grant 2016-IDG-1013 from the North Carolina Biotechnology Center (SM).