Bayesian Nonparametric Covariance Regression

Emily Fox

Duke University, Durham, NC USA.

David Dunson

Duke University, Durham, NC USA.

Summary. Although there is a rich literature on methods for allowing the variance in a univariate regression model to vary with predictors, time and other factors, relatively little has been done in the multivariate case. Our focus is on developing a class of nonparametric covariance regression models, which allow an unknown $p \times p$ covariance matrix to change flexibly with predictors. The proposed modeling framework induces a prior on a collection of covariance matrices indexed by predictors through priors for predictor-dependent loadings matrices in a factor model. In particular, the predictor-dependent loadings are characterized as a sparse combination of a collection of unknown dictionary functions (e.g, Gaussian process random functions). The induced covariance is then a regularized quadratic function of these dictionary elements. Our proposed framework leads to a highly-flexible, but computationally tractable formulation with simple conjugate posterior updates that can readily handle missing data. Theoretical properties are discussed and the methods are illustrated through simulations studies and an application to the Google Flu Trends data.

Keywords: **covariance estimation; Gaussian process; heteroscedastic regression; nonparametric Bayes; stochastic process.**

1. Introduction

Spurred by the increasing prevalence of high-dimensional datasets and the computational capacity to analyze them, capturing heteroscedasticity in multivariate processes has become a growing focus in many applied domains. For example, within the field of financial time series modeling, capturing the time-varying *volatility* and *co-volatility* of a collection of risky assets is key in devising a portfolio management scheme. Likewise, the spatial statistics community is often faced with multivariate measurements (e.g., temperature, precipitation, etc.) recorded at a large collection of locations, necessitating methodology to model the strong spatial (and spatio-temporal) variations in correlations. More generally, imagine that one has some arbitrary, potentially multivariate predictor space $\mathcal{X}$ and a collection of multivariate response vectors y. The problem of mean regression (i.e., $\mu(x) = E(y \mid x)$) has been well studied in both the univariate and multivariate settings. Although there is a rich literature on methods for allowing the variance in a univariate regression model to vary with predictors, there is a dearth of methodologies for the general case of multivariate covariance regression (i.e., $\Sigma(x) = \mathrm{cov}(y \mid x)$). The covariance matrix captures key correlations between the elements of the response vector, and the typical assumption of a homoscedastic model can have significant impact on inferences.

Historically, the problem of multivariate covariance regression has typically been addressed by standard regression operations on the unconstrained elements of the log or

Cholesky decomposition of the covariance (or precision) matrix. For example, Pourahmadi [1999] proposes to model elements of $\mathrm{chol}(\Sigma(x)^{-1})$ as a linear function of the predictors. The weights associated with the ith row have a nice interpretation in terms of the conditional distribution of y_i given $y_1, y_2, \ldots, y_{i-1}$; however, the model is not invariant to permutations of the elements of y which is problematic in applications where there does not exist a natural ordering. Alternatively, Chiu et al. [1996] consider modeling each element of $\log(\Sigma(x))$ as a linear function of the predictor. An issue with this formulation is in the interpretability of the model: a submatrix of $\Sigma(x)$ does not necessarily coincide with a submatrix of the matrix logarithm. Additionally, both the models of Pourahmadi [1999] and Chiu et al. [1996] involve a large number of parameters (specifically, $d \times p(p+1)/2$ assuming $x \in \Re^d$.) More recently, Hoff and Niu [2010] propose a covariance regression model in which $\Sigma(x) = A + Bxx'B'$ with A positive definite and B real. This model has interpretable parameters and may be equated with a latent factor model, leading to computational advantages. However, the model still has key limitations in (i) scaling to large p domains, and (ii) flexibility based on the parametric approach. Specifically, the model restricts the difference between $\Sigma(x)$ and the baseline matrix A to be rank 1. Higher rank models can be considered via extensions such as $\Sigma(x) = A + Bxx'B' + Cxx'C'$, but this dramatically increases the parameterization and requires definition of the maximal rank difference.

For volatility modeling where the covariate space is typically taken to be discrete time, heteroscedasticity has classically been captured via either variants of ARCH [Engle, 2002] or stochastic volatility models [Harvey et al., 1994]. The former directly specifies the volatility matrix as a linear combination of lagged volatilities and squared returns, which suffers from curse of dimensionality, and is typically limited to datasets with 5 or fewer dimensions. Alternatively, multivariate stochastic volatility models assume $\Sigma(t) = A\Gamma(t)A'$, with A real, $\Gamma(t) = \mathrm{diag}(\exp h_{1t}, \ldots, \exp h_{pt})$, and h_{it} independent autoregressive processes. See Chib et al. [2009] for a survey of such approaches. More recently, a number of papers have examined inducing covariance processes through variants of a Wishart process. Philipov and Glickman [2006a] take $\Sigma(t)^{-1} \mid \Sigma(t-1) \sim W(n, S_{t-1})$ with $S_{t-1} = 1/n(A^{1/2})(\Sigma(t-1)^{-1})^{\nu}(A^{1/2})'$†. Alternatively, a conditionally *non-central* Wishart distribution is induced on the precision matrix in Gouriéroux et al. [2009] by taking $\Sigma(t) = \sum_{k=1}^{p} x_{kt}x'_{kt}$, with each x_k independently a first order Gaussian autoregressive process. Key limitations of these types of Wishart processes are that posterior computations are extremely challenging, theory is lacking (e.g., simple descriptions of marginal distributions), and single parameters (e.g., n and ν) control the inter- and intra-temporal covariance relationships. Prado and West [2010] review alternative models of time-varying covariance matrices for dynamic linear models via discounting methods that maintain conjugacy. Central to all of the cited volatility models is the assumption of Markov dynamics, limiting the ability to capture long-range dependencies and often leading to spiky trajectories. Additionally, these methods assume a regular grid of observations that cannot easily accommodate missing values.

Within the spatial statistics community, the term *Wishart process* is typically used to specify a different formulation than those described herein for volatility modeling. Specifically, letting $\Sigma(s)$ denote the covariance of a p-dimensional observation at geographic location $s \in \mathcal{R}^2$, Gelfand et al. [2004] assume that $\Sigma(s) = A(s)A(s)' + \Sigma_0$ with Σ_0 diagonal and $T(s) = A(s)A(s)'$ following a *matric-variate Wishart process*. This Wishart process is such that $T(s)^{-1} = \Theta\xi(s)\xi(s)'\Theta'$ with $\xi_{\ell j} \sim \mathrm{GP}(0, c_j)$ independently for each ℓ, j and

†Extending to higher dimensions, Philipov and Glickman [2006b] apply this model to the covariance of a lower-dimensional latent factor in a standard latent factor model.

Θ typically taken to be diagonal. The induced distribution on $T(s)$ is then marginally inverse Wishart‡. Posterior computations in this model rely on Metropolis-Hastings proposals that do not scale well to dimensions p larger than 2-3 and cannot naturally accommodate missing data. In terms of spatio-temporal processes, Lopes et al. [2008] build upon a standard dynamic factor model to develop nonseparable and nonstationary space-time models. Specifically, the vector y_t of univariate observations y_{st} at spatial locations s is modeled as $y_t = \mu_t + \beta f_t + \epsilon_t$ with the components of the latent factors f_t independently evolving according to a first-order autoregressive process and columns of the factor loadings matrix β independently drawn from Markov random fields. Key assumptions of this formulation are that the observations evolve in discrete time on a regular grid, and that the dynamics of the spatio-temporal process are captured by independent random walks on the components of the latent factors.

In this paper, we present a Bayesian nonparametric approach to multivariate covariance regression that allows the covariance matrix to change flexibly with predictors and readily scales to high-dimensional datasets. The proposed modeling framework induces a prior on a collection of covariance matrices $\Sigma_{\mathcal{X}} = \{\Sigma(x), x \in \mathcal{X}\}$ through specification of a prior on a predictor-dependent latent factor model. In particular, the predictor-dependent loadings are characterized as a sparse combination of a collection of unknown dictionary functions (e.g, Gaussian process random functions). The induced covariance is then a regularized quadratic function of these dictionary elements. The proposed methodology has numerous advantages over previous approaches. By employing collections of continuous random functions, we allow for an irregular grid of observations. Similarly, we can easily cope with missing data within our framework without relying on imputing the missing values. Another fundamental property of the proposed methodology is the fact that our combined use of a shrinkage prior with a latent factor model enables us (in theory) to handle high-dimensional datasets (e.g., on the order of hundreds of dimensions) in the presence of a limited number of observations. Essential in being able to cope with such large datasets in practice is the fact that our computations are tractable, based on simple conjugate posterior updates. Finally, we are able to state theoretical properties of our proposed prior, such as large support.

The paper is organized as follows. In Section 2, we describe our proposed Bayesian nonparametric covariance regression model in addition to analyzing the theoretical properties of the model. Section 3 details the Gibbs sampling steps involved in our posterior computations. Finally, a number of simulation studies are examined in Section 4, with an application to the Google Trends flu dataset presented in Section 5.

2. Covariance Regression Priors

2.1. Notation and Motivation

Let $\Sigma(x)$ denote the $p \times p$ covariance matrix at "location" $x \in \mathcal{X}$. In general, x is an arbitrary, possibly multivariate predictor value. In dynamical modeling, x may simply represent a discrete time index (i.e., $\mathcal{X} = \{1, \dots, T\}$) or, in spatial modeling, a geographical location (i.e., $\mathcal{X} = \Re^2$). Another simple, tractable case is when x represents an ordered

‡More generally, Gelfand et al. [2004] develop a *spatial* coregionalization model such that $\text{cov}(y(s_1), y(s_2)) = \rho(s_1 - s_2)A(s_1)A(s_2)' + \Sigma_0$ (i.e., a model with spatial dependencies arising in both the covariance and *cross* covariance).

categorical predictor (i.e., $\mathcal{X} = \{1, \dots, N\}$). We seek a prior for $\Sigma_{\mathcal{X}} = \{\Sigma(x), x \in \mathcal{X}\}$, the collection of covariance matrices over the space of predictor values.

Letting $\Sigma_{\mathcal{X}} \sim \Pi_{\Sigma}$ our goal is to choose a prior Π_{Σ} for the collection of covariance matrices that has large support and leads to good performance in large p settings. By "good" we mean accurate estimation in small samples, taking advantage of shrinkage priors and efficient computation that scales well as p increases. We initially focus on the relatively simple setting in which

$$y_i \sim \mathcal{N}_p(\mu(x_i), \Sigma(x_i)) \tag{1}$$

independently for each i. Such a formulation could be extended to settings in which data are collected at repeated times for different subjects, as in multivariate longitudinal data analysis, by embedding the proposed model within a hierarchical framework. See Section 6 for a discussion.

2.2. Proposed Latent Factor Model

In large p settings, modeling a $p \times p$ covariance matrix $\Sigma(x)$ over an arbitrary predictor space $\mathcal{X}$ represents an enormous dimensional regression problem; we aim to reduce dimensionality for tractability in building a flexible nonparametric model for the predictor-dependent covariances. A popular approach for coping with such high dimensional (non-predictor-dependent) covariance matrices Σ in the presence of limited data is to assume that the covariance has a decomposition as $\Lambda\Lambda' + \Sigma_0$ where Λ is a $p \times k$ factor loadings matrix with $k << p$ and Σ_0 is a $p \times p$ diagonal matrix with non-negative entries. To build in predictor dependence, we assume a decomposition

$$\Sigma(x) = \Lambda(x)\Lambda(x)' + \Sigma_0, \tag{2}$$

where $\Lambda(x)$ is a $p \times k$ factor loadings matrix that is indexed by predictors x and where $\Sigma_0 = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$. Assuming initially for simplicity that $\mu(x) = 0$, such a decomposition is induced by marginalizing out a set of latent factors η_i from the following latent factor model:

$$\begin{aligned} y_i &= \Lambda(x_i)\eta_i + \epsilon_i \\ \eta_i &\sim \mathcal{N}_k(0, I_k), \quad \epsilon_i \sim \mathcal{N}_p(0, \Sigma_0). \end{aligned} \tag{3}$$

Here, $x_i = (x_{i1}, \dots, x_{iq})'$ is the predictor associated with the ith observation y_i.

Despite the dimensionality reduction introduced by the latent factor model of Eq. (3), modeling a $p \times k$ dimensional predictor-dependent factor loadings matrix $\Lambda(x)$ still represents a significant challenge for large p domains. To further reduce dimensionality, and following the strategy of building a flexible high-dimensional model from simple low-dimensional pieces, we assume that each element of $\Lambda(x)$ is a linear combination of a much smaller number of unknown dictionary functions $\xi_{\ell k} : \mathcal{X} \to \Re$. That is, we propose to let

$$\Lambda(x_i) = \Theta\xi(x_i), \tag{4}$$

where $\Theta \in \Re^{p \times L}$ is the matrix of coefficients relating the predictor-dependent factor loadings matrix to the set of dictionary functions comprising the $L \times k$ dimensional matrix $\xi(x)$. Typically, $k << p$ and $L << p$. Since we can write

$$[\Lambda(\cdot)]_{rs} = \sum_{\ell=1}^{L} \theta_{r\ell}\xi_{\ell s}(\cdot), \tag{5}$$

we see that the weighted sum of the sth column of dictionary functions $\xi_{\cdot s}(\cdot)$, with weights specified by the rth row of Θ, characterizes the impact of the sth latent factor on y_{ir}, the rth component of the response at predictor location x_i. By characterizing the elements of $\Lambda(x_i)$ as a linear combination of these flexible dictionary functions, we obtain a highly-flexible but computationally tractable formulation.

In marginalizing out the latent factors, we now obtain the following induced covariance structure

$$\operatorname{cov}(y_i \,|\, x_i = x) = \Sigma(x) = \Theta\xi(x)\xi(x)'\Theta' + \Sigma_0. \tag{6}$$

Note that the above decomposition of $\Sigma(x)$ is not unique and there are actually infinitely many such equivalent decompositions. For example, take $\Theta_1 = c\Theta$ and $\xi_1(\cdot) = (1/c)\xi(\cdot)$. Alternatively, consider $\xi_1(\cdot) = \xi(\cdot)P$ for any orthogonal matrix P or $\Theta_1 = [\Theta\ 0_{p\times d}]$ and $\xi_1 = [\xi;\, \xi_0]$ for any $d \times k$ matrix of dictionary functions ξ_0. One can also increase the dimension of the latent factors and take $\xi_1 = [\xi\ 0_{L\times d}]$. In standard (non-predictor-dependent) latent factor modeling, a common approach to ensure identifiability is to constrain the factor loadings matrix to be block lower triangular with strictly positive diagonal elements [Geweke and Zhou, 1996], though such a constraint induces order dependence among the responses [Aguilar and West, 2000, West, 2003, Lopes and West, 2004, Carvalho et al., 2008]. However, for tasks such as inference on the covariance matrix and prediction, identifiability of a unique decomposition is not necessary. Thus, we do not restrict ourselves to a unique decomposition of $\Sigma(x)$, allowing us to define priors with better computational properties.

Although we are not interested in identifying a unique decomposition of $\Sigma(x)$, we are interested in characterizing the class of covariance regressions $\Sigma(x)$ that can be decomposed as in Eq. (6). Lemma 2.1 states that for L and k sufficiently large, any covariance regression has such a decomposition. For $L, k \geq p$, let $\mathcal{X}_\xi$ denote the space of all $L \times k$ dimensional matrices of arbitrarily complex dictionary functions mapping from $\mathcal{X} \to \Re$, $\mathcal{X}_{\Sigma_0}$ be the space of all $p \times p$ diagonal matrices with non-negative entries, and $\mathcal{X}_\Theta$ be the space of all $p \times L$ dimensional matrices Θ such that $\Theta\Theta'$ has finite elements.

Lemma 2.1. *Given a symmetric positive semidefinite matrix $\Sigma(x) \succ 0$, $\forall x \in \mathcal{X}$, there exists $\{\xi(\cdot), \Theta, \Sigma_0\} \in \mathcal{X}_\xi \otimes \mathcal{X}_\Theta \otimes \mathcal{X}_{\Sigma_0}$ such that*

$$\Sigma(x) = \Theta\xi(x)\xi(x)'\Theta' + \Sigma_0, \quad \forall x \in \mathcal{X}. \tag{7}$$

Proof. *Assume without loss of generality that $\Sigma_0 = 0_{p\times p}$ and take $k, L \geq p$. Consider*

$$\Theta = [I_p\ 0_{p\times L-p}] \qquad \xi(x) = \begin{bmatrix} chol(\Sigma(x)) & 0_{p\times k-p} \\ 0_{L-p\times p} & 0_{L-p\times k-p} \end{bmatrix}. \tag{8}$$

Then, $\Sigma(x) = \Theta\xi(x)\xi(x)'\Theta'$, $\forall x \in \mathcal{X}$.

Now that we have established that there exist decompositions of $\Sigma(x)$ into the form specified by Equation (6), the question is whether we can specify a prior on the elements $\xi(\cdot)$, Θ, and Σ_0 that provides large support on such decompositions. This is explored in Section 2.3.

In order to generalize the model to also allow the mean $\mu(x)$ to vary flexibly with predictors, we can follow a nonparametric latent factor regression approach and let

$$\eta_i = \psi(x_i) + \nu_i, \quad \nu_i \sim \mathcal{N}_k(0, I_k), \tag{9}$$

where $\psi(x_i) = [\psi_1(x_i), \dots, \psi_k(x_i)]'$, and $\psi_j : \mathcal{X} \rightarrow \Re$ is an unknown function relating the predictors to the mean of the jth factor, for $j = 1, \dots, k$. These $\psi_j(\cdot)$ functions can be modeled in a related manner to the $\xi_{\ell k}(\cdot)$ functions described above. The induced mean of y_i conditionally on $x_i = x$ and marginalizing out the latent factors is then $\mu(x) = \Theta\xi(x)\psi(x)$. For simplicity, however, we focus our discussions on the case where $\mu(x) = 0$.

2.3. Prior Specification

Working within a Bayesian framework, we place independent priors on $\xi(\cdot)$, Θ, and Σ_0 in Eq. (6) to induce a prior on $\Sigma_{\mathcal{X}}$. Let Π_ξ, Π_Θ, and Π_{Σ_0} denote each of these independent priors, respectively. Recall that Π_Σ denotes the induced prior on $\Sigma_{\mathcal{X}}$.

Aiming to capture covariances that vary continuously over $\mathcal{X}$ combined with the goal of maintaining simple computations for inference, we specify the dictionary functions as

$$\xi_{\ell k}(\cdot) \sim \mathrm{GP}(0, c) \tag{10}$$

independently for all ℓ, k, with c a squared exponential correlation function having $c(\xi, \xi') = \exp(-\kappa||\xi - \xi'||_2^2)$.

To cope with the fact that the number of latent dictionary functions is a model choice we are required to make, we seek a prior Π_Θ that favors many values of Θ being close to zero so that we may choose L larger than the expected number of dictionary functions (also controlled by the latent factor dimension k). As proposed in Bhattacharya and Dunson [2010], we use the following shrinkage prior:

$$\begin{gathered} \theta_{j\ell} \mid \phi_{j\ell}, \tau_\ell \sim \mathcal{N}(0, \phi_{j\ell}^{-1}\tau_\ell^{-1}) \quad \phi_{j\ell} \sim \mathrm{Ga}(3/2, 3/2) \\ \delta_1 \sim \mathrm{Ga}(a_1, 1), \quad \delta_h \sim \mathrm{Ga}(a_2, 1),\ h \geq 2, \quad \tau_\ell = \prod_{h=1}^{\ell} \delta_h. \end{gathered} \tag{11}$$

Choosing $a_2 > 1$ implies that δ_h is greater than 1 in expectation so that τ_ℓ tends stochastically towards infinity as ℓ goes to infinity, thus shrinking the elements $\theta_{j\ell}$ toward zero increasingly as ℓ grows. The $\phi_{j\ell}$ precision parameters allow for flexibility in how the elements of Θ are shrunk towards zero by incorporating local shrinkage specific to each element of Θ, while τ_ℓ provides a global column-wise shrinkage factor.

Finally, we specify Π_{Σ_0} via the usual inverse gamma priors on the diagonal elements of Σ_0. That is,

$$\sigma_j^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma) \tag{12}$$

independently for each $j = 1, \dots, p$.

2.4. Theoretical Properties

In this section, we explore the theoretical properties of the proposed Bayesian nonparametric covariance regression model. In particular, we focus on the support of the induced prior

Π_Σ based on the priors Π_ξ, Π_Θ, and Π_{Σ_0} defined in Section 2.3. Large support implies that the prior can generate covariance regressions that are arbitrarily close to any function $\{\Sigma(x), x \in \mathcal{X}\}$ in a large class. Such a support property is the defining feature of a Bayesian nonparametric approach and cannot simply be assumed. Often, seemingly flexible models can have quite restricted support due to hidden constraints in the model and not to real prior knowledge that certain values are implausible. Although we have chosen a specific form for a shrinkage prior Π_Θ, we aim to make our statement of prior support as general as possible and thus simply assume that Π_Θ satisfies a set of two conditions given by Assumption 2.1 and Assumption 2.2. In Lemma 2.2, we show that the Π_Θ specified in Eq. (11) satisfies these assumptions. The proofs associated with the theoretical statements made in this section can be found in the Appendix.

ASSUMPTION 2.1. *Π_Θ is such that $\sum_\ell E[|\theta_{j\ell}|] < \infty$. This property ensures that the prior on the rows of Θ shrinks the elements towards zero fast enough as $\ell \to \infty$.*

ASSUMPTION 2.2. *Π_Θ is such that $\Pi_\Theta\,(rank(\Theta) = p) > 0$. That is, there is positive prior probability of Θ being full rank.*

The following theorem shows that, for $k \geq p$ and as $L \to \infty$, the induced prior Π_Σ places positive probability on the space of all covariance functions $\Sigma^*(x)$ that are continuous on $\mathcal{X}$.

THEOREM 2.1. *Let Π_Σ denote the induced prior on $\{\Sigma(x), x \in \mathcal{X}\}$ based on the specified prior $\Pi_\xi \otimes \Pi_\Theta \otimes \Pi_{\Sigma_0}$ on $\mathcal{X}_\xi \otimes \mathcal{X}_\Theta \otimes \mathcal{X}_{\Sigma_0}$. Assuming $\mathcal{X}$ compact, for all continuous $\Sigma^*(x)$ and for all $\epsilon > 0$,*

$$\Pi_\Sigma \left(\sup_{x \in \mathcal{X}} ||\Sigma(x) - \Sigma^*(x)||_2 < \epsilon \right) > 0. \tag{13}$$

Intuitively, the support on continuous covariance functions $\Sigma^*(x)$ arises from the continuity of the Gaussian process dictionary functions. However, since we are mixing over infinitely many such dictionary functions, we need the mixing weights specified by Θ to tend towards zero, and to do so "fast enough"—this is where Assumption 2.1 becomes important. See Theorem 2.2. The proof of Theorem 2.1 relies on the large support of Π_Σ at any point $x_0 \in \mathcal{X}$. Since each $\xi_{\ell k}(x_0)$ is independently Gaussian distributed (based on properties of the Gaussian process prior), $\xi(x_0)\xi(x_0)'$ is Wishart distributed. Conditioned on Θ, $\Theta\xi(x_0)\xi(x_0)'\Theta'$ is also Wishart distributed. More generally, for fixed Θ, $\Theta\xi(x)\xi(x)'\Theta'$ follows the matric-variate Wishart process of Gelfand et al. [2004]. Combining the large support of the Wishart distribution with that of the gamma distribution on the inverse elements of Σ_0 provides the desired large support of the induced prior Π_Σ at each predictor location x_0.

THEOREM 2.2. *For every finite k and $L \to \infty$ (or L finite), $\Lambda(\cdot) = \Theta\xi(\cdot)$ is almost surely continuous on $\mathcal{X}$.*

Lemma 2.2 specifies the conditions under which the prior Π_Θ specified in Eq. (11) satisfies Assumption 2.1, which provides a sufficient condition used in the proof of prior support.

LEMMA 2.2. *Based on the prior specified in Eq.* (11) *and choosing* $a_2 > 2$, *Assumption* 2.1 *is satisfied. That is,* $\sum_\ell E[|\theta_{j\ell}|] < \infty$.

It is also of interest to analyze the moments associated with the proposed prior. As detailed in the Appendix, the first moment can be derived based on the implied inverse gamma prior on the σ_j^2 combined with the fact that $\Theta\xi(x)\xi(x)'\Theta'$ is marginally Wishart distributed at every location x, with the prior on Θ specified in Equation (11).

LEMMA 2.3. *Let* μ_σ *denote the mean of* σ_j^2, $j = 1, \ldots, p$. *Then,*

$$E[\Sigma(x)] = diag\left(k\sum_\ell \phi_{1\ell}^{-1}\tau_\ell^{-1} + \mu_\sigma, \ldots, k\sum_\ell \phi_{p\ell}^{-1}\tau_\ell^{-1} + \mu_\sigma\right). \tag{14}$$

Since our goal is to develop a covariance regression model, it is natural to consider the correlation induced between an element of the covariance matrix at different predictor locations x and x'.

LEMMA 2.4. *Let* σ_σ^2 *denote the variance of* σ_j^2, $j = 1, \ldots, p$. *Then,*

$$cov(\Sigma_{ij}(x), \Sigma_{ij}(x')) = \begin{cases} kc(x,x')\left(5\sum_\ell \phi_{i\ell}^{-2}\tau_\ell^{-2} + (\sum_\ell \phi_{i\ell}^{-1}\tau_\ell^{-1})^2\right) + \sigma_\sigma^2 & i = j, \\ kc(x,x')\left(\sum_\ell \phi_{i\ell}^{-1}\phi_{j\ell}^{-1}\tau_\ell^{-2} + \sum_\ell \phi_{i\ell}^{-1}\tau_\ell^{-1}\sum_{\ell'}\phi_{j\ell'}^{-1}\tau_{\ell'}^{-1}\right) & i \neq j. \end{cases} \tag{15}$$

For any two elements $\Sigma_{ij}(x)$ *and* $\Sigma_{uv}(x')$ *with* $i \neq u$ *or* $j \neq v$,

$$cov(\Sigma_{ij}(x), \Sigma_{uv}(x')) = 0. \tag{16}$$

We can thus conclude that in the limit as the distance between the predictors tends towards infinity, the correlation decays at a rate defined by the Gaussian process kernel $c(x, x')$ with a limit:

$$\lim_{||x-x'||\to\infty} \text{cov}(\Sigma_{ij}(x), \Sigma_{uv}(x')) = \begin{cases} \sigma_\sigma^2 & i = j = u = v, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

It is perhaps counterintuitive that the correlation between $\Sigma_{ii}(x)$ and $\Sigma_{ii}(x')$ does not go to zero as the distance between the predictors x and x' tends to infinity. However, although the correlation between $\xi(x)$ and $\xi(x')$ goes to zero, the diagonal matrix Σ_0 does not depend on x or x' and thus retains the correlation between the diagonal elements of $\Sigma(x)$ and $\Sigma(x')$.

Equation (15) implies that the autocorrelation $ACF(x) = \text{corr}(\Sigma_{ij}(0), \Sigma_{ij}(x))$ is simply specified by $c(0, x)$. When we choose a Gaussian process kernel $c(x, x') = \exp(-\kappa||x - x'||_2^2)$, we have

$$ACF(x) = \exp(-\kappa||x||_2^2). \tag{18}$$

Thus, we see that the length-scale parameter κ directly determines the shape of the autocorrelation function.

Finally, one can analyze the stationarity properties of the proposed covariance regression prior.

LEMMA 2.5. *Our proposed covariance regression model defines a first-order stationary process in that $\Pi_\Sigma(\Sigma(x)) = \Pi_\Sigma(\Sigma(x'))$, $\forall x, x' \in \mathcal{X}$. Furthermore, the process is wide sense stationary: cov$(\Sigma_{ij}(x), \Sigma_{uv}(x'))$ solely depends upon $||x - x'||$.*

PROOF. *The first-order stationarity follows immediately from the stationarity of the Gaussian process dictionary elements $\xi_{\ell k}(\cdot)$ and recalling that $\Sigma(x) = \Theta\xi(x)\xi(x)'\Theta' + \Sigma_0$. Assuming a Gaussian process kernel $c(x, x')$ that solely depends upon the distance between x and x' (as in Section 2.3), Equations (15)- (16) imply that the defined process is wide sense stationary.*

3. Posterior Computation

3.1. Gibbs Sampling with a Fixed Truncation Level

Based on a fixed truncation level L^* and a latent factor dimension k^*, we propose a Gibbs sampler for posterior computation. The derivation of Step 1 is provided in the Appendix.

Step 1 Update each dictionary function $\xi_{\ell m}(\cdot)$ from the conditional posterior given $\{y_i\}$, Θ, $\{\eta_i\}$, Σ_0. We can rewrite the observation model for the jth component of the ith response as

$$y_{ij} = \sum_{m=1}^{k^*} \eta_{im} \sum_{\ell=1}^{L^*} \theta_{j\ell}\xi_{\ell m}(x_i) + \epsilon_{ij}. \tag{19}$$

Conditioning on $\xi(\cdot)^{-\ell m} = \{\xi_{rs}(\cdot), r \neq \ell, s \neq m\}$, our Gaussian process prior on the dictionary functions implies the following conditional posterior

$$\begin{bmatrix} \xi_{\ell m}(x_1) \\ \xi_{\ell m}(x_2) \\ \vdots \\ \xi_{\ell m}(x_n) \end{bmatrix} \mid \{y_i\}, \Theta, \eta, \xi(\cdot)^{-\ell m}, \Sigma_0 \sim \mathcal{N}_n\left(\tilde{\Sigma}_\xi \begin{bmatrix} \eta_{1m}\sum_{j=1}^p \theta_{j\ell}\sigma_j^{-2}\tilde{y}_{1j} \\ \vdots \\ \eta_{nm}\sum_{j=1}^p \theta_{j\ell}\sigma_j^{-2}\tilde{y}_{nj} \end{bmatrix}, \tilde{\Sigma}_\xi\right), \tag{20}$$

where $\tilde{y}_{ij} = y_{ij} - \sum_{(r,s)\neq(\ell,m)} \theta_{jr}\xi_{rs}(x_i)$ and, taking K to be the Gaussian process covariance matrix with $K_{ij} = c(x_i, x_j)$,

$$\tilde{\Sigma}_\xi^{-1} = K^{-1} + \text{diag}\left(\eta_{1m}^2 \sum_{j=1}^p \theta_{j\ell}^2\sigma_j^{-2}, \ldots, \eta_{nm}^2 \sum_{j=1}^p \theta_{j\ell}^2\sigma_j^{-2}\right). \tag{21}$$

Step 2 Next we sample each latent factor η_i given y_i, $\xi(\cdot)$, Θ, Σ_0. Recalling Eq. (3) and the fact that $\eta_i \sim \mathcal{N}_{k^*}(0, I_{k^*})$,

$$\begin{aligned} &\eta_i \mid y_i, \Theta, \xi(x_i), \Sigma_0 \\ &\quad \sim \mathcal{N}_{k^*}\left(\left(I + \xi(x_i)'\Theta'\Sigma_0^{-1}\Theta\xi(x_i)\right)^{-1}\xi(x_i)'\Theta'\Sigma_0^{-1}y_i, \left(I + \xi(x_i)'\Theta'\Sigma_0^{-1}\Theta\xi(x_i)\right)^{-1}\right). \end{aligned} \tag{22}$$

Step 3 Let $\theta_{j\cdot} = \begin{bmatrix}\theta_{j1} & \dots & \theta_{jL^*}\end{bmatrix}$. Recalling the $\mathrm{Ga}(a_\sigma, b_\sigma)$ prior on each precision parameter σ_j^{-2} associated with the diagonal noise covariance matrix Σ_0, standard conjugate posterior analysis yields the posterior

$$\sigma_j^{-2} \mid \{y_i\}, \Theta, \eta, \xi(\cdot) \sim \mathrm{Ga}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(y_{ij} - \theta_{j\cdot}\xi(x_i)\eta_i)^2\right). \tag{23}$$

Step 4 Conditioned on the hyperparameters ϕ and τ, the Gaussian prior on the elements of Θ specified in Eq. (11) combined with the likelihood defined by Eq. (3) imply the following posterior for each row of Θ:

$$\theta_{j\cdot} \mid \{y_i\}, \eta, \xi(\cdot), \phi, \tau \sim \mathcal{N}_{L^*}\left(\tilde{\Sigma}_\theta \tilde{\eta}' \sigma_j^{-2} \begin{bmatrix} y_{1j} \\ \vdots \\ y_{nj}\end{bmatrix}, \tilde{\Sigma}_\theta\right), \tag{24}$$

where $\tilde{\eta}' = \begin{bmatrix}\xi(x_1)\eta_1 & \xi(x_2)\eta_2 & \dots & \xi(x_n)\eta_n\end{bmatrix}$ and

$$\tilde{\Sigma}_\theta^{-1} = \sigma_j^{-2}\tilde{\eta}'\tilde{\eta} + \mathrm{diag}(\phi_{j1}\tau_1, \dots, \phi_{jL^*}\tau_{L^*}). \tag{25}$$

Step 5 Examining Eq. (11) and using standard conjugate analysis results in the following posterior for each local shrinkage hyperparameter $\phi_{j\ell}$ given $\theta_{j\ell}$ and τ_ℓ:

$$\phi_{j\ell} \mid \theta_{j\ell}, \tau_\ell \sim \mathrm{Ga}\left(2, \frac{3 + \tau_\ell \theta_{j\ell}^2}{2}\right). \tag{26}$$

Step 6 As in Bhattacharya and Dunson [2010], the global shrinkage hyperparameters are updated as

$$\begin{aligned}\delta_1 \mid \Theta, \tau^{(-1)} &\sim \mathrm{Ga}\left(a_1 + \frac{pL^*}{2}, 1 + \frac{1}{2}\sum_{\ell=1}^{L^*}\tau_\ell^{(-1)}\sum_{j=1}^{p}\phi_{j\ell}\theta_{j\ell}^2\right) \\ \delta_h \mid \Theta, \tau^{(-h)} &\sim \mathrm{Ga}\left(a_2 + \frac{p(L^* - h + 1)}{2}, 1 + \frac{1}{2}\sum_{\ell=1}^{L^*}\tau_\ell^{(-h)}\sum_{j=1}^{p}\phi_{j\ell}\theta_{j\ell}^2\right),\end{aligned} \tag{27}$$

where $\tau_\ell^{(-h)} = \prod_{t=1, t\neq h}^{\ell}\delta_t$ for $h = 1, \dots, p$.

3.2. Incorporating nonparametric mean $\mu(x)$

If one wishes to incorporate a latent factor regression model such as in Eq. (9) to induce a predictor-dependent mean $\mu(x)$, the MCMC sampling is modified as follows. Steps 1, 3, 4, 5, and 6 are exactly as before. Now, however, the sampling of η_i of Step 2 is replaced by a block sampling of $\psi(x_i)$ and ν_i. Specifically, let $\Omega_i = \Theta\xi(x_i)$. We can rewrite the observation model as $y_i = \Omega_i\psi(x_i) + \Omega_i\nu_i + \epsilon_i$. Marginalizing out ν_i, $y_i = \Omega_i\psi(x_i) + \omega_i$

with $\omega_i \sim \mathcal{N}(0, \tilde{\Sigma}_i \triangleq \Omega_i\Omega_i' + \Sigma_0)$. Assuming nonparametric mean vector components $\psi_\ell(\cdot) \sim \mathrm{GP}(0, c)$, the posterior of $\psi_\ell(\cdot)$ follows analogously to that of $\xi(\cdot)$ resulting in

$$\begin{bmatrix} \psi_\ell(x_1) \\ \psi_\ell(x_2) \\ \vdots \\ \psi_\ell(x_n) \end{bmatrix} \mid \{y_i\}, \psi(\cdot)^{-\ell}, \Theta, \eta, \xi(\cdot), \Sigma_0 \sim \mathcal{N}_n\left(\tilde{\Sigma}_\psi \begin{bmatrix} [\Omega_1]'_{\cdot\ell}\tilde{\Sigma}_1^{-1}\tilde{y}_1^{-\ell} \\ \vdots \\ [\Omega_n]'_{\cdot\ell}\tilde{\Sigma}_n^{-1}\tilde{y}_n^{-\ell} \end{bmatrix}, \tilde{\Sigma}_\psi\right), \tag{28}$$

where $\tilde{y}_i^{-\ell} = y_i - \sum_{(r\neq\ell)}[\Omega_i]_{\cdot r}\psi_r(x_i)$. Once again taking K to be the Gaussian process covariance matrix,

$$\tilde{\Sigma}_\xi^{-1} = K^{-1} + \mathrm{diag}\left([\Omega_1]'_{\cdot\ell}\tilde{\Sigma}_1^{-1}[\Omega_1]_{\cdot\ell}, \ldots, [\Omega_n]'_{\cdot\ell}\tilde{\Sigma}_n^{-1}[\Omega_n]_{\cdot\ell}\right). \tag{29}$$

Conditioned on $\psi(x_i)$, we consider $\tilde{y}_i = y_i - \Omega_i\psi(x_i) = \Omega_i\nu_i + \epsilon_i$. Then, using the fact that $\nu_i \sim \mathcal{N}(0, I_{k*})$,

$$\begin{aligned} &\nu_i \mid \tilde{y}_i, \psi(x_i), \Theta, \xi(x_i), \Sigma_0 \\ &\quad \sim \mathcal{N}_{k^*}\left(\left(I + \xi(x_i)'\Theta'\Sigma_0^{-1}\Theta\xi(x_i)\right)^{-1}\xi(x_i)'\Theta'\Sigma_0^{-1}\tilde{y}_i, \left(I + \xi(x_i)'\Theta'\Sigma_0^{-1}\Theta\xi(x_i)\right)^{-1}\right). \end{aligned} \tag{30}$$

3.3. Hyperparameter Sampling

One can also consider sampling the Gaussian process length-scale hyperparameter κ. Due to the linear-Gaussianity of the proposed covariance regression model, we can analytically marginalize the latent Gaussian process random functions in considering the posterior of κ. Once again taking $\mu(x) = 0$ for simplicity, our posterior is based on marginalizing the Gaussian process random vectors $\xi_{\ell m} = [\xi_{\ell m}(x_1) \ \ldots \ \xi_{\ell m}(x_n)]'$. Noting that

$$\begin{bmatrix} y_1' & y_2' & \ldots & y_n' \end{bmatrix}' = \sum_{\ell m}\left[\mathrm{diag}(\eta_{\cdot m}) \otimes \theta_{\cdot \ell}\right]\xi_{\ell m} + \begin{bmatrix} \epsilon_1' & \epsilon_2' & \ldots & \epsilon_n' \end{bmatrix}', \tag{31}$$

and letting K_κ denote the Gaussian process covariance matrix based on a length-scale κ,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \mid \kappa, \Theta, \eta, \Sigma_0 \sim \mathcal{N}_{np}\left(\sum_{\ell,m}\left[\mathrm{diag}(\eta_{\cdot m}) \otimes \theta_{\cdot\ell}\right]K_\kappa\left[\mathrm{diag}(\eta_{\cdot m}) \otimes \theta_{\cdot\ell}\right]' + \begin{bmatrix} \Sigma_0 & & & \\ & \Sigma_0 & & \\ & & \ddots & \\ & & & \Sigma_0 \end{bmatrix}\right). \tag{32}$$

We can then Gibbs sample κ based on a fixed grid and prior $p(\kappa)$ on this grid. Note, however, that computation of the likelihood specified in Eq. (32) requires evaluation of an np-dimensional Gaussian for each value κ specified in the grid. For large p scenarios, or when there are many observations y_i, this may be computationally infeasible. In such cases, a naive alternative is to iterate between sampling $\xi(\cdot)$ given K_κ and K_κ given $\xi(\cdot)$. However, this can lead to extremely slow mixing. Alternatively, one can consider employing the recent Gaussian process hyperparameter slice sampler of Adams and Murray [2011].

In general, because of the quadratic mixing over Gaussian process dictionary elements, our model is relatively robust to the choice of the length-scale parameter and the computational burden imposed by sampling κ is typically unwarranted. Instead, one can pre-select

Table 1. Computations required at each Gibbs sampling step.

Gibbs Update	Computation
Step 1	$L^* \times k^*$ draws from an n dimensional Gaussian
Step 2	n draws from a k^* dimensional Gaussian
Step 3	p draws from a gamma distribution
Step 4	p draws from an L^* dimensional Gaussian
Step 5	$p \times L^*$ draws from a gamma distribution
Step 6	L^* draws from a gamma distribution

a value for κ using a data-driven heuristic, which leads to a quasi-empirical Bayes approach. Recalling Equation (18), we have

$$-\log(ACF(x)) = \kappa ||x||_2^2. \tag{33}$$

Thus, if one can devise a procedure for estimating the autocorrelation function from the data, one can set κ accordingly. We propose the following.

1. For a set of evenly spaced knots $x_k \in \mathcal{X}$, compute the sample covariance $\hat{\Sigma}(x_k)$ from a local bin of data $y_{k-k_0:k+k_0}$ with $k_0 > p/2$.
2. Compute the Cholesky decomposition $C(x_k) = chol(\hat{\Sigma}(x_k))$.
3. Fit a spline through the elements of the computed $C(x_k)$. Denote the spline fit of the Cholesky by $\tilde{C}(x)$ for each $x \in \mathcal{X}$
4. For $i = 1, \dots, n$, compute a point-by-point estimate of $\Sigma(x_i)$ from the splines: $\Sigma(x_i) = \tilde{C}(x_i)\tilde{C}(x_i)'$.
5. Compute the autocorrelation function of each element $\Sigma_{ij}(x)$ of this kernel-estimated $\Sigma(x)$.
6. According to Equation (33), choose κ to best fit the most correlated $\Sigma_{ij}(x)$ (since less correlated components can be captured via weightings of dictionary elements with stronger correlation.)

3.4. Computational Considerations

In choosing a truncation level L^* and latent factor dimension k^*, there are a number of computational considerations. The Gibbs sampler outlined in Section 3.1 involves a large number of simulations from Gaussian distributions, each of which requires the inversion of an m-dimensional covariance matrix, with m the dimension of the Gaussian. For large m, this represents a large computational burden as the operation is, in general, $O(m^3)$. The computations required at each stage of the Gibbs sampler are summarized in Table 1. From this table we see that depending on the number of observations n and the dimension of these observations p, various combinations of L^* and k^* lead to more or less efficient computations.

In Bhattacharya and Dunson [2010], a method for adaptively choosing the number of factors in a non-predictor dependent latent factor model was proposed. One could directly apply such a methodology for adaptively selecting L^*. To handle the choice of k^*, one could consider an augmented formulation in which

$$\Lambda(x) = \Theta\xi(x)\Gamma, \tag{34}$$

where $\Gamma = \mathrm{diag}(\gamma_1, \dots, \gamma_k)$ is a diagonal matrix of parameters that shrink the columns of $\xi(x)$ towards zero. One can take these shrinkage parameters to be distributed as

$$\begin{aligned} \gamma_i \sim \mathcal{N}(0, \omega_i^{-1}), \quad \omega_i = \prod_{h=1}^{i} \zeta_h \\ \zeta_1 \sim \mathrm{Ga}(a_3, 1), \quad \zeta_h \sim \mathrm{Ga}(a_4, 1) \quad h = 2, \dots, k. \end{aligned} \tag{35}$$

For $a_4 > 1$, such a model shrinks the γ_i values towards zero for large indices i just as in the shrinkage prior on Θ. The γ_i close to zero provide insight into redundant latent factor dimensions. Computations in this augmented model are a straightforward extension of the Gibbs sampler presented in Section 3.1. Based on the inferred values of the latent γ_i parameters, one can design an adaptive strategy similar to that for L^*.

Note that in Step 1, the n-dimensional inverse covariance matrix $\tilde{\Sigma}_\xi^{-1}$ which needs to be inverted in order to sample $\xi_{\ell m}$ is a composition of a diagonal matrix and an inverse covariance matrix K^{-1} that has entries that tend towards zero as $||x_i - x_j||_2$ becomes large (i.e., for distant pairs of predictors.) That is, K^{-1} (and thus $\tilde{\Sigma}_\xi^{-1}$) is nearly band-limited, with a bandwidth dependent upon the Gaussian process parameter κ. Inverting a given $n \times n$ band-limited matrix with bandwidth $d << n$ can be efficiently computed in $O(m^2 d)$ [Kavcic and Moura, 2000] (versus the naive $O(m^3)$). Issues related to tapering the elements of K^{-1} to zero while maintaining positive-semidefiniteness are discussed in Zhang and Du [2008].

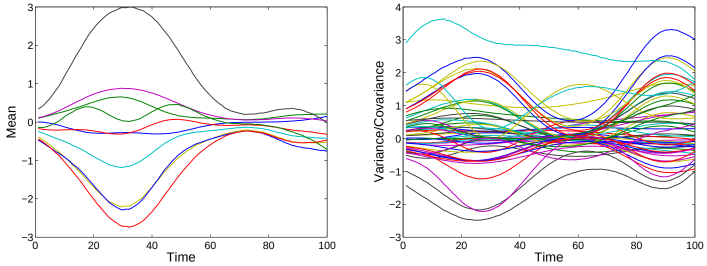
4. Simulation Example

In the following simulation examples, we aim to analyze the performance of the proposed Bayesian nonparametric covariance regression model relative to competing alternatives in terms of both covariance estimation and predictive performance. We initially consider the case in which $\Sigma(x)$ is generated from the assumed nonparametric Bayes model in Section 4.1 and 4.2, while in Section 4.3 we simulate from a parametric model and compare to a Wishart matrix discounting method [Prado and West, 2010] over a set of replicates.

4.1. Estimation Performance

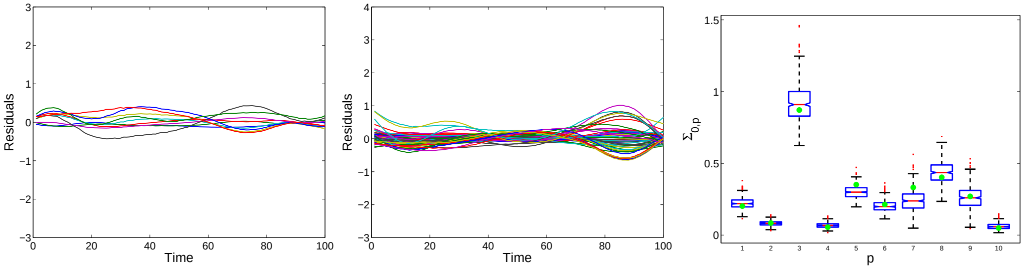
We simulated a dataset from the model as follows. The set of predictors is a discrete set $\mathcal{X} = \{1, \dots, 100\}$, with a 10-dimensional observation y_i generated for each $x_i \in \mathcal{X}$. The generating mechanism was based on weightings of a latent 5×4 dimensional matrix $\xi(\cdot)$ of Gaussian process dictionary functions (i.e, $L = 5$, $k = 4$), with length-scale $\kappa = 10$ and an additional nugget effect adding $1e^{-5} I_n$ to K. Here, we first scale the predictor space to $(0, 1]$. The additional latent mean dictionary elements $\psi(\cdot)$ were similarly distributed. The weights Θ were simulated as specified in Eq. (11) choosing $a_1 = a_2 = 10$. The precision parameters σ_j^{-2} were each drawn independently from a $\mathrm{Ga}(1, 0.1)$ distribution with mean 10. Figure 1 displays the resulting values of the elements of $\mu(x)$ and $\Sigma(x)$.

For inference, we set the hyperparameters as follows. We use truncation levels $k^* = L^* = 10$, which we found to be sufficiently large from the fact that the last few columns of the posterior samples of Θ were consistently shrunk close to 0. We set $a_1 = a_2 = 2$ and placed a $\mathrm{Ga}(1, 0.1)$ prior on the precision parameters σ_j^{-2}. The length-scale parameter κ was set from the data according to the heuristic described in Section 3.3 using 20 knots evenly spaced in $\mathcal{X} = \{1, \dots, 100\}$, and was determined to be 10 (after rounding).



(a) (b)

Fig. 1. Plot of each component of the (a) true mean vector $\mu(x)$ and (b) true covariance matrix $\Sigma(x)$ over the predictor space $\mathcal{X} = \{1, \ldots, 100\}$, taken here to represent a time index.



(a) (b) (c)

Fig. 2. Residuals between each component of the true and posterior mean of (a) the mean $\mu(x)$, and (b) covariance $\Sigma(x)$. The scaling of the axes matches that of Figure 1. (c) Box plot of posterior samples of the noise covariance terms σ_j^2 for $j = 1, \ldots, p$ compared to the true value (green).

Although experimentally we found that our sampler was insensitive to initialization in lower-dimensional examples such as the one analyzed here, we use the following more intricate initialization for consistency with later experiments on larger datasets in which mixing becomes more problematic. The predictor-independent parameters Θ and Σ_0 are sampled from their respective priors (first sampling the shrinkage parameters $\phi_{j\ell}$ and δ_h from their priors). The variables η_i and $\xi(x_i)$ are set via a data-driven initialization scheme in which an estimate of $\Sigma(x_i)$ for $i = 1, \ldots, n$ is formed using Steps 1-4 of Section 3.3. Then, $\Theta\xi(x_i)$ is taken to be a k^*-dimensional low-rank approximation to the Cholesky of the estimates of $\Sigma(x_i)$. The latent factors η_i are sampled from the posterior given in Equation (22) using this data-driven estimate of $\Theta\xi(x_i)$. Similarly, the $\xi(x_i)$ are initially taken to be spline fits of the pseudo-inverse of the low-rank Cholesky at the knot locations and the sampled Θ. We then iterate a couple of times between sampling: (i) $\xi(\cdot)$ given $\{y_i\}$, Θ, Σ_0, and the data-driven estimates of η, $\xi(\cdot)$; (ii) Θ given $\{y_i\}$, Σ_0, η, and the sampled $\xi(\cdot)$; (iii) Σ_0 given $\{y_i\}$, Θ, η, and $\xi(\cdot)$; and (iv) determining a new data-driven approximation to $\xi(\cdot)$ based on the newly sampled Θ. Results indistinguishable from those presented here were achieved (after a short burn-in period) by simply initializing each of Θ, $\xi(\cdot)$, Σ_0, η_i, and the shrinkage parameters $\phi_{j\ell}$ and δ_h from their respective priors.

We ran 10,000 Gibbs iterations and discarded the first 5,000 iterations. We then thinned the chain every 10 samples. The residuals between the true and posterior mean over all components are displayed in Figure 2(a) and (b). Figure 2(c) compares the posterior samples of the elements σ_j^2 of the noise covariance Σ_0 to the true values. Finally, in Figure 3 we

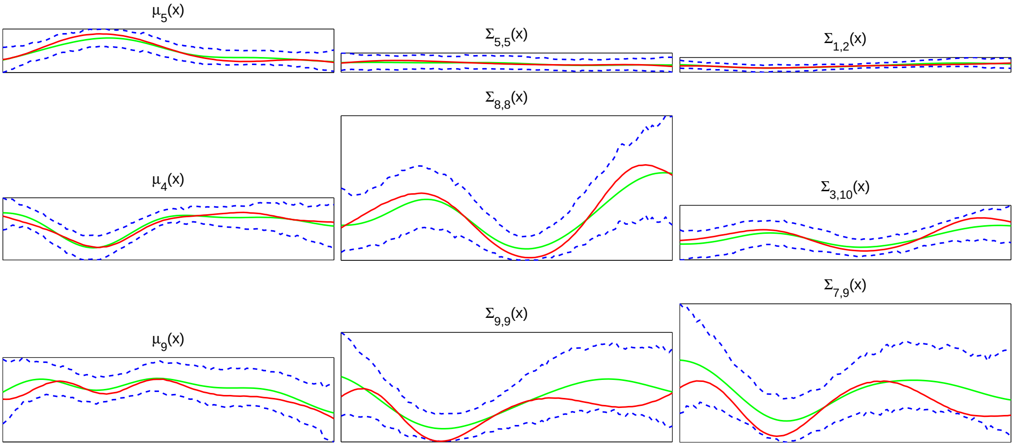


Fig. 3. Plots of truth (red) and posterior mean (green) for select components of the mean $\mu_p(x)$ (*left*), variances $\Sigma_{pp}(x)$ (*middle*), and covariances $\Sigma_{pq}(x)$ (*right*). The point-wise 95% highest posterior density intervals are shown in blue. The top row represents the component with the lowest $L2$ error between the truth and posterior mean. Likewise, the middle row represents median $L2$ error and the bottom row the worst $L2$ error. The size of the box indicates the relative magnitudes of each component.

display a select set of plots of the true and posterior mean of components of $\mu(x)$ and $\Sigma(x)$, along with the 95% highest posterior density intervals computed at each predictor value $x = 1, \ldots, 100$.

From the plots of Figures 2 and 3, we see that we are clearly able to capture heteroscedasticity in combination with a nonparametric mean regression. The true values of the mean and covariance components are (even in the worst case) contained within the 95% highest posterior density intervals, with these intervals typically small such that the overall interval bands are representative of the shape of the given component being modeled.

4.2. Predictive Performance

Capturing heteroscedasticity can significantly improve estimates of the predictive distribution of new observations or missing data. To explore these gains within our proposed Bayesian nonparametric covariance regression framework, we compare against two possible homoscedastic formulations that each assume $y \sim \mathcal{N}(\mu(x), \Sigma)$. The first is a standard Gaussian process mean regression model with each element of $\mu(x)$ an independent draw from $\text{GP}(0, c)$. The second builds on our proposed regularized latent factor regression model and takes $\mu(x) = \Theta\xi(x)\psi(x)$, with $\{\Theta, \xi(x), \psi(x)\}$ as defined in the heteroscedastic case. However, instead of having a predictor-dependent covariance $\Sigma(x) = \Theta\xi(x)\xi(x)'\Theta' + \Sigma_0$, the homoscedastic model assumes that Σ is an arbitrary covariance matrix constant over predictors. By comparing to this latter homoscedastic model, we can directly analyze the benefits of our heteroscedastic model since both share exactly the same mean regression formulation. For each of the homoscedastic models, we place an inverse Wishart prior on the covariance Σ.

We analyze the same simulation dataset as described in Section 4.1, but randomly remove approximately 5% of the observations. Specifically, independently for each element y_{ij} (i.e., the jth response component at predictor x_i) we decide whether to remove the observation

Table 2. Average Kullback-Leibler divergence $D_{KL}(P_{i,m}||Q_i)$, $i = 1, \dots, 100$, where $P_{i,m}$ and Q_i are the predictive distributions of all missing elements y_{ij} given the observed elements of y_i based on the mth posterior sample of and true parameters $\mu(x_i)$ and $\Sigma(x_i)$, respectively. We compare the predictive performance for two homoscedastic models to our covariance regression framework.

Model	*Average Posterior Predictive KL Divergence*
Homoscedastic Mean Regression	0.3409
Homoscedastic Latent Factor Mean Regression	0.2909
Heteroscedastic Mean Regression	**0.1216**

based on a Bernoulli(p_i) draw. We chose p_i to be a function of the matrix norm of the true covariance at x_i to slightly bias towards removing values from predictor regions with a tighter distribution. This procedure resulted in removing 48 of the 1000 response elements.

Table 2 compares the average Kullback-Leibler divergence $D_{KL}(P_{i,m}||Q_i)$, $i = 1, \dots, 100$, for the following definitions of $P_{i,m}$ and Q_i. The distribution Q_i is the predictive distribution of all missing elements y_{ij} given the observed elements of y_i under the true parameters $\mu(x_i)$ and $\Sigma(x_i)$. Likewise, $P_{i,m}$ is taken to be the predictive distribution based on the mth posterior sample of $\mu(x_i)$ and $\Sigma(x_i)$. In this scenario, the missing observations y_{ij} are imputed as an additional step in the MCMC computations§. The results, once again based on 10,000 Gibbs iterations and discarding the first 5,000 samples, clearly indicate that our Bayesian nonparametric covariance regression model provides more accurate predictive distributions. We additionally note that using a regularized latent factor approach to mean regression improves on the naive homoscedastic model in high dimensional datasets in the presence of limited data. Not depicted in this paper due to space constraints is the fact that the proposed covariance regression model also leads to improved estimates of the mean $\mu(x)$ in addition to capturing heteroscedasticity.

4.3. Model Mismatch

We now examine our performance over a set of replicates from a 30-dimensional *parametric* heteroscedastic model. To generate the covariance $\Sigma(x)$ over $\mathcal{X} = \{1, \dots, 500\}$, we chose a set of 5 evenly spaced knots $x_k = 1, 125, 250, 375, 500$ and generated

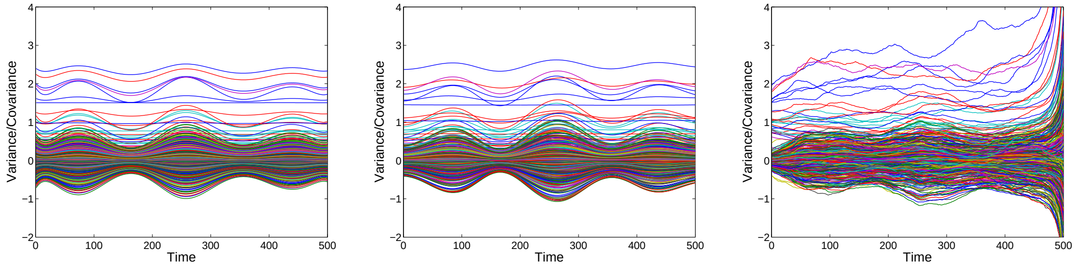
$$S(x_k) \sim \mathcal{N}(0, \Sigma_s) \tag{36}$$

with $\Sigma_s = \sum_{j=1}^{30} s_j s_j'$ and $s_j \sim \mathcal{N}([-29 \;\; -28 \;\; \dots \;\; 28 \;\; 29]', I_{30})$. This construction implies that $S(x_k)$ and $S(x_k')$ are correlated. We then fit a spline $\tilde{S}_{ij}(\cdot)$ independently through each element $S_{ij}(x_k)$ and evaluate this spline fit at $x = 1, \dots, 500$. The covariance is constructed as

$$\Sigma(x) = \alpha \tilde{S}(x) \tilde{S}(x)' + \Sigma_0, \tag{37}$$

where Σ_0 is a diagonal matrix with a truncated-normal prior, $\mathcal{TN}(0, 1)$, on its diagonal elements. The constant α is chosen to scale the maximum value of $\alpha \tilde{S}(x) \tilde{S}(x)'$ to 1. The resulting covariance is shown in Figure 4(a).

§Note that it is not necessary to impute the missing y_{ij} within our proposed Bayesian covariance regression model because of the conditional independencies at each Gibbs step. In Section 5, we simply sample based only on actual observations. Here, however, we impute in order to directly compare our performance to the homoscedastic models.



(a) (b) (c)

Fig. 4. (a) Plot of each component of the true covariance matrix $\Sigma(x)$ over the predictor space $\mathcal{X} = \{1, \dots, 500\}$, taken to represent a time index. Analogous plot for the mean estimate of $\Sigma(x)$ are shown for (b) our proposed Bayesian nonparametric covariance regression model based on Gibbs iterations 5000 to 10000, and (c) a Wishart matrix discounting model over 100 independent FFBS samples. Both mean estimates of $\Sigma(x)$ are for a single replicate $\{y_i^{(1)}\}$. Note that the scaling of (c) crops the large estimates of $\Sigma(x)$ for x near 500.

Each replicate $m = 1, \dots, 30$ of this parametric heteroscedastic model is generated as

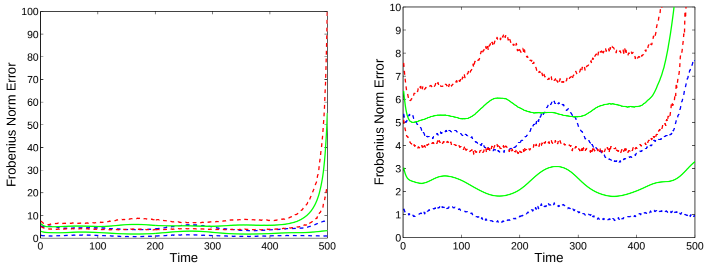
$$y_i^{(m)} \sim \mathcal{N}(0, \Sigma(x_i)). \tag{38}$$

Our hyperparameters and initialization scheme are exactly as in Section 4. The only difference is that we use truncation levels $k^* = L^* = 5$ based on an initial analysis with $k^* = L^* = 17$. For each replicate, we once again run 10,000 Gibbs iterations and thin the chain by examining every 10th sample. A mean estimate of $\Sigma(x)$ is displayed in Figure 4(b). In Figure 5, we plot the mean and 95% highest posterior density intervals of the Frobenius norm $||\Sigma^{(\tau,m)}(x) - \Sigma(x)||_2$ aggregated over iterations $\tau = 9,000, \dots, 10,000$ and replicates $m = 1, \dots, 30$. The average norm error over $\mathcal{X}$ is around 3, which is equivalent to each element of the inferred $\Sigma^{(\tau,m)}(x)$ deviating from the true $\Sigma(x)$ by 0.1. Since the covariance elements are approximately in the range of $[-1, 1]$ and the variances in $[0, 3]$, these norm error values indicate very good estimation performance.

We compare our performance to that of the Wishart matrix discounting model (see Section 10.4.2 of Prado and West [2010]), which is commonly used in stochastic volatility modeling of financial time series. Let $\Phi_t = \Sigma_t^{-1}$. The Wishart matrix discounting model is a discrete-time covariance evolution model that accounts for the slowly changing covariance by discounting the cumulated information. Specifically, assume $\Phi_{t-1} \mid y_{1:t-1}, \beta \sim W(h_{t-1}, D_{t-1}^{-1})$, with $D_t = \beta D_{t-1} + y_t y_t'$ and $h_t = \beta h_{t-1} + 1$. The discounting model then specifies

$$\Phi_t \mid y_{1:t-1}, \beta \sim W(\beta h_{t-1}, (\beta D_{t-1})^{-1}) \tag{39}$$

such that $E[\Phi_t \mid y_{1:t-1}] = E[\Phi_{t-1} \mid y_{1:t-1}] = h_{t-1} D_{t-1}^{-1}$, but with certainty discounted by a factor determined by β. The update with observation y_t is conjugate, maintaining a Wishart posterior on Φ_t. A limitation of this construction is that it constrains $h_t > p - 1$ (or h_t integral) implying that $\beta > (p-2)/(p-1)$. We set $h_0 = 40$ and $\beta = 1 - 1/h_0$ such that $h_t = 40$ for all t and ran the forward filtering backward sampling (FFBS) algorithm outlined in Prado and West [2010], generating 100 independent samples. A mean estimate of $\Sigma(x)$ is displayed in Figure 4(c) and the Frobenius norm error results are depicted in Figure 5. Within the region $x = 1, \dots, 400$, we see that the error of the Wishart matrix discounting



(a) (b)

Fig. 5. (a) Plot of the mean and 95% highest posterior density intervals of the Frobenius norm $||\Sigma^{(\tau,m)}(x) - \Sigma(x)||_2$ for the proposed Bayesian nonparametric covariance regression model (blue and green) and the Wishart matrix discounting model (red and green). The results are aggregated over 100 posterior samples and replicates $m = 1, \ldots, 30$. For the Bayesian nonparametric covariance regression model, these samples are taken at iterations $\tau = [9000 : 10 : 10000]$. (b) Analogous plot, but zoomed in to more clearly see the differences over the range of $x = 1, \ldots, 400$.

method is approximately twice that of our proposed methodology. Furthermore, towards the end of the time series (interpreting $\mathcal{X}$ as representing a batch of time), the estimation error is especially poor due to errors accumulated in forward filtering. Increasing h_t mitigates this problem, but shrinks the model towards homoscedasticity. In general, the formulation is sensitive to the choice of h_t, and in high-dimensional problems this degree of freedom is forced to take large (or integral) values.

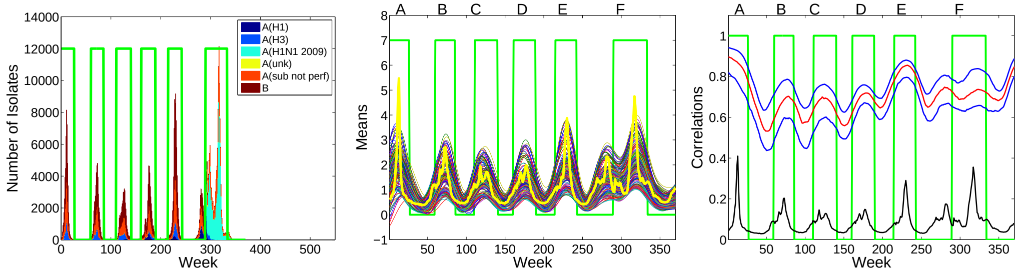
5. Applications

We applied our Bayesian nonparametric covariance regression model to the problem of capturing spatio-temporal structure in influenza rates in the United States (US). Surveillance of influenza has been of growing interest following a series of pandemic scares (e.g., SARS and avian flu) and the 2009 H1N1 pandemic, previously known as “swine flu”. Although influenza pandemics have a long history, such as the 1918-1919 “Spanish flu”, 1957-1958 “Asian flu”, and 1968-1969 “Hong Kong flu”, a convergence of factors are increasing the current public interest in influenza surveillance. These include both practical reasons such as the rapid rate by which geographically distant cases of influenza can spread worldwide, along with other driving factors such as an increased media coverage.

5.1. CDC Influenza Monitoring

The surveillance of influenza within the US is coordinated by the Centers for Disease Control and Prevention (CDC), which collects data from a large network of diagnostic laboratories, hospitals, clinics, individual healthcare providers, and state health departments (see http://www.cdc.gov/flu/weekly/). The approximately 3,000 participating outpatient sites, collectively referred to as the US Outpatient Influenza-Like Illness Surveillance Network (ILINet), provide the CDC with key information about rates of influenza-like illness (ILI)[¶]. The CDC consolidates the ILINet observed cases and produces reports for 10 geographic regions in addition to a US aggregate rate based on a population-based weighted average of

[¶] An influenza-like illness (ILI) is defined as any case of a person having over 100 degrees Fahrenheit fever along with a cough and/or sore throat in absence of any other known cause.



(a) (b) (c)

Fig. 6. (a) Plot of the number of isolates tested positive by the WHO and NREVSS over the period of September 29, 2003 to May 23, 2010. The isolates are divided into virus subtypes, specifically influenza A (H3N2, H1 = {H1N2 and H1N1}, 2009 H1N1) and influenza B. The isolates where subtyping was not performed or was not possible are also indicated. (b) Plot of posterior means of the nonparametric mean function $\mu_j(x)$ for each of the 183 states and regions in the Google Trends Flu dataset. The thick yellow line indicates the Google Flu Trends estimate of the United States influenza rates. (c) For New York, the 25th, 50th, and 75th quantiles of correlation with the 182 other states and regions based on the posterior mean $\hat{\Sigma}(x)$ of the covariance function. The black line is a scaled version of the United States influenza rates, as in (b), shown for easy comparison. The green line shown in plots (a)-(c) indicates the time periods determined to flu events. Specifically, Event A corresponds to the 2003-2004 flu season (flu shot shortage), Event B the 2004-2005 season, Event C the 2005-2006 season (avian flu scare), Event D the 2006-2007 season, Event E the 2007-2008 season (severe), and Event F the 2009-2010 season (2009 H1N1 or "swine flu").

state-level rates. The CDC weekly flu reports are typically released after a 1-2 week delay and are subject to retroactive adjustments based on corrected ILINet reports.

A plot of the number of isolates tested positive by the WHO and NREVSS from 2003-2010 is shown in Figure 6(a). From these data and the CDC weekly flu reports, we defined a set of six events (Events A-F) corresponding to the 2003-2004, 2004-2005, 2005-2006, 2006-2007, 2007-2008, and 2009-2010 flu seasons, respectively. The 2003-2004 flu season began earlier than normal, and coincided with a flu vaccination shortage in many states. For the vaccination that was available, the CDC found that it was "not effective or had very low effectiveness" (http://www.cdc.gov/media/pressrel/fs040115.htm). The 2004-2005 and 2007-2008 flu seasons were more severe than the 2005-2006 and 2006-2007 seasons. However, the 2005-2006 season coincided with an avian flu (H5N1) scare in which Dr. David Narbarro, Senior United Nations System Coordinator for Avian and Human Influenza, was famously quoted as predicting that an avian flu pandemic would lead to 5 million to 150 million deaths. Finally, the 2009-2010 flu season coincides with the emergence of the 2009 H1N1 ("swine flu") subtype[‖] in the United States.

5.2. *Google Flu Trends Dataset*

To aid in a more rapid response to influenza activity, a team of researchers at Google devised a model based on Google user search queries that is predictive of CDC ILI rates [Ginsberg et al., 2008]—that is, the probability that a random physician visit is related to an influenza-

[‖] According to the CDC, "Antigenic characterization of 2009 influenza A (H1N1) viruses indicates that these viruses are only distantly related antigenically and genetically to seasonal influenza A (H1N1) viruses". See http://www.cdc.gov/flu/weekly/weeklyarchives2009-2010/weekly20.htm.

like illness. The *Google Flu Trends* methodology was devised as follows. From the hundreds of billions of individual searches from 2003-2008, time series of state-based weekly query rates were created for the 50 million most common search terms. The predictive performance of a regression on the logit-transformed query rates was examined for each of the 50 million candidates and a ranked list was produced that rewarded terms predictive of rates exhibiting similar regional variations to that of the CDC data. A massive variable selection procedure was then performed to find the optimal combination of query words (based on best fit against out-of-sample ILI data), resulting in a final set of 45 *ILI-related queries*. Using the 45 ILI-related queries as the explanatory variable, a region-independent univariate linear model was fit to the weekly CDC ILI rates from 2003-2007. This model is used for making estimates in any region based on the ILI-related query rates from that region. The results were validated against the CDC data both on training and test data, with the Google reported US and regional rates closely tracking the actual reported rates.

A key advantage of the Google data (available at http://www.google.org/flutrends/) is that the ILI rate predictions are available 1 to 2 weeks before the CDC weekly reports are published. Additionally, a user's IP address is typically connected with a specific geographic area and can thus provide information at a finer scale than the 10-regional and US aggregate reporting provided by the CDC. Finally, the Google reports are not subject to revisions. One important note is that the Google Flu Trends methodology aims to hone in on searches and rates of such searches that are indicative of influenza activity. A methodology based directly on raw search queries might instead track general interest in influenza, waxing and waning quickly with various media events.

We analyze the Google Flu Trends data from the week of September 28, 2003 through the week of October 24, 2010, providing 370 observation vectors y_i. Each observation vector is 183-dimensional with elements consisting of Google estimated ILI rates at the US national level, the 50 states, 10 U.S. Department of Health & Human Services surveillance regions, and 122 cities. It is important to note, however, that there is substantial missing data with entire blocks of observations unavailable (as opposed to certain weeks sporadically being omitted). At the beginning of the examined time frame only 114 of the 183 regions were reporting. By the end of Year 1, there were 130 regions. These numbers increased to 173, 178, 180, and 183 by the end of Years 2, 3, 4, and 5, respectively. The high-dimensionality and missing data structure make the Google Flu Trends dataset challenging to analyze in its entirety with existing heteroscedastic models. As part of an exploratory data analysis, in Figure 7 we plot sample estimates of the geographic correlation structure between the states during an event period for four representative states. Specifically, we first subtract a moving average estimate of the mean (window size 10) and then aggregate the data over Events B-F, omitting Event A due to the quantity of missing data. Because of the dimensionality of the data (183 dimensions) and the fact that there are only 157 event observations, we simply consider the state-level observations (plus District of Columbia), reducing the dimensionality to 51. The limited data also impedes our ability to perform time-specific sample estimates of geographic correlations.

5.3. Heteroscedastic Modeling of Google Flu Trends

Our proposed heteroscedastic model allows one to capture both spatial and temporal changes in correlation structure, providing an important additional tool in predicting influenza rates. We specifically consider $y_i \sim \mathcal{N}(\mu(x_i), \Sigma(x_i))$ with the nonparametric function $\mu(x_i) = \Theta\xi(x_i)\psi(x_i)$ defining the mean of the ILI rates in each of the 183 regions. For

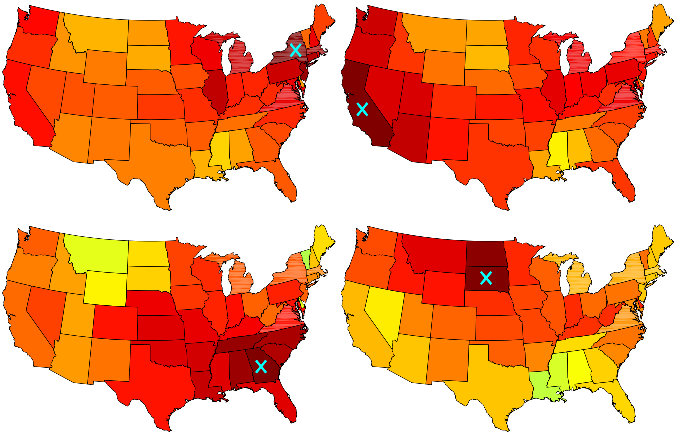


Fig. 7. For each of four geographically distinct states (New York, California, Georgia, and South Dakota), plots of correlations between the state and all other states based on the sample covariance estimate from aggregating the state-level data during the event periods B-F after subtracting a moving average estimate of the mean. Event A was omitted due to an insufficient number of states reporting. Note that South Dakota is missing 58 of the 157 Event B-F observations.

a given week x_i, the spatial correlation structure is captured by the covariance $\Sigma(x_i) = \Theta\xi(x_i)\xi(x_i)'\Theta' + \Sigma_0$. Temporal changes are implicitly modeled through the proposed covariance regression framework that allows for continuous variations in $\Sigma(x_i)$. Dukić et al. [2009] also examine portions of the Google Flu Trends data, but with the goal of on-line tracking of influenza rates on either a national, state, or regional level. Specifically, they employ a state-space model with particle learning. Our goal differs considerably. We aim to jointly analyze the full 183-dimensional data, as opposed to univariate modeling. Through such joint modeling, we can uncover important spatial dependencies lost when analyzing components of the data individually. Such spatial information can be key in predicting influenza rates based on partial observations from select regions or in retrospectively imputing missing data.

There are a few crucial points to note. The first is that no geographic information is provided to our model. Instead, the spatial structure is uncovered simply from analyzing the raw 183-dimensional time series and patterns therein. Second, because of the substantial quantity of missing data, imputing the missing values as in Section 4.2 is less ideal than simply updating our posterior based solely on the data that is available. The latter is how we chose to analyze the Google Flu Trends data—our ability to do so without introducing any approximations is a key advantage of our proposed methodology.

The results presented in Figures 6 and 8 clearly demonstrate that we are able to capture both spatial and temporal changes in correlations in the Google Flu Trends data, even in the presence of substantial missing information. We preprocessed the data by scaling the entire dataset by one over the largest variance of any of the 183 time series. The hyperparameters were set as in the simulation study of Section 4, except with larger truncation levels $L^* = 10$ and $k^* = 20$ and with the Gaussian process length-scale hyperparameter set to $\kappa = 100$ to account for the time scale (weeks) and the rate at which ILI incidences change. Once again, by examining posterior samples of Θ, we found that the chosen truncation level was sufficiently large. We ran 10 chains each for 10,000 MCMC iterations, discarded the first 5,000 for burn-in, and then thinned the chains by examining every 10 samples. Each

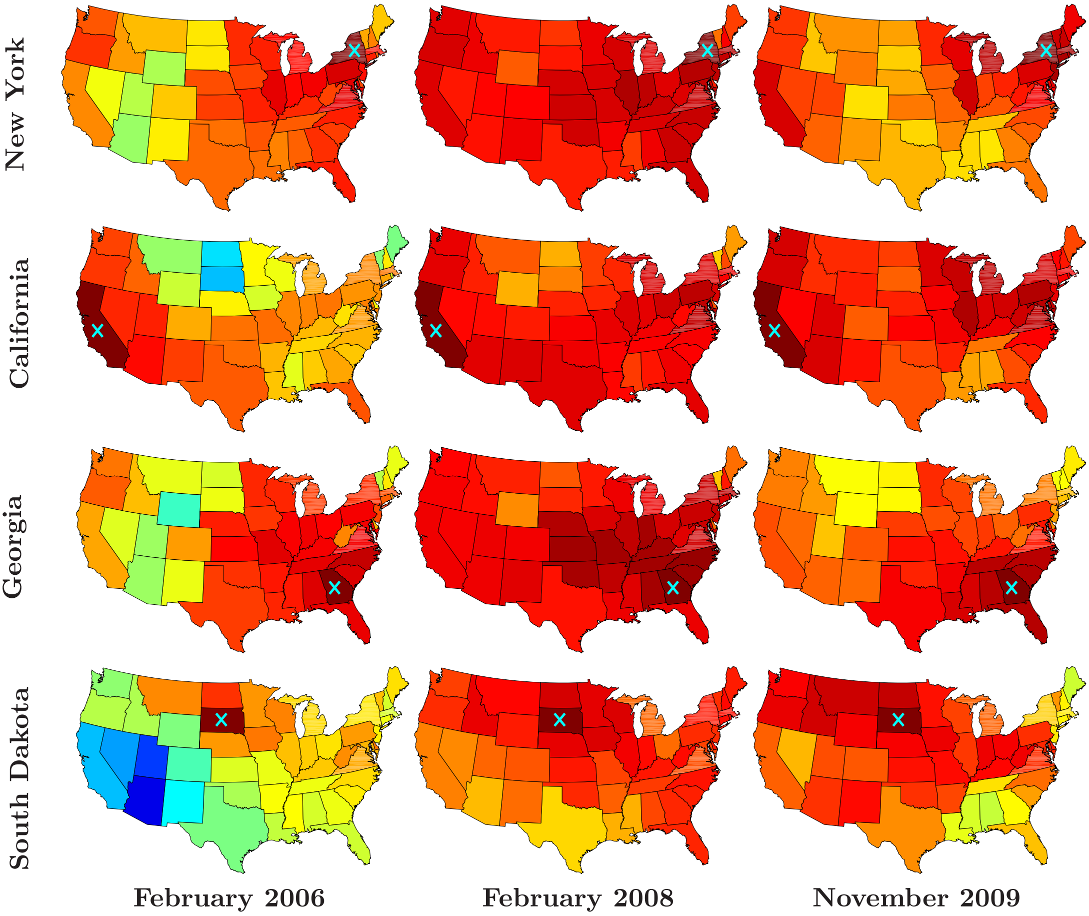


Fig. 8. For each of four geographically distinct states (New York, California, Georgia, and South Dakota) and each of three key dates (February 2006 of Event C, February 2008 of Event E, and November 2009 of Event F), plots of correlations between the state and all other states based on the posterior mean $\hat{\Sigma}(x)$ of the covariance function. The plots clearly indicate spatial structure captured by $\Sigma(x)$, and that these spatial dependencies change over time. Note that no geographic information was included in our model. Compare to the maps of Figure 7.

chain was initialized with parameters sampled from the prior. To assess convergence, we performed the modified Gelman-Rubin diagnostic of Brooks and Gelman [1998] on the MCMC samples of the variance terms $\Sigma_{jj}(x_i)$ for j corresponding to the state indices of New York, California, Georgia, and South Dakota, and x_i corresponding to the midpoint of each of the 12 event and non-event time windows**. These elements are spatially and geographically disparate, with South Dakota corresponding to an element with substantial missing data. Of the 48 resulting variables examined (4 states and 12 time points), 40 had potential scale reduction factors $R^{1/2} < 1.2$ and most $R^{1/2} < 1.1$. The variables with larger $R^{1/2}$ (all less than 1.4) corresponded to two non-event time periods. We also performed hyperparameter sensitivity, doubling the length-scale parameter to $\kappa = 200$ (implying less

**Our tests used the code provided at http://www.lce.hut.fi/research/mm/mcmcdiag/.

temporal correlation) and using a larger truncation level of $L^* = k^* = 20$ with less stringent shrinkage hyperparameters $a_1 = a_2 = 2$ (instead of $a_1 = a_2 = 10$). The results were very similar to those presented in this section, with all conclusions remaining the same.

In Figure 6(b), we plot the posterior mean of the 183 components of $\mu(x)$, showing trends that follow the Google estimated US national ILI rate. For New York, in Figure 6(c) we plot the 25th, 50th, and 75th quantiles of correlation with the 182 other states and regions based on the posterior mean $\hat{\Sigma}(x)$ of the covariance function. From this plot, we immediately notice that regions become more correlated during flu seasons, as we would expect. The specific geographic structure of these correlations is displayed in Figure 8, where we see key changes with the specific flu event. In the more mild 2005-2006 season, we see much more local correlation structure than the more severe 2007-2008 season (which still maintains stronger regional than distant correlations.) The November 2009 H1N1 event displays overall regional correlation structure and values similar to the 2007-2008 season, but with key geographic areas that are less correlated. Note that some geographically distant states, such as New York and California, are often highly correlated as we might expect based on their demographic similarities and high rates of travel between them. Interestingly, the strong local spatial correlation structure for South Dakota in February 2006 has been inferred before any data are available for that state. Actually, no data are available for South Dakota from September 2003 to November 2006. Despite this missing data, the inferred correlation structures over these years are fairly consistent with those of neighboring states and change in manners similar to the flu-to-non-flu changes inferred after data for South Dakota are available.

Comparing the maps of Figure 8 to those of the sample-based estimates in Figure 7, we see much of the same correlation structure, which at a high level validates our findings. Since the sample-based estimates aggregate data over Events B-F (containing those displayed in Figure 8), they tend to represent a time-average of the event-specific correlation structure we uncovered. Note that due to the dimensionality of the dataset, the sample-based estimates are based solely on state-level measurements and thus are unable to harness the richness (and crucial redundancy) provided by the other regional reporting agencies. Furthermore, since there are a limited number of per-event observations, the naive approach of forming sample covariances based on local bins of data is infeasible. The high-dimensionality and missing data structure of this dataset also limit our ability to compare to alternative methods such as those cited in Section 1—none yield results directly comparable to the full analysis we have provided here. Instead, they are either limited to examination of the small subset of data for which all observations are present and/or a lower-dimensional selection (or projection) of observations. On the other hand, our proposed algorithm can readily utilize all information available to model the heteroscedasticity present here. (Compare to the common GARCH models, which cannot handle missing data and are limited to typically no more than 5 dimensions.) In terms of computational complexity, each of our chains of 10,000 Gibbs iterations based on a naive implementation in MATLAB (R2010b) took approximately 12 hours on a machine with four Intel Xeon X5550 Quad-Core 2.67GHz processors and 48 GB of RAM.

6. Discussion

In this paper, we have presented a Bayesian nonparametric approach to covariance regression which allows an unknown $p \times p$ dimensional covariance matrix $\Sigma(x)$ to vary flexibly over $x \in \mathcal{X}$, where $\mathcal{X}$ is some arbitrary (potentially multivariate) predictor space. Foundational to our formulation is quadratic mixing over a collection of dictionary elements, assumed herein to be Gaussian process random functions, defined over $\mathcal{X}$. By inducing a prior on $\Sigma_{\mathcal{X}} = \{\Sigma(x), x \in \mathcal{X}\}$ through a shrinkage prior on a predictor-dependent latent factor model, we are able to scale to the large p domain. Our proposed methodology also yields computationally tractable algorithms for posterior inference via fully conjugate Gibbs updates—this is crucial in our being able to analyze high-dimensional multivariate datasets. We demonstrated the utility of our Bayesian covariance regression framework through both simulated studies and analysis of the Google Trends flu dataset, the latter having nearly 200 dimensions.

There are many possible interesting and relatively direct extensions of the proposed covariance regression framework. The most immediate are those that fall into the categories of (i) addressing the limitations of our current assumption of Gaussian marginals, and (ii) scaling to datasets with large numbers of observations.

In longitudinal studies or spatio-temporal datasets, one is faced with repeated collections of observations over the predictor space. These collections are clearly not independent. To cope with such data, one could envision embedding the current framework within a hierarchical model (e.g., to model random effects on a mean), or otherwise use the proposed methodology as a building block in more complicated models. Additionally, it would be trivial to extend our framework to accommodate multivariate categorical responses, or joint categorical and continuous responses, by employing the latent variable probit model of Albert and Chib [1993]. This would lead, for example, to a more flexible class of multivariate probit models in which the correlation between variables can change with time and other factors. For computation, the use of parameter expansion allows us to simply modify our current MCMC algorithm to include a data augmentation step for imputing the underlying continuous variables. Imposing the constraints on the covariance could be deferred to a post-processing stage. Another interesting direction for future research is to consider embedding our covariance regression model in a Gaussian copula model. One possible means of accomplishing this is through a variant of the approach proposed by Hoff [2007], which avoids having to completely specify the marginal distributions.

As discussed in Section 3.4, our sampler relies on $L * \times k*$ draws from an n-dimensional Gaussian (i.e., posterior draws of our Gaussian process random dictionary functions). For very large n, this becomes infeasible in practice since computations are, in general, $O(n^3)$. Standard tools for scaling up Gaussian process computation to large datasets, such as covariance tapering [Kaufman et al., 2008, Du et al., 2009] and the predictive process [Banerjee et al., 2008], can be applied directly in our context. Additionally, one might consider using the integrated nested Laplace approximations of Rue et al. [2009] for computations. The size of the dataset (both in terms of n and p) also dramatically affects our ability to sample the Gaussian process length-scale hyperparameter κ since our proposed method relies on samples from an np-dimensional Gaussian. See Section 3.3 for details and possible methods of addressing this issue. If it is feasible to perform inference over the length-scale parameter, one can consider implicitly including a test for homoscedasticity by considering κ taking values in the extended reals (i.e., $\Re \bigcup \{\infty\}$) and thus allowing our formulation to collapse on the simpler model if $\kappa = \infty$.

Finally, we note that there are scenarios in which the functional data itself is covariance-valued, such as in diffusion tensor imaging. In this case, each voxel in an image consists of a 3×3 covariance matrix that has potential spatio-temporal dependencies. Specifically, take $\Sigma_{ij}(t)$ to represent the observed covariance matrix for subject i at pixel j and time t.

Here, one could imagine replacing the Gaussian process dictionary elements with splines and embedding this model within a hierarchical structure to allow variability among subjects while borrowing information.

As we readily see, the presented Bayesian nonparametric covariance regression framework easily lends itself to many interesting directions for future research with the potential for dramatic impact in many applied domains.

Acknowledgements

The authors would like to thank Surya Tokdar for helpful discussions on the proof of prior support for the proposed covariance regression formulation.

A. Proofs of Theorems and Lemmas

Proof (Proof of Theorem 2.1). Since $\mathcal{X}$ is compact, for every $\epsilon_0 > 0$ there exists an open covering of ϵ_0-balls $B_{\epsilon_0}(x_0) = \{x : ||x - x_0||_2 < \epsilon_0\}$ with a finite subcover such that $\bigcup_{x_0 \in \mathcal{X}_0} B_{\epsilon_0}(x_0) \supset \mathcal{X}$, where $|\mathcal{X}_0| = n$. Then,

$$\Pi_\Sigma \left(\sup_{x \in \mathcal{X}} ||\Sigma(x) - \Sigma^*(x)||_2 < \epsilon \right) = \Pi_\Sigma \left(\max_{x_0 \in \mathcal{X}_0} \sup_{x \in B_{\epsilon_0}(x_0)} ||\Sigma(x) - \Sigma^*(x)||_2 < \epsilon \right). \tag{40}$$

Define $Z(x_0) = \sup_{x \in B_{\epsilon_0}(x_0)} ||\Sigma(x) - \Sigma^*(x)||_2$. Since

$$\Pi_\Sigma \left(\max_{x_0 \in \mathcal{X}_{\prime}} Z(x_0) < \epsilon \right) > 0 \iff \Pi_\Sigma \left(Z(x_0) < \epsilon \right) > 0, \, \forall x_0 \in \mathcal{X}_0, \tag{41}$$

we only need to look at each ϵ_0-ball independently, which we do as follows.

$$\begin{aligned}
&\Pi_\Sigma \left(\sup_{x \in B_{\epsilon_0}(x_0)} ||\Sigma(x) - \Sigma^*(x)||_2 < \epsilon \right) \\
&\geq \Pi_\Sigma \left(||\Sigma(x_0) - \Sigma^*(x_0)||_2 + \sup_{x \in B_{\epsilon_0}(x_0)} ||\Sigma^*(x_0) - \Sigma^*(x)||_2 + \sup_{x \in B_{\epsilon_0}(x_0)} ||\Sigma(x_0) - \Sigma(x)||_2 < \epsilon \right) \\
&\geq \Pi_\Sigma \left(||\Sigma(x_0) - \Sigma^*(x_0)||_2 < \epsilon/3 \right) \\
&\qquad \cdot \Pi_\Sigma \left(\sup_{x \in B_{\epsilon_0}(x_0)} ||\Sigma^*(x_0) - \Sigma^*(x)||_2 < \epsilon/3 \right) \cdot \Pi_\Sigma \left(\sup_{x \in B_{\epsilon_0}(x_0)} ||\Sigma(x_0) - \Sigma(x)||_2 < \epsilon/3 \right),
\end{aligned} \tag{42}$$

where the first inequality comes from repeated uses of the triangle inequality, and the second inequality follows from the fact that each of these terms is an independent event. We evaluate each of these terms in turn. The first follows directly from the assumed continuity of $\Sigma^*(\cdot)$. The second will follow from a statement of (almost sure) continuity of $\Sigma(\cdot)$ that arises from the (almost sure) continuity of the $\xi_{\ell k}(\cdot) \sim \mathrm{GP}(0, c)$ and the shrinkage prior on $\theta_{\ell k}$ (i.e., $\theta_{\ell k} \to 0$ almost surely as $\ell \to \infty$, and does so "fast enough".) Finally, the third will follow from the support of the conditionally Wishart prior on $\Sigma(x_0)$ at every fixed $x_0 \in \mathcal{X}$.

Based on the continuity of $\Sigma^*(\cdot)$, for all $\epsilon/3 > 0$ there exists an $\epsilon_{0,1} > 0$ such that

$$||\Sigma^*(x_0) - \Sigma^*(x)||_2 < \epsilon/3, \quad \forall ||x - x_0||_2 < \epsilon_{0,1}. \tag{43}$$

Therefore, $\Pi_\Sigma\left(\sup_{x\in B_{\epsilon_{0,1}}(x_0)} ||\Sigma^*(x_0)-\Sigma^*(x)||_2<\epsilon/3\right)=1$.

Based on Theorem 2.2, each element of $\Lambda(\cdot)\triangleq\Theta\xi(\cdot)$ is almost surely continuous on $\mathcal{X}$ assuming k finite. Letting $g_{jk}(x)=[\Lambda(x)]_{jk}$,

$$[\Lambda(x)\Lambda(x)']_{ij}=\sum_{m=1}^{k} g_{im}(x)g_{jm}(x), \quad \forall x\in\mathcal{X}. \tag{44}$$

Eq. (44) represents a finite sum over pairwise products of almost surely continuous functions, and thus results in a matrix $\Lambda(x)\Lambda(x)'$ with elements that are almost surely continuous on $\mathcal{X}$. Therefore, $\Sigma(x)=\Lambda(x)\Lambda(x)'+\Sigma_0=\Theta\xi(x)\xi(x)'\Theta'+\Sigma_0$ is almost surely continuous on $\mathcal{X}$. We can then conclude that for all $\epsilon/3>0$ there exists an $\epsilon_{0,2}>0$ such that

$$\Pi_\Sigma\left(\sup_{x\in B_{\epsilon_{0,2}}(x_0)} ||\Sigma(x_0)-\Sigma(x)||_2<\epsilon/3\right)=1. \tag{45}$$

To examine the third term, we first note that

$$\begin{aligned}&\Pi_\Sigma\left(||\Sigma(x_0)-\Sigma^*(x_0)||_2<\epsilon/3\right)\\ &\qquad=\Pi_\Sigma\left(||\Theta\xi(x_0)\xi(x_0)'\Theta'+\Sigma_0-\Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}+\Sigma_0^*||_2<\epsilon/3\right),\end{aligned} \tag{46}$$

where $\{\xi^*(x_0),\Theta^*,\Sigma_0^*\}$ is any element of $\mathcal{X}_\xi\otimes\mathcal{X}_\Theta\otimes\mathcal{X}_{\Sigma_0}$ such that $\Sigma^*(x_0)=\Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}+\Sigma_0^*$. Such a factorization exists by Lemma 2.1. We can then bound this prior probability by

$$\begin{aligned}&\Pi_\Sigma\left(||\Sigma(x_0)-\Sigma^*(x_0)||_2<\epsilon/3\right)\\ &\geq\Pi_\Sigma\left(||\Theta\xi(x_0)\xi(x_0)'\Theta'-\Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}||_2<\epsilon/6\right)\Pi_{\Sigma_0}\left(||\Sigma_0-\Sigma_0^*||_2<\epsilon/6\right)\\ &\geq\Pi_\Sigma\left(||\Theta\xi(x_0)\xi(x_0)'\Theta'-\Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}||_2<\epsilon/6\right)\Pi_{\Sigma_0}\left(||\Sigma_0-\Sigma_0^*||_\infty<\epsilon/(6\sqrt{\pi})\right),\end{aligned} \tag{47}$$

where the first inequality follows from the triangle inequality, and the second from the fact that for all $A\in\Re^{p\times p}$, $||A||_2\leq\sqrt{p}||A||_\infty$, with the sup-norm defined as $||A||_\infty=\max_{1\leq i\leq p}\sum_{i=1}^{p}|a_{ij}|$. Since $\Sigma_0=diag(\sigma_1^2,\dots,\sigma_p^2)$ with $\sigma_i^2\overset{i.i.d.}{\sim}\mathrm{Ga}(a_\sigma,b_\sigma)$, the support of the gamma prior implies that

$$\Pi_{\Sigma_0}\left(||\Sigma_0-\Sigma_0^*||_\infty<\epsilon/(6\sqrt{\pi})\right)=\Pi_{\Sigma_0}\left(\max_{1\leq i\leq p}|\sigma_i^2-\sigma_i^{*2}|<\epsilon/(6\sqrt{\pi})\right)>0. \tag{48}$$

Recalling that $[\xi(x_0)]_{\ell k}=\xi_{\ell k}(x_0)$ with $\xi_{\ell k}(x_0)\overset{i.i.d.}{\sim}\mathcal{N}(0,1)$ and taking $\Theta\in\Re^{p\times L}$ with $\mathrm{rank}(\Theta)=p$,

$$\Theta\xi(x_0)\xi(x_0)'\Theta'\mid\Theta\sim\mathrm{W}(k,\Theta\Theta'). \tag{49}$$

When $k\geq p$, $\Theta\xi(x_0)\xi(x_0)'\Theta'$ is invertible (i.e., full rank) with probability 1.

By Assumption 2.2, there is positive probability under Π_Θ on the set of Θ such that $\mathrm{rank}(\Theta)=p$. Since $\Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}$ is an arbitrary symmetric positive semidefinite matrix in $\Re^{p\times p}$, and based on the support of the Wishart distribution,

$$\Pi_\Sigma\left(||\Theta\xi(x_0)\xi(x_0)'\Theta'-\Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}||_2<\epsilon/6\right)>0. \tag{50}$$

We thus conclude that $\Pi_\Sigma\left(||\Sigma(x_0) - \Sigma^*(x_0)||_2 < \epsilon/3\right) > 0$.

For every $\Sigma^*(\cdot)$ and $\epsilon > 0$, let $\epsilon_0 = \min(\epsilon_{0,1}, \epsilon_{0,2})$ with $\epsilon_{0,1}$ and $\epsilon_{0,2}$ defined as above. Then, combining the positivity results of each of the three terms in Eq. (42) completes the proof.

Proof (Proof of Theorem 2.2). We can represent each element of $\Lambda(\cdot)$ as follows:

$$[\Lambda(\cdot)]_{jk} = \lim_{L\to\infty}\left[\begin{bmatrix}\theta_{11} & \theta_{12} & \dots & \theta_{1L}\\ \theta_{21} & \theta_{22} & \dots & \theta_{2L}\\ \vdots & \vdots & \ddots & \vdots\\ \theta_{p1} & \theta_{p2} & \dots & \theta_{pL}\end{bmatrix}\begin{bmatrix}\xi_{11}(\cdot) & \xi_{12}(\cdot) & \dots & \xi_{1k}(\cdot)\\ \xi_{21}(\cdot) & \xi_{22}(\cdot) & \dots & \xi_{2k}(\cdot)\\ \vdots & \vdots & \ddots & \vdots\\ \xi_{L1}(\cdot) & \xi_{L2}(\cdot) & \dots & \xi_{Lk}(\cdot)\end{bmatrix}\right]_{jk} = \sum_{\ell=1}^{\infty}\theta_{j\ell}\xi_{\ell k}(\cdot). \tag{51}$$

If $\xi_{\ell k}(x)$ is continuous for all ℓ, k and $s_n(x) = \sum_{\ell=1}^{\infty}\theta_{j\ell}\xi_{\ell k}(x)$ uniformly converges almost surely to some $g_{jk}(x)$, then $g_{jk}(x)$ is almost surely continuous. That is, if for all $\epsilon > 0$ there exists an N such that for all $n \geq N$

$$\Pr\left(\sup_{x\in\mathcal{X}}|g_{jk}(x) - s_n(x)| < \epsilon\right) = 1, \tag{52}$$

then $s_n(x)$ converges uniformly almost surely to $g_{jk}(x)$ and we can conclude that $g_{jk}(x)$ is continuous based on the definition of $s_n(x)$. To show almost sure uniform convergence, it is sufficient to show that there exists an M_n with $\sum_{n=1}^{\infty} M_n$ almost surely convergent and

$$\sup_{x\in\mathcal{X}}|\theta_{jn}\xi_{nk}(x)| \leq M_n. \tag{53}$$

Let $c_{nk} = \sup_{x\in\mathcal{X}}|\xi_{nk}(x)|$. Then,

$$\sup_{x\in\mathcal{X}}|\theta_{jn}\xi_{nk}(x)| \leq |\theta_{jn}|c_{nk}. \tag{54}$$

Since $\xi_{nk}(\cdot) \overset{i.i.d.}{\sim} \mathrm{GP}(0, c)$ and $\mathcal{X}$ is compact, $c_{nk} < \infty$ and $E[c_{nk}] = \bar{c}$ with $\bar{c}$ finite. Defining $M_n = |\theta_{jn}|c_{nk}$,

$$E_{\Theta,c}\left[\sum_{n=1}^{\infty}M_n\right] = E_\Theta\left[E_{c|\Theta}\left[\sum_{n=1}^{\infty}|\theta_{jn}|c_{nk} \mid \Theta\right]\right] = E_\Theta\left[\sum_{n=1}^{\infty}|\theta_{jn}|\bar{c}\right] = \bar{c}\sum_{n=1}^{\infty}E_\Theta\left[|\theta_{jn}|\right], \tag{55}$$

where the last equality follows from Fubini's theorem. Based on Assumption 2.1, we conclude that $E[\sum_{n=1}^{\infty}M_n] < \infty$ which implies that $\sum_{n=1}^{\infty}M_n$ converges almost surely.

Proof (Proof of Lemma 2.2). Recall that $\theta_{j\ell} \sim \mathcal{N}(0, \phi_{j\ell}^{-1}\tau_\ell^{-1})$ with $\phi_{j\ell} \sim \mathrm{Ga}(3/2, 3/2)$ and $\tau_\ell = \prod_{h=1}^{\ell}\delta_h$ for $\delta_1 \sim \mathrm{Ga}(a_1, 1)$, $\delta_h \sim \mathrm{Ga}(a_2, 1)$. Using the fact that if $x \sim \mathcal{N}(0, \sigma)$ then $E[|x|] = \sigma\sqrt{2/\pi}$ and if $y \sim \mathrm{Ga}(a, b)$ then $1/y \sim \text{Inv-Ga}(a, 1/b)$ with $E[1/y] = 1/(b\cdot(a-1))$, we derive that

$$\begin{aligned}\sum_{\ell=1}^{\infty}E_\theta[|\theta_{j\ell}|] &= \sum_{\ell=1}^{\infty}E_{\phi,\tau}[E_{\theta|\phi,\tau}[|\theta_{j\ell}| \mid \phi_{j\ell}, \tau_\ell]] = \sqrt{\frac{2}{\pi}}\sum_{\ell=1}^{\infty}E_{\phi,\tau}[\phi_{j\ell}^{-1}\tau_\ell^{-1}]\\ &= \sqrt{\frac{2}{\pi}}\sum_{\ell=1}^{\infty}E_\phi[\phi_{j\ell}^{-1}]E_\tau[\tau_\ell^{-1}] = \frac{4}{3}\sqrt{\frac{2}{\pi}}\sum_{\ell=1}^{\infty}E_\delta\left[\prod_{h=1}^{\ell}\frac{1}{\delta_h}\right] = \frac{1}{a_1-1}\frac{4}{3}\sqrt{\frac{2}{\pi}}\sum_{\ell=1}^{\infty}\frac{1}{a_2-1}^{\ell-1}.\end{aligned} \tag{56}$$

When $a_2 > 2$, we conclude that $\sum_\ell E[|\theta_{j\ell}|] < \infty$.

Proof (Proof of Lemma 2.3). Recall that $\Sigma(x) = \Theta\xi(x)\xi(x)'\Theta' + \Sigma_0$ with $\Sigma_0 = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$. The elements of the respective matrices are independently distributed as $\theta_{i\ell} \sim \mathcal{N}(0, \phi_{i\ell}^{-1}\tau_\ell^{-1})$, $\xi_{\ell k}(\cdot) \sim \mathrm{GP}(0, c)$, and $\sigma_i^{-2} \sim \mathrm{Gamma}(a_\sigma, b_\sigma)$. Let μ_σ and σ_σ^2 represent the mean and variance of the implied inverse gamma prior on σ_i^2, respectively. In all of the following, we first condition on Θ and then use iterated expectations to find the marginal moments.

The expected covariance matrix at any predictor location x is simply derived as

$$\begin{aligned} E[\Sigma(x)] &= E[E[\Sigma(x) \mid \Theta]] = E[E[\Theta\xi(x)\xi(x)'\Theta' \mid \Theta]] + \mu_\sigma I_p \\ &= kE[\Theta\Theta'] + \mu_\sigma I_p \\ &= \mathrm{diag}(k\sum_\ell \phi_{1\ell}^{-1}\tau_\ell^{-1} + \mu_\sigma, \dots, k\sum_\ell \phi_{p\ell}^{-1}\tau_\ell^{-1} + \mu_\sigma). \end{aligned}$$

Here, we have used the fact that conditioned on Θ, $\Theta\xi(x)\xi(x)'\Theta'$ is Wishart distributed with mean $k\Theta\Theta'$ and

$$\begin{aligned} E[\Theta\Theta']_{ij} &= \sum_\ell \sum_{\ell'} E[\theta_{i\ell}\theta_{j\ell'}] = \sum_\ell E[\theta_{i\ell}^2]\delta_{ij} \\ &= \sum_\ell \mathrm{var}(\theta_{i\ell})\delta_{ij} = \left(\sum_\ell \phi_{i\ell}^{-1}\tau_\ell^{-1}\right)\delta_{ij}. \end{aligned}$$

Proof (Proof of Lemma 2.4). One can use the conditionally Wishart distribution of $\Theta\xi(x)\xi(x)'\Theta'$ to derive $\mathrm{cov}(\Sigma_{ij}(x), \Sigma_{uv}(x))$. Specifically, let $S = \Theta\xi(x)\xi(x)'\Theta'$. Then $S = \sum_{n=1}^k z^{(n)}z^{(n)'}$ with $z^{(n)} \mid \Theta \sim \mathcal{N}(0, \Theta\Theta')$ independently for each n. Then, using standard Gaussian second and fourth moment results,

$$\begin{aligned} \mathrm{cov}(\Sigma_{ij}(x), \Sigma_{uv}(x) \mid \Theta) &= \mathrm{cov}(S_{ij}, S_{uv} \mid \Theta) + \sigma_\sigma^2\delta_{ijuv} \\ &= \sum_{n=1}^k E[z_i^{(n)}z_j^{(n)}z_u^{(n)}z_v^{(n)} \mid \Theta] - E[z_i^{(n)}z_j^{(n)} \mid \Theta]E[z_u^{(n)}z_v^{(n)} \mid \Theta] + \sigma_\sigma^2\delta_{ijuv} \\ &= k((\Theta\Theta')_{iu}(\Theta\Theta')_{jv} + (\Theta\Theta')_{iv}(\Theta\Theta')_{ju}) + \sigma_\sigma^2\delta_{ijuv}. \end{aligned}$$

Here, $\delta_{ijuv} = 1$ if $i = j = u = v$ and is 0 otherwise. Taking the expectation with respect to Θ yields $\mathrm{cov}(\Sigma_{ij}(x), \Sigma_{uv}(x))$. However, instead of looking at one slice of the predictor space, what we are really interested in is how the correlation between elements of the covariance matrix changes with predictors. Thus, we work directly with the latent Gaussian processes to derive $\mathrm{cov}(\Sigma_{ij}(x), \Sigma_{uv}(x'))$. Let

$$g_{in}(x) = \sum_\ell \theta_{i\ell}\xi_{\ell n}(x), \tag{57}$$

implying that $g_{in}(x)$ is independent of all $g_{im}(x')$ for any $m \neq n$ and all $x' \in \mathcal{X}$. Since each $\xi_{\ell n}(\cdot)$ is distributed according to a zero mean Gaussian process, $g_{in}(x)$ is zero mean. Using this definition, we condition on Θ (which is dropped in the derivations for notational simplicity) and write

$$\begin{aligned}\operatorname{cov}(\Sigma_{ij}(x),\Sigma_{uv}(x') \mid \Theta) &= \sum_{n=1}^{k}\operatorname{cov}(g_{in}(x)g_{jn}(x), g_{un}(x'), g_{vn}(x')) + \sigma_\sigma^2\delta_{ijuv}\\ &= \sum_{n=1}^{k} E[g_{in}(x)g_{jn}(x)g_{un}(x'), g_{vn}(x')] - E[g_{in}(x)g_{jn}(x)]E[g_{un}(x'), g_{vn}(x')] + \sigma_\sigma^2\delta_{ijuv}\end{aligned}$$

We replace each $g_{kn}(x)$ by the form in Equation (57), summing over different dummy indices for each. Using the fact that $\xi_{\ell n}(x)$ is independent of $\xi_{\ell' n}(x')$ for any $\ell \neq \ell'$ and that each $\xi_{\ell n}(x)$ is zero mean, all cross terms in the resulting products cancel if a $\xi_{\ell n}(x)$ arising from one $g_{kn}(x)$ does not share an index ℓ with at least one other $\xi_{\ell n}(x)$ arising from another $g_{pn}(x)$. Thus,

$$\begin{aligned}\operatorname{cov}(\Sigma_{ij}(x),\Sigma_{uv}(x') \mid \Theta) = \sum_{n=1}^{k}\sum_{\ell}&\theta_{i\ell}\theta_{j\ell}\theta_{u\ell}\theta_{v\ell}E[\xi_{\ell n}^2(x)\xi_{\ell n}^2(x')]\\ &+ \sum_{\ell}\theta_{i\ell}\theta_{u\ell}E[\xi_{\ell n}(x)\xi_{\ell n}(x')]\sum_{\ell'\neq\ell}\theta_{j\ell'}\theta_{v\ell'}E[\xi_{\ell' n}(x)\xi_{\ell' n}(x')]\\ &+ \sum_{\ell}\theta_{i\ell}\theta_{j\ell}E[\xi_{\ell n}^2(x)]\sum_{\ell'\neq\ell}\theta_{u\ell'}\theta_{v\ell'}E[\xi_{\ell' n}^2(x')]\\ &- \sum_{\ell}\theta_{i\ell}\theta_{j\ell}E[\xi_{\ell n}^2(x)]\sum_{\ell'}\theta_{u\ell'}\theta_{v\ell'}E[\xi_{\ell' n}^2(x')] + \sigma_\sigma^2\delta_{ijuv}\end{aligned}$$

The Gaussian process moments are given by

$$\begin{aligned}E[\xi_{\ell n}^2(x)] &= 1\\ E[\xi_{\ell n}(x)\xi_{\ell n}(x')] &= E[E[\xi_{\ell n}(x) \mid \xi_{\ell n}(x')]\xi_{\ell n}(x')] = c(x,x')E[\xi_{\ell n}^2(x')] = c(x,x')\\ E[\xi_{\ell n}^2(x)\xi_{\ell n}^2(x')] &= E[E[\xi_{\ell n}^2(x) \mid \xi_{\ell n}(x')]\xi_{\ell n}^2(x')]\\ &= E[\{(E[\xi_{\ell n}(x) \mid \xi_{\ell n}(x')])^2 + \operatorname{var}(\xi_{\ell n}(x) \mid \xi_{\ell n}(x'))\}\xi_{\ell n}^2(x')]\\ &= c^2(x,x')E[\xi_{\ell n}^4(x')] + (1-c^2(x,x'))E[\xi_{\ell n}^2(x')] = 2c^2(x,x')+1,\end{aligned}$$

from which we derive that

$$\begin{aligned}\operatorname{cov}(\Sigma_{ij}(x),\Sigma_{uv}(x') \mid \Theta) &= k\Bigg\{(2c^2(x,x')+1)\sum_{\ell}\theta_{i\ell}\theta_{j\ell}\theta_{u\ell}\theta_{v\ell} + c^2(x,x')\sum_{\ell}\theta_{i\ell}\theta_{u\ell}\sum_{\ell'\neq\ell}\theta_{j\ell'}\theta_{v\ell'}\\ &\quad + \sum_{\ell}\theta_{i\ell}\theta_{j\ell}\sum_{\ell'\neq\ell}\theta_{u\ell'}\theta_{v\ell'} - \sum_{\ell}\theta_{i\ell}\theta_{j\ell}\sum_{\ell'}\theta_{u\ell'}\theta_{v\ell'}\Bigg\} + \sigma_\sigma^2\delta_{ijuv}\\ &= kc^2(x,x')\left\{\sum_{\ell}\theta_{i\ell}\theta_{j\ell}\theta_{u\ell}\theta_{v\ell} + \sum_{\ell}\theta_{i\ell}\theta_{u\ell}\sum_{\ell'}\theta_{j\ell'}\theta_{v\ell'}\right\} + \sigma_\sigma^2\delta_{ijuv}.\end{aligned}$$

An iterated expectation with respect to Θ yields the following results. When $i \neq u$ or $j \neq v$, the independence between $\theta_{i\ell}$ (or $\theta_{j\ell}$) and the set of other $\theta_{k\ell}$ implies that

$\mathrm{cov}(\Sigma_{ij}(x), \Sigma_{uv}(x')) = 0$. When $i = u$ and $j = v$, but $i \neq j$,

$$\begin{aligned}\mathrm{cov}(\Sigma_{ij}(x), \Sigma_{ij}(x')) &= kc^2(x,x') \left\{ \sum_{\ell} E[\theta_{i\ell}^2]E[\theta_{j\ell}^2] + \sum_{\ell} E[\theta_{i\ell}^2] \sum_{\ell'} E[\theta_{j\ell'}^2] \right\} \\ &= kc^2(x,x') \left\{ \sum_{\ell} \phi_{i\ell}^{-1}\phi_{j\ell}^{-1}\tau_\ell^{-2} + \sum_{\ell} \phi_{i\ell}^{-1}\tau_\ell^{-1} \sum_{\ell'} \phi_{j\ell'}^{-1}\tau_{\ell'}^{-1} \right\}.\end{aligned}$$

Finally, when $i = j = u = v$,

$$\begin{aligned}\mathrm{cov}(\Sigma_{ii}(x), \Sigma_{ii}(x')) &= kc^2(x,x') \left\{ 2\sum_{\ell} E[\theta_{i\ell}^4] + \sum_{\ell} E[\theta_{i\ell}^2] \sum_{\ell' \neq \ell} E[\theta_{i\ell'}^2] \right\} + \sigma_\sigma^2 \\ &= kc^2(x,x') \left\{ 6\sum_{\ell} \phi_{i\ell}^{-2}\tau_\ell^{-2} + \sum_{\ell} \phi_{i\ell}^{-1}\tau_\ell^{-1} \sum_{\ell' \neq \ell} \phi_{i\ell'}^{-1}\tau_{\ell'}^{-1} \right\} + \sigma_\sigma^2 \\ &= kc^2(x,x') \left\{ 5\sum_{\ell} \phi_{i\ell}^{-2}\tau_\ell^{-2} + (\sum_{\ell} \phi_{i\ell}^{-1}\tau_\ell^{-1})^2 \right\} + \sigma_\sigma^2.\end{aligned}$$

B. Derivation of Gibbs Sampler

In this Appendix, we derive the conditional distribution for sampling the Gaussian process dictionary elements. Combining Eq. (3) and Eq. (4), we have that

$$y_i = \Theta \begin{bmatrix} \xi_{11}(x_i) & \xi_{12}(x_i) & \dots & \xi_{1k}(x_i) \\ \xi_{21}(x_i) & \xi_{22}(x_i) & \dots & \xi_{2k}(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{L1}(x_i) & \xi_{L2}(x_i) & \dots & \xi_{Lk}(x_i) \end{bmatrix} \eta_i + \epsilon_i = \Theta \begin{bmatrix} \sum_{m=1}^k \xi_{1m}(x_i)\eta_{im} \\ \vdots \\ \sum_{m=1}^k \xi_{Lm}(x_i)\eta_{Lm} \end{bmatrix} + \epsilon_i \tag{58}$$

implying that

$$y_{ij} = \sum_{\ell=1}^{L} \sum_{m=1}^{k} \theta_{j\ell}\eta_{im}\xi_{\ell m}(x_i) + \epsilon_{ij}. \tag{59}$$

Conditioning on $\xi(\cdot)^{-\ell m}$, we rewrite Eq. (58) as

$$y_i = \eta_{im} \begin{bmatrix} \theta_{1\ell} \\ \vdots \\ \theta_{p\ell} \end{bmatrix} \xi_{\ell m}(x_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim \mathcal{N}\left(\mu_{\ell m}(x_i) \triangleq \begin{bmatrix} \sum_{(r,s)\neq(\ell,m)} \theta_{1r}\eta_{is}\xi_{rs}(x_i) \\ \vdots \\ \sum_{(r,s)\neq(\ell,m)} \theta_{pr}\xi_{rs}(x_i) \end{bmatrix}, \Sigma_0 \right). \tag{60}$$

Let $\theta_{\cdot\ell} = \begin{bmatrix} \theta_{1\ell} & \dots & \theta_{p\ell} \end{bmatrix}'$. Then,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \eta_{1m}\theta_{\cdot\ell} & 0 & \dots & 0 \\ 0 & \eta_{2m}\theta_{\cdot\ell} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_{nm}\theta_{\cdot\ell} \end{bmatrix} \begin{bmatrix} \xi_{\ell m}(x_1) \\ \xi_{\ell m}(x_2) \\ \vdots \\ \xi_{\ell m}(x_n) \end{bmatrix} + \begin{bmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \vdots \\ \tilde{\epsilon}_n \end{bmatrix} \tag{61}$$

Defining $A_{\ell m} = \mathrm{diag}(\eta_{1m}\theta_{\cdot\ell}, \dots, \eta_{nm}\theta_{\cdot\ell})$, our Gaussian process prior on the dictionary elements $\xi_{\ell m}(\cdot)$ implies the following conditional posterior

$$\begin{aligned}\begin{bmatrix}\xi_{\ell m}(x_1)\\ \xi_{\ell m}(x_2)\\ \vdots \\ \xi_{\ell m}(x_n)\end{bmatrix} \mid \{y_i\}, \Theta, \eta, \xi(\cdot), \Sigma_0 &\sim \mathcal{N}\left(\tilde{\Sigma}A'_{\ell m}\begin{bmatrix}\Sigma_0^{-1} & & \\ & \ddots & \\ & & \Sigma_0^{-1}\end{bmatrix}\begin{bmatrix}\tilde{y}_1\\ \vdots \\ \tilde{y}_n\end{bmatrix}, \tilde{\Sigma}\right)\\ &= \mathcal{N}\left(\tilde{\Sigma}\begin{bmatrix}\eta_{1m}\sum_{j=1}^p \theta_{j\ell}\sigma_j^{-2}\tilde{y}_{1j}\\ \vdots \\ \eta_{nm}\sum_{j=1}^p \theta_{j\ell}\sigma_j^{-2}\tilde{y}_{nj}\end{bmatrix}, \tilde{\Sigma}\right), \end{aligned} \tag{62}$$

where $\tilde{y}_i = y_i - \mu_{\ell m}(x_i)$ and, taking K to be the matrix of correlations $K_{ij} = c(x_i, x_j)$ defined by the Gaussian process parameter κ,

$$\tilde{\Sigma}^{-1} = K^{-1} + A'_{\ell m}\begin{bmatrix}\Sigma_0^{-1} & & \\ & \ddots & \\ & & \Sigma_0^{-1}\end{bmatrix}A_{\ell m} = K^{-1} + \mathrm{diag}\left(\eta_{1m}^2\sum_{j=1}^p \theta_{j\ell}^2\sigma_j^{-2}, \dots, \eta_{nm}^2\sum_{j=1}^p \theta_{j\ell}^2\sigma_j^{-2}\right). \tag{63}$$

References

R.P. Adams and I. Murray. Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, volume 23, 2011.

O. Aguilar and M. West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357, 2000.

J.H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

S. Banerjee, A.E. Gelfand, A.O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 70(4):825–848, 2008.

A. Bhattacharya and D.B. Dunson. Sparse Bayesian infinite factor models. *under revision at Biometrika*, 2010.

S.P. Brooks and A. Gelman. General methods for general methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7: 434–455, 1998.

C.M. Carvalho, J.E. Lucas, Q. Wang, J. Chang, J.R. Nevins, and M. West. High-dimensional sparse factor modelling - Applications in gene expression genomics. *Journal of the American Statistical Association*, 103, 2008.

S. Chib, Y. Omori, and M. Asai. Multivariate stochastic volatility. *Handbook of Financial Time Series*, pages 365–400, 2009.

T.Y.M. Chiu, T. Leonard, and K.W. Tsui. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210, 1996.

J. Du, H. Zhang, and V.S. Mandrekarm. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *the Annals of Statistics*, 37(6A):3330–3361, 2009.

V. Dukić, H.F. Lopes, and N.G. Polson. Tracking flu epidemics using google flu trends and particle learning. Technical report, University of Chicago Booth School of Business, 2009.

R. Engle. New frontiers for ARCH models. *Journal of Applied Econometrics*, 17(5):425–446, 2002.

A.E. Gelfand, A.M. Schmidt, S. Banerjee, and C.F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13(2):263–312, 2004.

J. Geweke and G. Zhou. Measuring the pricing error of the arbitrage pricing theory. *Review of Financial Studies*, 9(2):557–587, 1996.

J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2008.

C. Gouriéroux, J. Jasiak, and R. Sufana. The Wishart autoregressive process of multivariate stochastic volatility. *Journal of Econometrics*, 150(2):167–181, 2009.

A.C. Harvey, E. Ruiz, and N. Shephard. Multivariate stochastic variance models. *Review of Economic Studies*, 61:247–264, 1994.

P.D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1(1):265–283, 2007.

P.D. Hoff and X. Niu. A covariance regression model. Technical report, Department of Statistics, University of Washington, March 2010.

C.G. Kaufman, M.J. Schervish, and D.W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103 (484):1545–1555, 2008.

A. Kavcic and J.M.F. Moura. Matrix with banded inverses: Algorithms and factorization of Gauss-Markov processes. *IEEE Transactions on Information Theory*, 46(4):1495–1509, 2000.

H.F. Lopes and M. West. Bayesian model assessment in factor analysis. *Statistica Sinica*, 14(1):41–68, 2004.

H.F. Lopes, E. Salazar, and D. Gamerman. Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4):759–792, 2008.

A. Philipov and M.E. Glickman. Multivariate stochastic volatility via Wishart processes. *Journal of Business & Economic Statistics*, 24(3):313–328, 2006a.

A. Philipov and M.E. Glickman. Factor multivariate stochastic volatility via Wishart processes. *Econometric Reviews*, 25(2-3):311–334, 2006b.

M. Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

R. Prado and M. West. *Time Series: Modeling, Computation, and Inference*. Chapman & Hall / CRC, Boca Raton, FL, 2010.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society, Series B*, 71(2):319–392, 2009.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, 7:723–732, 2003.

H. Zhang and J. Du. Covariance tapering in spatial statistics. *Positive Definite Functions: From Schoenberg to Space-Time Challenges*, 181–196, 2008.