Journal of Machine Learning Research 16 (2015) 2501-2542 Submitted 2/15; Revised 7/15; Published 12/15

Bayesian Nonparametric Covariance Regression

Emily B. Fox EBFOX@STAT.WASHINGTON.EDU
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA

David B. Dunson DUNSON@STAT.DUKE.EDU
Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA

Editor: Edoardo M. Airoldi

Abstract

Capturing predictor-dependent correlations amongst the elements of a multivariate response vector is fundamental to numerous applied domains, including neuroscience, epidemiology, and finance. Although there is a rich literature on methods for allowing the variance in a univariate regression model to vary with predictors, relatively little has been done in the multivariate case. As a motivating example, we consider the Google Flu Trends data set, which provides indirect measurements of influenza incidence at a large set of locations over time (our predictor). To accurately characterize temporally evolving influenza incidence across regions, it is important to develop statistical methods for a time-varying covariance matrix. Importantly, the locations provide a redundant set of measurements and do not yield a sparse nor static spatial dependence structure. We propose to reduce dimensionality and induce a flexible Bayesian nonparametric covariance regression model by relating these location-specific trajectories to a lower-dimensional subspace through a latent factor model with predictor-dependent factor loadings. These loadings are in terms of a collection of basis functions that vary nonparametrically over the predictor space. Such low-rank approximations are in contrast to sparse precision assumptions, and are appropriate in a wide range of applications. Our formulation aims to address three challenges: scaling to large p domains, coping with missing values, and allowing an irregular grid of observations. The model is shown to be highly flexible, while leading to a computationally feasible implementation via Gibbs sampling. The ability to scale to large p domains and cope with missing values is fundamental in analyzing the Google Flu Trends data.

Keywords: covariance regression, dictionary learning, Gaussian process, latent factor model, nonparametric Bayes, time series

1. Introduction

Spurred by the increasing prevalence of high-dimensional data sets 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. Within neuroscience, there is interest in analyzing the time-varying coactivation patterns in brain activity, referred to as *functional connectivity*.

As a motivating example, we focus on the problem of modeling the changing correlations in flu activity amongst a large collection of regions in the United States as a function of time. The *Google Flu Trends* data set (available at `http://www.google.org/flutrends/`) provides estimates of flu activity in 183 regions on a weekly basis. The regions consist of the U.S. national level, 50 states, 10 regions, and 122 cities. A common strategy for modeling such data are Markov random fields (cf. Mugglin et al., 2002) (and relatedly, the kriging exploratory flu analysis of Sakai et al. (2004).) However, in addition to assuming (temporal) homoscedasticity, a limitation of such approaches is the typical reliance on a locally defined neighborhood structure that does not directly capture potential long-range dependencies (e.g., between New York and California.) Indeed, influenza spread can occur rapidly between non-contiguous regions (e.g., by air travel (Brownstein et al., 2006).) From exploratory data analysis, we find that the flu data does not yield a sparse graphical model structure. Instead, the redundancy between time series (e.g., Los Angeles and California) is naturally modeled via *low-rank approximations* that embed the observed trajectories in a low-dimensional subspace. Beyond its dimensionality, another challenge posed by this data set is the extent of missing data. For example, 25% of regions do not report data in the first year. The existing influenza modeling approaches described above rely on imputing such missing values, which we aim to avoid. The data attributes presented by the Google Flu Trends data set—redundancy in high dimensions, changing correlations, missing observations—are common to many applications.

In general terms, let $\boldsymbol{y} = (y_1, \ldots, y_p)' \in \Re^p$ denote a multivariate response and $\boldsymbol{x} = (x_1, \ldots, x_q)' \in \mathcal{X} \subset \Re^q$ an arbitrary multivariate predictor (e.g., *time*, *space*, etc.). In the flu analysis, p is the number of regions and $q = 1$ with $\boldsymbol{x}$ representing a scalar time index. A typical focus is on capturing the conditional mean $\mathrm{E}(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{\mu}(\boldsymbol{x})$, assuming a *homoscedastic* model with conditional covariance $\mathrm{cov}(\boldsymbol{y}|\boldsymbol{x}) = \Sigma$. Recall that this covariance matrix captures key correlations between the elements of the response vector (e.g., flu activity in the various regions). In our exploratory analysis of the flu data in Appendix G, the residuals from a smoothing spline fit indicate that a model of i.i.d. errors across time is inappropriate for this data. In such cases, an assumption of homoscedasticity can have significant ramifications on inferences (e.g., predictive accuracy) as we demonstrate in Sections 4 and 5.2.2. It is possible to decrease residual correlation through a more intricate mean model, but the complexities of doing so motivate us to instead turn to modeling the conditional covariance. In particular, our focus is on developing Bayesian methods that allow not only $\mathrm{E}(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{\mu}(\boldsymbol{x})$ but also $\mathrm{cov}(\boldsymbol{y}|\boldsymbol{x}) = \Sigma(\boldsymbol{x})$ to change flexibly with $\boldsymbol{x} \in \mathcal{X}$.

Classical strategies for estimating $\Sigma(\boldsymbol{x})$ rely on standard regression methods applied to the elements of the log or Cholesky decomposition of $\Sigma(\boldsymbol{x})$ or $\Sigma(\boldsymbol{x})^{-1}$ (Chiu et al., 1996; Pourahmadi, 1999; Leng et al., 2010; Zhang and Leng, 2012). This involves fitting $p(p+1)/2$ separate regression models, and hence these methods are ill-suited to high-dimensional applications due to the curse of dimensionality. Hoff and Niu (2012) instead proposed modeling $\Sigma(\boldsymbol{x})$ as a quadratic function of $\boldsymbol{x}$ plus a baseline positive definite matrix. The mapping from predictors to covariance assumes a parametric form, thus limiting the

model's expressivity. A nonparametric Nadaraya-Watson kernel estimator was proposed by Yin et al. (2010). Their approach is only appropriate for random $\boldsymbol{x}$ (i.e., not time series) and the kernel is required to be symmetric with a single bandwidth for all elements of $\Sigma(\boldsymbol{x})$. The result is a kernel estimator that may not be locally adaptive. For time series, heteroscedastic modeling has a long history (Chib et al., 2009), with the main approaches being multivariate generalized autoregressive conditional heteroscedasticity (GARCH) (Engle, 2002) (limited to applications with $p \leq 5$), multivariate stochastic volatility models (Harvey et al., 1994), and Wishart processes (Philipov and Glickman, 2006a,b; Gouriéroux et al., 2009). Central to the cited volatility models are assumptions of (i) Markov dynamics, limiting the ability to capture long-range dependencies, (ii) observation times that are equally spaced with no missing values, (iii) challenges in model fitting, and (iv) limited theory to justify flexibility.

We instead propose a Bayesian nonparametric approach to simultaneously modeling $\boldsymbol{\mu}(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$. Using low-rank approximations as a parsimonious modeling technique when p is not small, we consider latent factor models with *predictor-dependent factor loadings*. In particular, we characterize the loadings as a sparse combination of unknown basis functions, with Gaussian processes providing a convenient prior for basis elements varying nonparametrically over $\mathcal{X}$. The induced covariance is then a regularized quadratic function of these basis elements. The proposed approach is provably flexible and admits a latent variable representation with simple conjugate posterior updates, which facilitates tractable posterior computation in moderate to high dimensions. In addition to being able to state theoretical properties of our proposed prior—such as large support integral to a Bayesian nonparametric approach—the proposed methodology has numerous practical advantages over previous covariance regression frameworks:

1. *Scaling to high dimensions in the presence of limited data* (via structured latent factor models)
2. *Handling irregular grid of observations* (via continuous functions as basis elements)
3. *Tractable computations* (via simple conjugate posterior updates)
4. *Coping with ignorable missing data* (no data imputation required)
5. *Robustness to outlying observations* (via sharing information in the latent basis).

Importantly, our framework enables analytic marginalization of missing data from the complete data likelihood, and without introducing extra dependencies amongst the remaining variables. The benefits of this analytic marginalization are two-fold: (1) we do not spend computational resources imputing the missing values, and (2) compared to the otherwise dramatically increased Markov chain Monte Carlo (MCMC) state space that includes the missing values, we can improve convergence and mixing rates through marginalization (Liu et al., 1994). Combined with the model's flexible sharing of information via the latent basis functions, we are able to handle data sets with substantial missing data, such as in the flu application of Section 5. Finally, the Google Flu Trends estimates are based on user search queries, and as such are susceptible to the types of malicious attacks that Google regularly guards against in other domains. Our model is well-geared for handling some forms of these situations: the inherent redundancy and borrowing of information across locations provides robustness to limited amounts of inaccurate estimates. Note that these inaccurate estimates may not be malicious in nature, but instead represent outliers arising from unusual spurs in search activity and poorly calibrated models (Cook et al., 2011). As long as these errors do not form systematic or stochastic trends or explosive processes (Fuller, 2009), our model

appears to be robust, as we demonstrate in Section 5.2.3. This is in contrast to approaches that look at rates in individual locations (e.g., Dukić et al. (2012)).

An earlier version of this work appeared in a technical report (Fox and Dunson, 2011); the current version provides significant additions including revised proofs, an extended model presentation, new experiments on the Google Flu Trends data, and an extensive model assessment. The recent work of Durante et al. (2014) builds on our framework and has shown great promise, but with a focus on time series applications and without handling missing data or scaling to large p domains.

The paper is organized as follows. In Section 2, we describe our proposed Bayesian nonparametric covariance regression model and analyze 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 Flu Trends data set presented in Section 5.

2. Covariance Regression Priors

In this section, we consider the specific form for our Bayesian nonparametric covariance regression. Section 2.1 examines our assumed covariance structuring whereas Section 2.2 details our prior specification for the various model components.

2.1 Model Specification

We focus on a multivariate Gaussian nonparametric mean-covariance regression model

$$\boldsymbol{y}_i = \boldsymbol{\mu}(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}_p(0, \Sigma(\boldsymbol{x}_i)), \quad i = 1, \ldots, n, \tag{1}$$

with $\boldsymbol{x}_i \in \mathcal{X}$, $\mathcal{X}$ a compact subset of $\Re^q$, and the $\boldsymbol{\epsilon}_i$s independent. We focus on $\boldsymbol{x}$ non-random. In the flu application, $q = 1$ with $\{x_1, \ldots, x_n\}$ a set of week indices and $\boldsymbol{y}_i = \log \boldsymbol{r}_i$, the vector of log Google-estimated ILI rates in the 183 regions ($p = 183$) at time x_i. To cope with large p, we take model (1) to be induced through the factor model

$$\boldsymbol{y}_i = \Lambda(\boldsymbol{x}_i)\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim \mathrm{N}_k(\boldsymbol{\psi}(\boldsymbol{x}_i), I_k), \quad \boldsymbol{\epsilon}_i \sim \mathrm{N}_p(0, \Sigma_0) \tag{2}$$

where $\Lambda(\boldsymbol{x})$ is a $p \times k$ *factor loadings matrix* specific to predictor value $\boldsymbol{x}$, $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ik})'$ are *latent factors* associated with observation $\boldsymbol{y}_i$, and $\Sigma_0 = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

A latent factor model harnesses a lower-dimensional description of the observations, assuming $k \ll p$. $\boldsymbol{\psi}(\boldsymbol{x})$ captures the evolution of the latent factors whereas $\Lambda(\boldsymbol{x})$ dictates a low-rank evolution to the conditional covariance of the response vector. In particular, marginalizing out $\boldsymbol{\eta}_i$, the mean and covariance regression models are expressed as

$$\boldsymbol{\mu}(\boldsymbol{x}) = \Lambda(\boldsymbol{x})\boldsymbol{\psi}(\boldsymbol{x}), \quad \Sigma(\boldsymbol{x}) = \Lambda(\boldsymbol{x})\Lambda(\boldsymbol{x})' + \Sigma_0. \tag{3}$$

To make this concrete, in our flu application, $\boldsymbol{\eta}_i$ captures a small latent set of flu responses (not necessarily standard ILI rates) at week i, $\boldsymbol{\psi}(x)$ the evolution of these latent responses, and $\Lambda(x_i)$ a low-rank description of the spatial correlations at week i. The motivation for modeling the mean as in (3) arises from a desire to have a parsimonious model in large p domains. This is in contrast to, for example, a model $\boldsymbol{y}_i = \Lambda(\boldsymbol{x}_i)\boldsymbol{\eta}_i + \boldsymbol{\mu}(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$ where $\boldsymbol{\eta}_i \sim N_k(0, I_k)$ and $\boldsymbol{\mu}(\boldsymbol{x}_i)$ is a p-dimensional mean regression.

Before specifying our priors for each of the components in (2), we first place our formulation within the context of dynamic latent factor models.

2.1.1 Relationship to Dynamic Latent Factor Models

A standard latent factor model characterizes independent observations $\boldsymbol{y}_i$ via independent latent factors $\boldsymbol{\eta}_i$:

$$\boldsymbol{y}_i = \Lambda\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\eta}_i \sim N_k(0, I_k), \quad \boldsymbol{\epsilon}_i \sim N_p(0, \Sigma_0). \tag{4}$$

Marginalizing the latent factors $\boldsymbol{\eta}_i$ yields $\boldsymbol{y}_i \sim N_p(0, \Sigma)$ with $\Sigma = \Lambda\Lambda' + \Sigma_0$. The ideas of latent factor analysis have also been applied to the time-series domain by assuming a latent factor *process*. Such *dynamic latent factor models* have a rich history. Typically, the dynamics of the latent factors are assumed to follow a simple Markov evolution with a time-invariant parameterization (West, 2003; Lopes et al., 2008):

$$\begin{aligned} \boldsymbol{\eta}_i &= \Gamma\boldsymbol{\eta}_{i-1} + \boldsymbol{\nu}_i, \quad \boldsymbol{\nu}_i \sim N_k(0, I_k) \\ \boldsymbol{y}_i &= \Lambda\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim N_p(0, \Sigma_0), \end{aligned} \tag{5}$$

where $\Gamma \in \Re^{k\times k}$ is the dynamic matrix for the latent factor evolution. Assuming a stationary process on $\boldsymbol{\eta}_i$, then $\boldsymbol{y}_i \sim N_p(0, \Sigma)$ with $\Sigma = \Lambda\Sigma_\eta\Lambda' + \Sigma_0$. Here, Σ_η denotes the marginal covariance of $\boldsymbol{\eta}_i$. If we restrict our attention to cases in which $\boldsymbol{x}_i$ is a discrete time index, as in our flu application, then our proposed model of (2) can be related to the class of dynamic latent factor models as follows. The latent factor evolution is governed by $\boldsymbol{\psi}$ rather than a standard linear autoregression: $\boldsymbol{\eta}_i = \boldsymbol{\psi}(x_i) + \boldsymbol{\nu}_i$, $\boldsymbol{\nu}_i \sim N_k(0, I_k)$. In Section 2.2, we specify $\boldsymbol{\psi}$ via Gaussian processes, providing a nonparametric evolution in *continuous* time. Importantly, the factor loadings matrix $\Lambda(x)$ also evolves in time: $\boldsymbol{y}_i = \Lambda(x_i)\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i$ with conditional covariance $\Sigma(x) = \Lambda(x)\Lambda(x)' + \Sigma_0$. Again, this analogy relies on assuming $\boldsymbol{x}$ represents time. The formulation of (2) is proposed for general predictors $\boldsymbol{x} \in \mathcal{X}$.

2.2 Prior specification

To capture the evolution of $\boldsymbol{\psi}(\boldsymbol{x})$ and $\Lambda(\boldsymbol{x})$, we use Gaussian processes as a set of basis functions. We first briefly review Gaussian processes and then describe how this basis is used in our model.

2.2.1 Gaussian Processes

A Gaussian process provides a distribution over real-valued functions $f : \mathcal{X} \to \Re$, with the property that the function evaluated at any finite collection of points is jointly Gaussian. The Gaussian process, denoted $\mathrm{GP}(m, c)$, is uniquely defined by its *mean function* m and *covariance kernel* c. In particular, $f \sim \mathrm{GP}(m, c)$ if and only if for all n and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$,

$$p(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)) \sim N_n(\boldsymbol{\mu}, K), \tag{6}$$

with $\boldsymbol{\mu} = [m(\boldsymbol{x}_1), \ldots, m(\boldsymbol{x}_n)]$ and K the $n \times n$ *Gram matrix* with entries $K_{ij} = c(\boldsymbol{x}_i, \boldsymbol{x}_j)$. The properties (e.g., continuity, smoothness, periodicity, etc.) of functions drawn from a

given Gaussian process are determined by the covariance kernel. One example leading to smooth functions is the squared exponential, or *Gaussian*, kernel:

$$c(\boldsymbol{x}, \boldsymbol{x}') = d \exp(-\kappa ||\boldsymbol{x} - \boldsymbol{x}'||_2^2), \tag{7}$$

where d is a *scale* hyperparameter and κ the *bandwidth*, which determines the extent of the correlation in f over $\mathcal{X}$. See Rasmussen and Williams (2006) for further details.

2.2.2 Latent Factor Mean Process

Letting $\boldsymbol{\psi}(\boldsymbol{x}) = \{\psi_1(\boldsymbol{x}), \ldots, \psi_k(\boldsymbol{x})\}$, we specify independent Gaussian process priors for each ψ_h as a convenient and flexible choice. In particular, $\psi_h \sim \text{GP}(0, c_\psi)$ with $c_\psi(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\kappa_\psi ||\boldsymbol{x} - \boldsymbol{x}'||_2^2)$ a squared exponential covariance kernel. We assume unit variance for reasons of identifiability seen in (3) through the multiplication of the latent factors with $\Lambda(\boldsymbol{x})$.

2.2.3 Idiosyncratic Noise

We choose independent inverse gamma priors for the diagonal elements of Σ_0 by letting $\sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$. The off-diagonal elements are deterministically set to zero.

2.2.4 Factor Matrix Process

Specifying a prior for $\Lambda(\boldsymbol{x})$ is more challenging, as naive approaches, such as independent Gaussian process priors for each element of the $p \times k$ matrix, may have poor performance in large p application domains even for small k. Likewise, the computational demands for considering $p \times k$ Gaussian processes can be prohibitive depending on the choice of p, k, n (see Section 3). Instead, we take the factor loadings to be a weighted combination of a much smaller set of basis elements ξ_{lh},

$$\Lambda(\boldsymbol{x}) = \Theta \xi(\boldsymbol{x}), \quad \Theta \in \Re^{p \times L}, \quad \xi(\boldsymbol{x}) = \{\xi_{lh}(\boldsymbol{x}), l = 1, \ldots, L, h = 1, \ldots, k\}, \tag{8}$$

where Θ is a matrix of coefficients that maps the $L \times k$ array of basis functions $\xi(\boldsymbol{x})$ to the predictor-dependent loadings matrix $\Lambda(\boldsymbol{x})$. Typically, $k \ll p$ and $L \ll p$. Again, k defines the factor dimension (i.e., assumed subspace that captures the statistical variability) whereas L controls the size of the basis for any fixed choice of k. We once again choose independent Gaussian process priors $\xi_{lh} \sim \text{GP}(0, c)$, with $c(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\kappa ||\boldsymbol{x} - \boldsymbol{x}'||_2^2)$ a squared exponential covariance kernel. The choice of unit variance Gaussian processes again arises for reasons of identifiability, but now with the multiplication with Θ.

To allow for an adaptive choice of the basis size, we in theory let $L \to \infty$ and employ the shrinkage prior of Bhattacharya and Dunson (2011) for Θ,

$$\theta_{jl} \sim \text{N}(0, \phi_{jl}^{-1} \tau_l^{-1}), \quad \phi_{jl} \sim \text{Ga}(\gamma/2, \gamma/2), \quad \tau_l = \prod_{h=1}^{l} \delta_h, \tag{9}$$

with ϕ_{jl} a local precision specific in element j, l, and τ_l a column-specific multiplier, which is assigned a multiplicative gamma process prior to favor increasing shrinkage of elements in later columns by letting $\delta_1 \sim \text{Ga}(a_1, 1)$ and $\delta_h \sim \text{Ga}(a_2, 1)$, $h \geq 2$, with $a_2 > 1$. If a column

of Θ is shrunk towards zero, the corresponding row of the basis $\xi(\boldsymbol{x})$ has insignificant effect in defining $\{\boldsymbol{\mu}(\boldsymbol{x}), \Sigma(\boldsymbol{x})\}$. Our chosen prior specification increasingly shrinks columns with column index, effectively truncating Θ. That is, despite an arbitrarily large L, the effective dimension of the basis is much smaller, providing our desired dimensionality reduction. In practice, of course, a finite truncation $\bar{L}$ is chosen. See Appendix E for a discussion on other possible decompositions of $\Lambda(\boldsymbol{x})$ and prior specifications.

At any point $\boldsymbol{x} \in \mathcal{X}$, the different $\xi_{lk}(\boldsymbol{x})$s are independently Gaussian distributed, and hence $\xi(\boldsymbol{x})\xi(\boldsymbol{x})'$ is Wishart distributed. Conditioned on Θ, $\Theta\xi(\boldsymbol{x})\xi(\boldsymbol{x})'\Theta'$ is also Wishart distributed and, as $\boldsymbol{x}$ varies, follows the matrix-variate Wishart process of Gelfand et al. (2004) with Wilson and Ghahramani (2011) recently considering a related specification. However, these alternative specifications do not have the dimensionality reduction structure, which is key to the performance of our approach in moderate to high dimensions. Furthermore, they do not provide the theoretical statements of large support we show in Section 2.4 nor a framework for coping with missing data. Marginalizing over the prior for Θ, one obtains a type of adaptively scaled mixture of Wishart processes that has fundamentally different behavior than the Wishart. Our prior is also somewhat related to the spatial dynamic factor model of Lopes et al. (2008), though their focus is on space-time dependence in univariate observations. Finally, following our early technical report version of this paper (Fox and Dunson, 2011), Fosdick and Hoff (2014) examine factor-structured separable covariance models for general M-array data. Considering a 2-array of space and time, the model assumes a spatial structure $\Lambda_s\Lambda_s' + \Sigma_{0,s}$ and temporal structure $\Lambda_t\Lambda_t' + \Sigma_{0,t}$. That is, the model is low rank in both space and time. In contrast, our covariance decomposition at any predictor $\boldsymbol{x}$ assumes the factor structure $\Lambda(\boldsymbol{x})\Lambda(\boldsymbol{x})' + \Lambda_0$ for the response vector (e.g., indexed by spatial location); however, the dependence between predictors $\boldsymbol{x}$ and $\boldsymbol{x}'$ (e.g., across time) is described via a stochastic *process*.

2.2.5 Identifiability

The factorizations for $\boldsymbol{\mu}(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$ are not unique but instead we obtain a many-to-one specification. It is not necessary to enforce identifiability constraints, as our focus is on inducing a prior for $\boldsymbol{\mu}(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$ that favors an effectively low-dimensional representation without constraining the possible changes in the mean and covariance with predictors beyond minimal regularity conditions.

2.3 Parsimony of Covariance Decomposition

Through the chosen covariance decomposition $\Sigma(\boldsymbol{x}) = \Theta\xi(\boldsymbol{x})\xi(\boldsymbol{x})'\Theta' + \Sigma_0$ specified in (3) and (8), we have transformed the problem of modeling $p(p+1)/2$ predictor dependent elements to one of modeling $p \times (L+1)$ non-predictor dependent elements (comprising Θ and Σ_0) plus $L \times k$ predictor dependent elements (comprising $\xi(\cdot)$). A substantial reduction in parameterization occurs when $k \ll p$ and $L \ll p$. Such an assumption is appropriate in modeling a large class of covariance regressions $\Sigma(\boldsymbol{x})$ that arise when analyzing real data.

For arbitrary Θ and $\xi(\cdot)$, this parameterization still scales poorly to large data sets. It is only through the implied regularization effect of our chosen prior specification that a parsimonious model arises (even for L large, as previously discussed). Specifically, the continuity of the latent Gaussian process basis elements $\xi_{\ell k}$ combined with the shrinkage

properties of the prior on Θ forms a flexible, adaptive hierarchical structure that borrows information and can collapse on an effectively lower dimensional structure.

Another important aspect of the covariance decomposition is the implied *transfer of knowledge* property that allows us to cope with substantial missing or corrupted data. Let x_m correspond to a point in the predictor space at which the jth response component y_{mj} is missing or corrupted. In our model, the estimates of $\Sigma_{j\cdot}(\boldsymbol{x}_m) = \Theta_{j\cdot}\xi(\boldsymbol{x}_m)\xi(\boldsymbol{x}_m)'\Theta' + \Sigma_{0j\cdot}$ are improved by the fact that (i) the rows $\Theta_{j\cdot}$ are informed by all available observations y_{ij} at predictor locations $\boldsymbol{x}_i \neq \boldsymbol{x}_m$, and (ii) the latent basis functions $\xi(\boldsymbol{x}_m)$ are informed by the available response components y_{mk}, $k \neq j$, at the predictor location $\boldsymbol{x}_m$ and at nearby locations via the continuity of the basis functions. By employing a small collection of latent basis elements with non-predictor-dependent weights, our model better copes with limited data and is more robust to corrupted values than one in which the elements of $\Sigma(\boldsymbol{x})$ are modeled independently.

2.4 Properties

Our proposed Bayesian nonparametric covariance regression framework of Section 2 yields various important theoretical properties, such as large prior support and stationarity, which we examine here.

2.4.1 Large Support

We induce a prior $\{\Sigma(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\} \sim \Pi_\Sigma$ through priors Π_ξ, Π_Θ and Π_{Σ_0} for ξ, Θ and Σ_0, respectively. In this section, we explore the properties of the induced prior Π_Σ. Most fundamentally, we establish that this prior has large support in Theorem 2. Large support implies that the prior can generate a covariance regression function $\Sigma : \mathcal{X} \to \mathcal{P}_p^+$ arbitrarily close to any continuous function $\Sigma^* : \mathcal{X} \to \mathcal{P}_p^+$, with $\mathcal{P}_p^+$ the space of $p \times p$ positive semidefinite matrices. 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. The proofs associated with the theoretical statements made in this section can be found in Appendix A.

We start by introducing a notion of *k-decomposability* of a covariance regression function $\Sigma(\boldsymbol{x})$.

Definition 1 $\Sigma : \mathcal{X} \to \mathcal{P}_p^+$ *is said to be* **k-*decomposable*** *if* $\Sigma(\boldsymbol{x}) = \Lambda(\boldsymbol{x})\Lambda(\boldsymbol{x})' + \Sigma_0$ *for* $\Lambda(\boldsymbol{x}) \in \Re^{p\times k}$, $\Sigma_0 \in \mathcal{X}_{\Sigma_0}$, *and for all* $\boldsymbol{x} \in \mathcal{X}$.

In Appendix A, we show that such a decomposition always exists for k sufficiently large. Now assume our model $\Sigma(\boldsymbol{x}) = \Theta\xi(\boldsymbol{x})\xi(\boldsymbol{x})'\Theta' + \Sigma_0$ with priors Π_ξ and Π_{Σ_0} as specified in Section 2.2. For Π_Θ, we aim to make our statement of prior support as general as possible and thus simply assume that Π_Θ satisfies the following two conditions. The proof that Assumptions 2.1 and 2.2 are satisfied by our shrinkage prior (9) is provided in Appendix A.

Assumption 2.1 Π_Θ *is such that* $\sum_\ell E(|\theta_{j\ell}|) < \infty$, *ensuring that the prior for* 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.*

Our main result on prior support now follows.

Theorem 2 *Let Π_Σ denote the induced prior on $\{\Sigma(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\}$ based on $\Pi_\xi \otimes \Pi_\Theta \otimes \Pi_{\Sigma_0}$, with Π_Θ satisfying Assumptions 2.1 and 2.2. Assume $\mathcal{X}$ is compact. Then, for all continuous functions $\Sigma^* : \mathcal{X} \to \mathcal{P}_p^+$ that are k^*-decomposable and for all $\epsilon > 0$ and $k \geq k^*$,*

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

Informally, Theorem 2 states that there is positive prior probability of random covariance regressions $\Sigma(\boldsymbol{x})$ that stay within an L_2 ϵ-ball of any specified continuous $\Sigma^*(\boldsymbol{x})$ everywhere over the predictor space $\mathcal{X}$. Intuitively, the support on continuous covariance functions $\Sigma^*(\boldsymbol{x})$ arises from the continuity of the Gaussian process basis functions. However, since we are mixing over infinitely many such basis 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. We also rely on the large support of Π_Σ at any point $\boldsymbol{x}_0 \in \mathcal{X}$. Combining the large support of the Wishart distribution for $\Theta\xi(\boldsymbol{x}_0)\xi(\boldsymbol{x}_0)'\Theta'$ (Θ fixed) 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 $\boldsymbol{x}_0$.

Remark 3 *Our theory holds for $L \to \infty$ (an arbitrarily large set of latent basis functions); however, our large support result only relies on choosing $L = p$. Assuming Σ^* is k^*-decomposable with $k^* \ll p$ such that we can select $k \ll p$, this still represents a reduction in parameterization relative to a full model necessitating $p \times p$ basis functions. The reliance on $L = p$ is to be able to capture* any *k^*-decomposable Σ^*. We can further introduce a concept of* **L-decomposability** *where $\Sigma(\boldsymbol{x}) = \Lambda(\boldsymbol{x})\Lambda(\boldsymbol{x})' + \Sigma_0$ with $\Lambda(\boldsymbol{x}) = \Theta\xi(\boldsymbol{x})$ for $\Theta \in \Re^{p \times L}$ and for all $\boldsymbol{x} \in \mathcal{X}$, which represents a second factor assumption. Assuming $L \ll p$ is likely reasonable for large p. Then for a (k^*, L^*)-decomposable Σ^*, and choosing $k > k^*$ and $L > L^*$ (rather than relying on $L = p$), the theory of large support follows straightforwardly.*

Even when selecting $L = p$, due to our shrinkage prior for Θ of Section 2.2.4, we find in practice that many columns tend to be shrunk to zero *a posteriori* such that choosing a truncation $\bar{L} \ll p$ suffices. See Sections 4 and 5.2.4.

2.4.2 Moments and Stationarity

To better understand the relationship between our hyperparameter settings and resulting covariance regressions, it is useful to analyze the moments of $\{\Sigma(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\} \sim \Pi_\Sigma$. Lemma 4 provides the prior mean and Lemma 5 the covariance between elements of $\Sigma(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x}')$. As the distance between $\boldsymbol{x}$ and $\boldsymbol{x}'$ increases, the correlation decreases at a rate depending on the Gaussian process covariance kernel $c(\boldsymbol{x}, \boldsymbol{x}')$.

Lemma 4 *Let μ_σ denote the mean of σ_j^2, $j = 1, \ldots, p$. Then,*

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

That is, the expected covariance at $\boldsymbol{x}$ is diagonal with expected variance elements depending on our latent dimension k.

Lemma 5 *Let σ^2_σ denote the variance of σ^2_j, $j = 1, \ldots, p$. Then,*

$$cov(\Sigma_{ij}(\boldsymbol{x}), \Sigma_{ij}(\boldsymbol{x}')) = \begin{cases} k\, c(\boldsymbol{x}, \boldsymbol{x}') \left(5 \sum_\ell \phi_{i\ell}^{-2} \tau_\ell^{-2} + (\sum_\ell \phi_{i\ell}^{-1} \tau_\ell^{-1})^2\right) + \sigma^2_\sigma & i = j, \\ k\, c(\boldsymbol{x}, \boldsymbol{x}') \left(\sum_\ell \phi_{i\ell}^{-1} \phi_{j\ell}^{-1} \tau_\ell^{-2} + \sum_\ell \phi_{i\ell}^{-1} \tau_\ell^{-1} \sum_{\ell'} \phi_{j\ell'}^{-1} \tau_{\ell'}^{-1}\right) & i \neq j. \end{cases} \tag{10}$$

For $\Sigma_{ij}(\boldsymbol{x})$ and $\Sigma_{uv}(\boldsymbol{x}')$ with $i \neq u$ or $j \neq v$, $cov(\Sigma_{ij}(\boldsymbol{x}), \Sigma_{uv}(\boldsymbol{x}')) = 0$.

Here, we see how our Gaussian process covariance function $c(\boldsymbol{x}, \boldsymbol{x}')$ controls the dependence over $\mathcal{X}$ in an interpretable, linear fashion.

From Lemma 5, the autocorrelation $ACF(\boldsymbol{x}) = \text{corr}(\Sigma_{ij}(0), \Sigma_{ij}(\boldsymbol{x}))$ is simply specified by $c(0, \boldsymbol{x})$. When we choose a Gaussian process kernel $c(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\kappa ||\boldsymbol{x} - \boldsymbol{x}'||_2^2)$, we have

$$ACF(\boldsymbol{x}) = \exp(-\kappa ||\boldsymbol{x}||_2^2). \tag{11}$$

Thus, the length-scale parameter κ directly determines the shape of the autocorrelation function. This property aids in the selection of κ via a data-driven mechanism (i.e., a quasi-empirical Bayes approach), as outlined in Appendix C. One can also consider selecting κ using methods akin to those proposed by Higdon et al. (2008); Paulo (2005).

Finally, Lemma 6 shows that the stochastic process Σ has stationarity properties, an often desirable property of a covariance process specification since Σ itself captures heteroscedasticity in the observation process.

Lemma 6 *The process $\{\Sigma(\boldsymbol{x}), \boldsymbol{x} \in \mathcal{X}\} \sim \Pi_\Sigma$ is first-order stationary in that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $\Pi_\Sigma(\Sigma(\boldsymbol{x})) = \Pi_\Sigma(\Sigma(\boldsymbol{x}'))$. Furthermore, assuming a stationary covariance function $c(\boldsymbol{x}, \boldsymbol{x}')$, the process is wide sense stationary: $cov(\Sigma_{ij}(\boldsymbol{x}), \Sigma_{uv}(\boldsymbol{x}'))$ solely depends upon $||\boldsymbol{x} - \boldsymbol{x}'||$.*

3. Posterior Computation via Gibbs Sampling

Based on a fixed truncation level $\bar{L}$ and a latent factor dimension $\bar{k}$, we propose a Gibbs sampler for posterior computation. For the model of Section 2, the full joint probability is given by $p_{obs} \cdot p_{params} \cdot p_{hypers}$ where

$$\begin{aligned} p_{obs} &= \prod_{i=1}^n \left[p(\boldsymbol{y}_i \mid \Theta, \xi, \boldsymbol{\eta}_i, \Sigma_0) \prod_{k=1}^{\bar{k}} p(\eta_{ik} \mid \psi_k) \right] \quad p_{hypers} = \prod_{\ell=1}^{\bar{L}} \left[p(\tau_\ell) \prod_{j=1}^p p(\phi_{j\ell}) \right] \\ p_{params} &= \prod_{k=1}^{\bar{k}} \left[p(\psi_k) \prod_{\ell=1}^{\bar{L}} p(\xi_{\ell k}) \right] \prod_{j=1}^p \left[p(\sigma_j^2) \prod_{\ell=1}^{\bar{L}} p(\theta_{j\ell} \mid \phi_{j\ell}, \tau_\ell) \right] \end{aligned} \tag{12}$$

The resulting sampler is outlined in Steps 1-5 below. Step 1 is derived in Appendix B. In this section, we equivalently represent the latent factor process of (2) as $\boldsymbol{\eta}_i = \boldsymbol{\psi}(\boldsymbol{x}_i) + \boldsymbol{\nu}_i$, with $\boldsymbol{\nu}_i \sim \text{N}_k(0, I_k)$.

Step 1. Update each basis function $\xi_{\ell m}$ 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 $y_{ij} = \sum_{m=1}^{\bar{k}} \eta_{im} \sum_{\ell=1}^{\bar{L}} \theta_{j\ell}\xi_{\ell m}(x_i) + \epsilon_{ij}$. Conditioning on $\xi^{-\ell m} = \{\xi_{rs}, r \neq \ell, s \neq m\}$,

$$\begin{pmatrix} \xi_{\ell m}(\boldsymbol{x}_1) \\ \vdots \\ \xi_{\ell m}(\boldsymbol{x}_n) \end{pmatrix} \mid \{\boldsymbol{y}_i\}, \{\boldsymbol{\eta}_i\}, \Theta, \xi^{-\ell m}, \Sigma_0 \sim \mathrm{N}_n \left(\tilde{\Sigma}_\xi \begin{pmatrix} \eta_{1m} \sum_{j=1}^p \theta_{j\ell}\sigma_j^{-2}\tilde{y}_{1j} \\ \vdots \\ \eta_{nm} \sum_{j=1}^p \theta_{j\ell}\sigma_j^{-2}\tilde{y}_{nj} \end{pmatrix}, \tilde{\Sigma}_\xi \right),$$

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

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

Step 2. Sample each latent factor mean function ψ_l. Letting $\Omega_i = \Theta\xi(\boldsymbol{x}_i)$, we have $\boldsymbol{y}_i = \Omega_i\boldsymbol{\psi}(\boldsymbol{x}_i) + \Omega_i\boldsymbol{\nu}_i + \boldsymbol{\epsilon}_i$. Marginalizing out $\boldsymbol{\nu}_i$, $\boldsymbol{y}_i = \Omega_i\boldsymbol{\psi}(\boldsymbol{x}_i) + \boldsymbol{\omega}_i$ with $\boldsymbol{\omega}_i \sim \mathrm{N}(0, \tilde{\Sigma}_i = \Omega_i\Omega_i' + \Sigma_0)$. Assuming $\psi_\ell \sim \mathrm{GP}(0, c)$, the posterior of ψ_ℓ follows analogously to that of $\xi_{\ell m}$ resulting in

$$\begin{pmatrix} \psi_l(\boldsymbol{x}_1) \\ \vdots \\ \psi_l(\boldsymbol{x}_n) \end{pmatrix} \mid \{\boldsymbol{y}_i\}, \{\boldsymbol{\eta}_i\}, \boldsymbol{\psi}^{-l}, \Theta, \xi, \Sigma_0 \sim \mathrm{N}_n \left(\tilde{\Sigma}_\psi \begin{pmatrix} \Omega_{1l}'\tilde{\Sigma}_1^{-1}\tilde{\boldsymbol{y}}_1^{-l} \\ \vdots \\ \Omega_{nl}'\tilde{\Sigma}_n^{-1}\tilde{\boldsymbol{y}}_n^{-l} \end{pmatrix}, \tilde{\Sigma}_\psi \right),$$

where $\tilde{\boldsymbol{y}}_i^{-l} = \boldsymbol{y}_i - \sum_{(r\neq l)} \Omega_{ir}\boldsymbol{\psi}_r(\boldsymbol{x}_i)$, Ω_{il} is the lth column vector of Ω_i, and

$$\tilde{\Sigma}_\psi^{-1} = K^{-1} + \mathrm{diag}\left(\Omega_{1l}'\tilde{\Sigma}_1^{-1}\Omega_{1l}, \ldots, \Omega_{nl}'\tilde{\Sigma}_n^{-1}\Omega_{nl} \right).$$

Step 3. Sample $\boldsymbol{\nu}_i$. Defining $\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i - \Omega_i\boldsymbol{\psi}(\boldsymbol{x}_i)$ such that $\tilde{\boldsymbol{y}}_i = \Omega_i\boldsymbol{\nu}_i + \boldsymbol{\epsilon}_i$, we draw $\boldsymbol{\nu}_i$ given $\tilde{\boldsymbol{y}}_i, \boldsymbol{\psi}(\boldsymbol{x}_i), \Theta, \xi(\boldsymbol{x}_i), \Sigma_0$ from the conditional posterior,

$$\mathrm{N}_{\bar{k}} \left(\left\{ I + \xi(\boldsymbol{x}_i)'\Theta'\Sigma_0^{-1}\Theta\xi(\boldsymbol{x}_i) \right\}^{-1} \xi(\boldsymbol{x}_i)'\Theta'\Sigma_0^{-1}\tilde{\boldsymbol{y}}_i, \left\{ I + \xi(\boldsymbol{x}_i)'\Theta'\Sigma_0^{-1}\Theta\xi(\boldsymbol{x}_i) \right\}^{-1} \right).$$

Step 4. Sample σ_j^2. Letting $\theta_{j\cdot}$ denote the jth row vector of Θ, we draw

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

Step 5. Sample $\theta_{j\cdot}$. The conditional posterior on the row vectors of Θ is

$$\theta_{j\cdot} \mid \{\boldsymbol{y}_i\}, \{\boldsymbol{\eta}_i\}, \xi, \phi, \tau \sim \mathrm{N}_{\bar{L}} \left(\tilde{\Sigma}_\theta\tilde{\boldsymbol{\eta}}'\sigma_j^{-2}(y_{1j}, \ldots, y_{nj})', \tilde{\Sigma}_\theta \right),$$

where $\tilde{\boldsymbol{\eta}} = \{\xi(\boldsymbol{x}_1)\boldsymbol{\eta}_1, \ldots, \xi(\boldsymbol{x}_n)\boldsymbol{\eta}_n\}'$ and $\tilde{\Sigma}_\theta^{-1} = \sigma_j^{-2}\tilde{\boldsymbol{\eta}}'\tilde{\boldsymbol{\eta}} + \mathrm{diag}(\phi_{j1}\tau_1, \ldots, \phi_{j\bar{L}}\tau_{\bar{L}})$.

Step 6. Finally, for the hyperparameters in the shrinkage prior for Θ, we have

$$\phi_{jl} \mid \theta_{jl}, \tau_l \sim \mathrm{Ga}\left(2, \frac{\gamma + \tau_l\theta_{jl}^2}{2} \right)$$

$$\delta_1 \mid \Theta, \tau^{(-1)} \sim \text{Ga}\left(a_1 + \frac{p\bar{L}}{2}, 1 + \frac{1}{2}\sum_{l=1}^{\bar{L}} \tau_l^{(-1)} \sum_{j=1}^{p} \phi_{j\ell}\theta_{jl}^2\right)$$

$$\delta_h \mid \Theta, \tau^{(-h)} \sim \text{Ga}\left(a_2 + \frac{p(\bar{L}-h+1)}{2}, 1 + \frac{1}{2}\sum_{l=1}^{\bar{L}} \tau_l^{(-h)} \sum_{j=1}^{p} \phi_{jl}\theta_{jl}^2\right),$$

where $\tau_l^{(-h)} = \prod_{t=1,t\neq h}^{l} \delta_t$ for $h = 1, \ldots, p$.

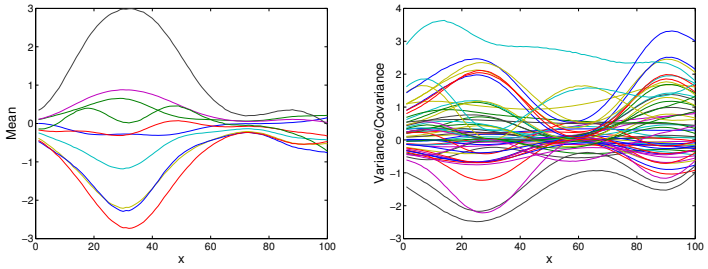
Each of the above steps is straightforward to implement involving sampling from standard distributions. We have observed good rates of convergence and mixing in our considered applications (see Section 5). As with other models involving Gaussian processes, computational bottlenecks can arise as n increases due to $O(n^3)$ matrix computation. Standard computational approaches can be used for dealing with this problem, as discussed in Section 6. We find inferences to be somewhat robust to the Gaussian process covariance parameter κ due the quadratic mixing over the basis functions. In the applications described below, we estimate κ from the data as an empirical Bayes approach, with details in Appendix C.

4. Simulation Example

We assess the performance of the proposed approach in terms of both covariance estimation and predictive performance. In Case 1 we simulated from the proposed model, while in Case 2 we simulated from a parametric model. In Case 1, we let $\mathcal{X} = \{1, \ldots, 100\}$, $p = 10$, $L = 5$, $k = 4$, $a_1 = a_2 = 10$, $\gamma = 3$, $a_\sigma = 1$, $b_\sigma = 0.1$ and $\kappa_\psi = \kappa = 10$ in the Gaussian process after scaling $\mathcal{X}$ to $(0, 1]$ with an additional nugget of $1e^{-5}I_n$ added to K. Figure 1 displays the resulting values of the elements of $\boldsymbol{\mu}(x)$ and $\Sigma(x)$. For inference, we use truncation levels $\bar{k} = \bar{L} = 10$, 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 $a_1 = a_2 = 2$, $\gamma = 3$, and placed a $\text{Ga}(1, 0.1)$ prior on the precision parameters σ_j^{-2}. The length-scale parameter κ was set from the data according to the heuristic described in Appendix C, and was determined to be 10 (after rounding). Details on initialization are available in Appendix D. We simulated 10,000 Gibbs iterations, discarded the first 5,000 and saved every 10th iteration.

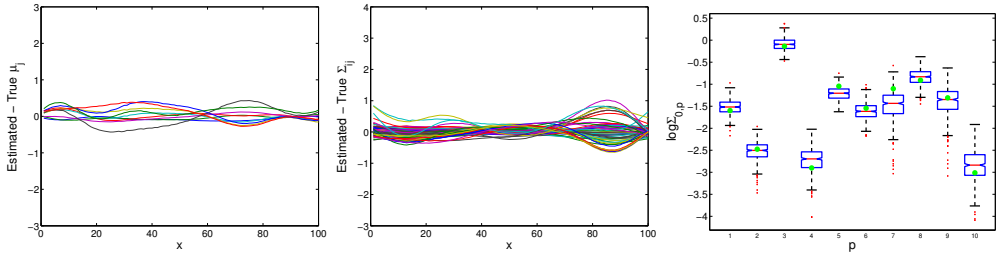
The residuals between the true and posterior mean over all components are displayed in Figure 2(a) and (b). Figure 2(c) compares the posterior samples of the elements σ_j^2 of the residual covariance Σ_0 to the true values. In Figure 3 we display a select set of plots of the true and posterior mean of components of $\boldsymbol{\mu}(x)$ and $\Sigma(x)$, along with the 95% highest posterior density intervals computed pointwise. From 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 all contained within the 95% highest posterior density intervals, with these intervals typically narrow.

For the same simulated data set, we assessed predictive performance compared to homoscedastic models $\boldsymbol{y} \sim \text{N}_p(\boldsymbol{\mu}(x), \Sigma)$, with $\mu_j(x)$ either arising as independent $\text{GP}(0, c)$ draws or through a latent factor regression model with $\boldsymbol{\mu}(x) = \Theta\xi(x)\boldsymbol{\psi}(x)$ just as in the heteroscedastic formulation; in both cases, Σ was assigned an inverse-Wishart prior. By



(a) (b)

Figure 1: Plot of each component of the (a) true mean vector $\boldsymbol{\mu}(x)$ and (b) true covariance matrix $\Sigma(x)$ over $\mathcal{X} = \{1, \ldots, 100\}$.



(a) (b) (c)

Figure 2: Differences between each component of the true and posterior mean of (a) the mean $\boldsymbol{\mu}(x)$, and (b) covariance $\Sigma(x)$. The y-axis scale matches that of Figure 1. (c) Box plot of posterior samples of $\log(\sigma_j^2)$ for $j = 1, \ldots, p$ compared to the true value (green).

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. To generate a hold out sample, we removed 48 of the 1,000 observations by deleting observations y_{ij} with probability p_i, where p_i was chosen to vary with x_i to slightly favor removal in regions with more concentrated conditional response distributions.

We first calculated the average Kullback-Leibler divergence between the estimated and true predictive distribution of the missing elements y_{ij} given the observed elements of y_i. The average values were 0.341, 0.291 and 0.122 for the homoscedastic mean regression, homoscedastic latent factor mean regression and heteroscedastic latent factor mean regression, respectively. In this scenario, the missing observations y_{ij} are imputed as an additional step in the MCMC computations.[1] The results clearly indicate that our Bayesian nonparametric covariance regression model provides more accurate predictive distributions. We also observed improvements in estimating the mean $\boldsymbol{\mu}(x)$ for the heteroscedastic approach.

1. 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.

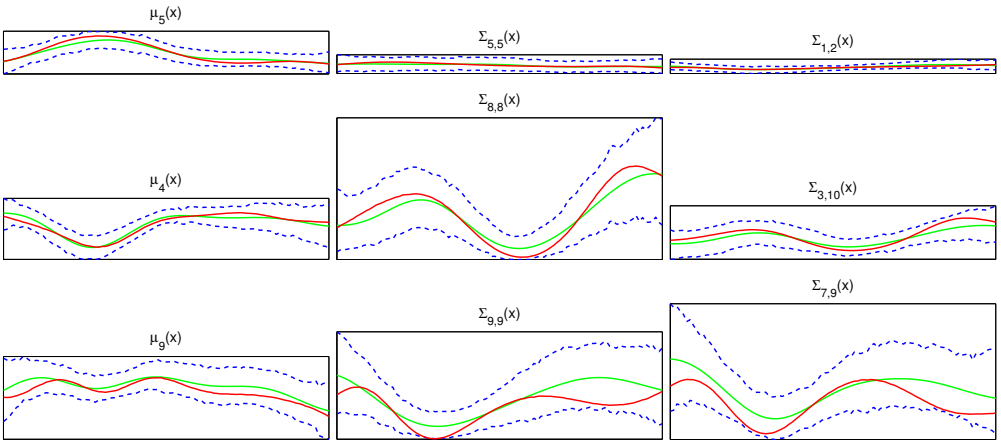


Figure 3: Plots of truth (red) and posterior mean (green) for select components of the mean $\mu_p(x)$ (*left*), variances $\Sigma_{pp}(x)$ (*middle*), and covariances $\Sigma_{pq}(x)$ (*right*). The point-wise 95% highest posterior density intervals are shown in blue. The top row represents the component 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.

In Case 2, we generated 30 replicates from a 30-dimensional parametric heteroscedastic model with $y \sim \mathrm{N}_p(0, \Sigma(x))$ and $\mathcal{X} = \{1, \ldots, 500\}$. To generate $\Sigma(x)$, we chose a set of 5 evenly spaced knots x_k and generated $S(x_k) \sim \mathrm{N}(0, \Sigma_s)$, with $\Sigma_s = \sum_{j=1}^{30} s_j s_j'$ and $s_j \sim \mathrm{N}((-29, -27, \ldots, 27, 29)', I_{30})$. The covariance is constructed as $\Sigma(x) = \alpha \tilde{S}(x)\tilde{S}(x)' + \Sigma_0$, $x = 1, \ldots, 500$, where $\tilde{S}(x)$ is a spline fit to the $S(x_k)$ and Σ_0 is a diagonal matrix with a $\mathrm{N}(0, 1)$ truncated to be positive on its diagonal elements. The constant α is chosen to scale the maximum value of $\alpha \tilde{S}(x)\tilde{S}(x)'$ to 1.

Our hyperparameters and initialization scheme are as in Case 1, but we use truncation levels $\bar{k} = \bar{L} = 5$ based on an initial analysis with $\bar{k} = \bar{L} = 17$. A posterior mean estimate of $\Sigma(x)$ is displayed in Figure 4(c). Compare to the true $\Sigma(x)$ shown in Figure 4(a). Figure 4(b) shows the mean and 95% highest posterior density intervals of the log Frobenius norm $\log ||\Sigma^{(\tau,m)}(x) - \Sigma(x)||_2$ over Gibbs iterations τ and replicates $m = 1, \ldots, 30$. The average (un-logged) 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 values indicate good estimation performance. We compare to a Wishart matrix discounting approach of Prado and West (2010), which is commonly used in stochastic volatility modeling. Details on our implementation are included in Appendix F. From Figures 4(b) and (d), Wishart discounting has substantially worse performance, with estimation error particularly large at high xs due to accumulation of errors in forward filtering.

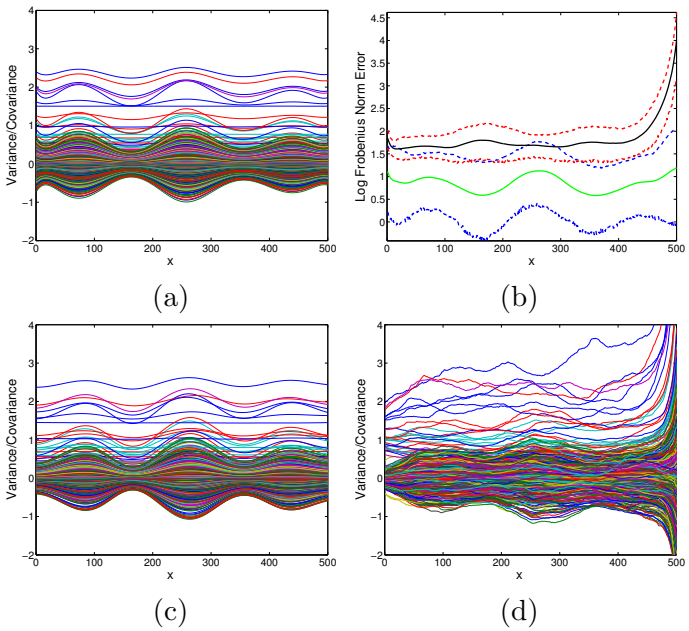


Figure 4: (a) Plot of each component of the true $\Sigma(x)$ over $\mathcal{X} = \{1, \ldots, 500\}$; (c) corresponding posterior means for our proposed approach; and (d) results for a Wishart discounting method, with (c)–(d) based on a single simulation replicate. (b) Mean and 95% highest posterior density intervals of the log Frobenius norm, $\log ||\Sigma^{(\tau,m)}(x) - \Sigma(x)||_2$, for the proposed approach (blue and green) and Wishart discounting (red and black). Results are aggregated over 100 posterior samples and replicates $m = 1, \ldots, 30$.

5. Analysis of Spatio-temporal Trends in Flu

We now turn to our analysis of the Google Flu Trends data, described in detail in Section 5.1. Our focus is on applying our Bayesian nonparametric covariance regression model to capture the heteroscedasticity noted in the exploratory analysis of Appendix G. We also examine how our modeling approach is robust to (i) inaccuracies in the mean model, (ii) missing data, and (iii) outlying estimates.

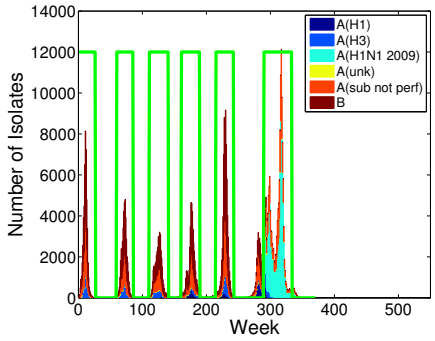
Surveillance of influenza has been of growing interest following a series of pandemic scares (e.g., SARS and avian flu) and the 2009 H1N1 ("swine flu") pandemic. Although influenza pandemics have a long history, a convergence of factors—such as the rapid rate by which geographically distant cases of influenza can spread worldwide—have increased the current public interest in influenza surveillance. A number of papers have recently analyzed the temporal (Martínez-Beneito et al., 2008) and spatio-temporal dynamics of influenza transmission (Stark et al., 2012; Dukić et al., 2012; Hooten et al., 2010; Sakai et al., 2004; Viboud et al., 2004; Mugglin et al., 2002). These approaches focus on data from a modest number of locations, and make restrictive assumptions about the spatial dependence structure, which itself may evolve temporally. Our focus is on addressing these limitations.

For example, Dukić et al. (2012) 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. Additionally, the inherent redundancy and borrowing of information across locations provided by our model should lead to robustness to inaccuracies of flu estimates caused by malicious attacks to the Google infrastructure or unaccounted for sudden spikes in web searches (see Section 5.2.3). Hooten et al. (2010) consider the temporal dynamics of the state-level Google estimates, building on a susceptible–infected–recovered (SIR) model to capture the complexities of intra- and inter-state dynamics of flu dispersal. Such a model aims to captures the intricate mechanistic structure of flu transmission, whereas our goals are focused primarily on fit using metrics such as predictive performance, with an eye towards scalability and robustness. Our exploratory data analysis of Appendix G shows that even with a very flexible and well-fit mean model, temporally changing spatial structure persist in the residuals motivating a heteroscedastic approach. In Section 4, we demonstrated that actually modeling such heteroscedasticity can improve predictive performance. Here, we show that the model of Section 2 can effectively capture such time-varying correlations in region-specific Google-estimated ILI rates, even when considering 183 regions jointly and in the presence of significant missing data.

5.1 Influenza Monitoring and Google Flu Trends

The surveillance of rates of influenza-like illness (ILI) within the United States is coordinated by the Centers for Disease Control and Prevention (CDC), which consolidates data from a large network of diagnostic laboratories, hospitals, clinics, individual healthcare providers, and state health departments. The CDC produces weekly reports (`http://www.cdc.gov/flu/weekly/`) for 10 geographic regions and a U.S. aggregate rate. A plot of the number of isolates tested positive by the WHO and NREVSS from September 28, 2003 to October 24, 2010 is shown in Figure 5 (left). 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. See the specific dates listed in Figure 5 (right). The 2003-2004 flu season began earlier than normal, and coincided with a flu vaccination shortage in many states. Additionally, the CDC found that the vaccination was "not effective or had very low effectiveness" (CDC, 2004). Finally, the 2009-2010 flu season coincides with the emergence of the 2009 H1N1 ("swine flu") subtype in the U.S..

To aid in a more rapid response to influenza activity, researchers at Google devised a model in collaboration with the CDC based on Google user search queries that is meant to be predictive of CDC ILI rates, measured as cases per 100,000 physician visits (Ginsberg et al., 2008). The *Google Flu Trends* methodology was devised based on a two-stage procedure: (i) a massive variable selection procedure was used to select a subset of search queries, and (ii) using these queries as the explanatory variable, region-independent univariate linear



Event	***S***tart Date	***E***nd Date
A	Sept. 28, 2003	Mar. 21, 2004
B	Nov. 14, 2004	May 8, 2005
C	Nov. 6, 2005	May 28, 2006
D	Oct. 22, 2006	May 6, 2007
E	Nov. 4, 2007	May 11, 2008
F	Apr. 12, 2009	Feb. 7, 2010

Figure 5: *Left:* Number of isolates of Influenza A and B tested positive by the WHO and NREVSS over the period of September 29, 2003 to May 23, 2010, with Influenza A broken down into various subtypes. The green line indicates the time periods determined to be flu events. *Right:* Corresponding date ranges for flu events A-F.

models were fit to the weekly CDC ILI rates from 2003-2007. The fitted models are then used for making estimates in any region based on the ILI-related query rates from that region. A key advantage of the Google data 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 U.S. aggregate reporting provided by the CDC.

There has, however, been significant recent debate about the accuracy of the Google Flu Trend estimates (Butler, 2013; Lazer et al., 2014; Harris, 2014). For this paper, we take a backseat in this discussion and simply use this data set to demonstrate the potential impact of our methods in this domain. Revised Google-estimated ILI rates could likewise be used in our framework, as could other recent sources of rapid ILI estimates, e.g., using Twitter data (Lamb et al., 2013; Achrekar et al., 2012) or platforms that incorporate user-contributed reported cases (e.g., `https://flunearyou.org`). Regardless, as we demonstrate in Section 5.2.3, our formulation provides some robustness to inaccurate estimates.

5.1.1 Data Description and Key Features

We analyze the Google Flu Trends data—produced on a weekly basis—from September 28, 2003 through October 24, 2010, totaling 370 weeks. These data provide ILI estimates in 183 regions, consisting of the U.S. national level, 50 states, 10 U.S. Department of Health & Human Services surveillance regions, and 122 cities. For our modeling, we take our observation vectors $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$ to be the log of the Google-estimated ILI rates in the $p = 183$ regions at week i. We denote the untransformed rates by $\boldsymbol{r}_i = (r_{i1}, \ldots, r_{ip})$. Our predictor x_i is simply a discrete time index indicating the current week ($x_i = i$, $i = 1, \ldots, 370$).

Since the Google model fits regions independently, it is not the case that city counts add to regional counts which add to state counts, and so on. That is, the dimensions of $\boldsymbol{y}_i$ are not deterministic functions of each other. There is, however, inherent redundancy (e.g., between

the estimated ILI rates for California and Los Angeles) that is naturally accommodated by a latent factor approach. Another important note is 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.

5.2 Analysis via Bayesian Nonparametric Covariance Regression

We apply our Bayesian nonparametric covariance regression model as follows: $\log \boldsymbol{r}_i \sim N(\boldsymbol{\mu}(x_i), \Sigma(x_i))$. Recall that $\boldsymbol{r}_i$ simply stacks all region-specific measurements r_{ij} into a 183-dimensional vector for each week x_i. The spatial conditional correlation structure at week x_i is then captured by the covariance $\Sigma(x_i) = \Theta\xi(x_i)\xi(x_i)'\Theta' + \Sigma_0$ and the mean by $\boldsymbol{\mu}(x_i) = \Theta\xi(x_i)\boldsymbol{\psi}(x_i)$. Temporal changes are implicitly modeled through the proposed mean-covariance regression framework that allows for continuous variations in $\{\boldsymbol{\mu}(x_i), \Sigma(x_i)\}$ via our Gaussian-process-based formulation. As such, we can also examine $\{\boldsymbol{\mu}(x), \Sigma(x)\}$ for unobserved time points $x \in \mathcal{X}$ occurring between the weekly measurements.

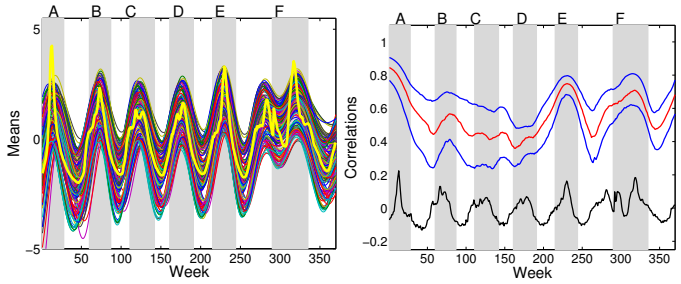
We emphasize that our model does not explicitly encode any spatial structure between the regions (comprising the dimensions of the response vector $\boldsymbol{y}_i$), which is in contrast to many spatial and spatio-temporal models that build in a notion of neighborhood structure. This is motivated both by the fact that, as we see in the correlation maps of the exploratory data analysis in Appendix G, the definition of "neighborhood" is not necessarily straightforward to encode using Euclidean distance since geographically distant regions might have significant correlation[2]. Likewise, this structure need not remain fixed across time. Finally, the full set of 183 regions—comprised of cities, states, regions, and the U.S. national level—represents a type of multiresolution spatial description of flu activity. Although multiresolution-based spatial structures could be imposed based on known relationships, the inherent redundancy of these observations in this task is very well accommodated by a latent factor model. As we have shown, such a structure is very simple to work with computationally and enables our ability to straightforwardly cope with missing data without imputing these values. We could consider a model that combines latent factor and neighborhood based approaches, leading to low-rank plus sparse precision forms for the covariance. This is a topic that has received considerable recent attention (Chandrasekaran et al., 2012). We leave this as a direction of future research.

Details on our model and MCMC setup are provided in Section 5.2.4.

5.2.1 Qualitative Assessment

We begin by producing correlation map snapshots similar to those of the exploratory data analysis in Appendix G, but here with an ability to examine instantaneous correlations that utilize (i) all 183 regions jointly and (ii) the entire time course. In contrast, the analysis of Appendix G reduces dimensionality to state-level, aggregates data amongst flu versus non-flu events to cope with data scarcity, and discards data prior to Event B due to significant missing values. The results presented in Figures 6 and 7 clearly demonstrate that

2. Perhaps this effect arises from air travel (Brownstein et al., 2006), which was found to be a statistically significant driver in the state-level model of Hooten et al. (2010).



(a) (b)

Figure 6: (a) Plot of posterior means of the nonparametric mean function $\mu_j(x)$ for each of the 183 Google Flu Trends regions. The thick yellow line indicates the empirical mean of the log Google-estimated ILI rates, $\log r_{ij}$, across regions j. (b) For New York, the 25th, 50th, and 75th quantiles of correlation with the 182 other regions based on the posterior mean estimate of $\Sigma(x)$. The black line is a scaled version of the log Google-estimated United States ILI rate. The shaded gray regions indicate the time periods determined to be flu events (see Figure 5).

we are able to capture temporal changes in the spatial correlations of the Google Flu Trends data, even in the presence of substantial missing information. In Figure 6(b), we plot the posterior mean of the 183 components of $\boldsymbol{\mu}(x)$, showing trends that follow the empirical mean Google-estimated ILI rate. Although this mean model provides a slightly worse fit than the smoothing splines, our quantitative assessment of Section 5.2.2 demonstrates that modeling heteroscedasticity allows for a well-calibrated joint model. That is, we are robust to our simple choice for the mean regression function. (We note that more complicated mean models could be used within this framework, but this analysis demonstrates the flexibility of joint mean-covariance modeling.) 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 estimate of $\Sigma(x)$. From this plot, we immediately notice the time-varying correlations.

The specific time-varying geographic structure of the inferred correlations is displayed in Figure 7. Qualitatively, we see changes in the residual structure not just between flu and non-flu periods as in Appendix G, but also between flu events. 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. The 2006-2007 season is rather typical, with correlation maps very similar to those of the exploratory data analysis in Figure 12. Note that some geographically distant states, such as New York and California, are often highly correlated. 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. (See the movies

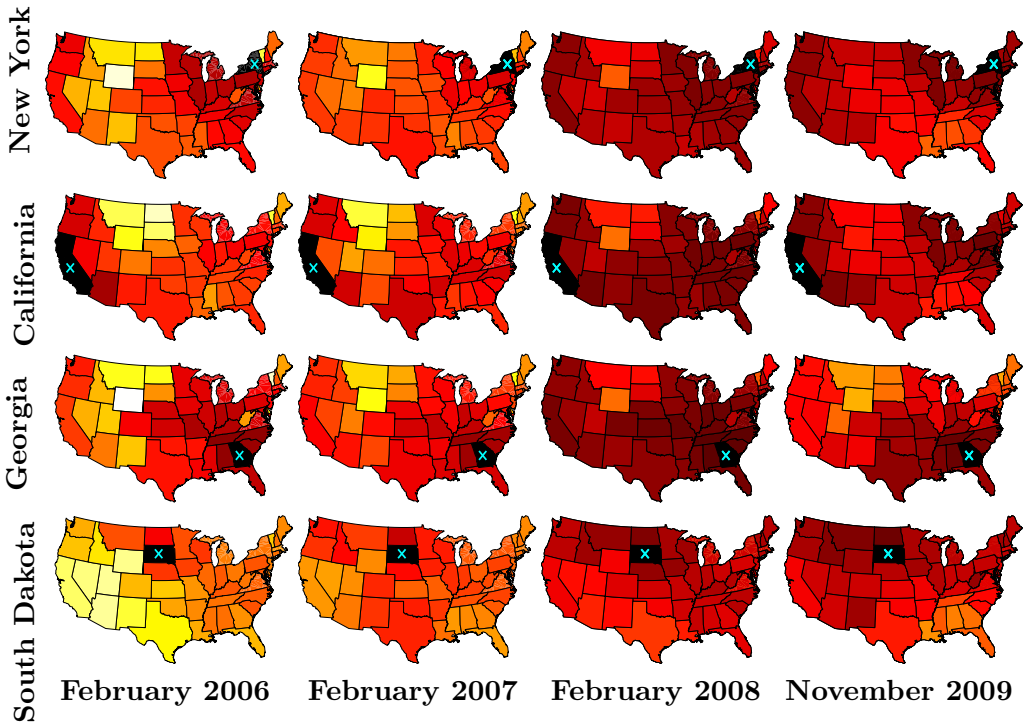


Figure 7: For the states in Figure 12 and each of four key dates (February 2006 of Event C, February 2007 of Event D, February 2008 of Event E, and November 2009 of Event F), correlation maps based on the posterior mean estimate of $\Sigma(x)$ using samples $[5000 : 10 : 10000]$ from 10 chains. The color scale is exactly the same as in Figure 12. The plots 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.

provided in the Online Appendix.) This is enabled by the transfer of knowledge property described in Section 2.3. In particular, the row of Θ corresponding to South Dakota is informed by all of South Dakota's available data while the latent GP basis elements $\xi_{\ell k}$ are informed by all of the other regions' data, in addition to assumed continuity of $\xi_{\ell k}$ which shares information across time.

Comparing the maps of Figure 7 to those of the sample-based estimates in Figure 12, 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 7), they tend to represent a time-average of the event-specific correlation structure we uncovered. Note that due to the dimensionality of the data set, 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. The high-dimensionality and missing data structure of this data set 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

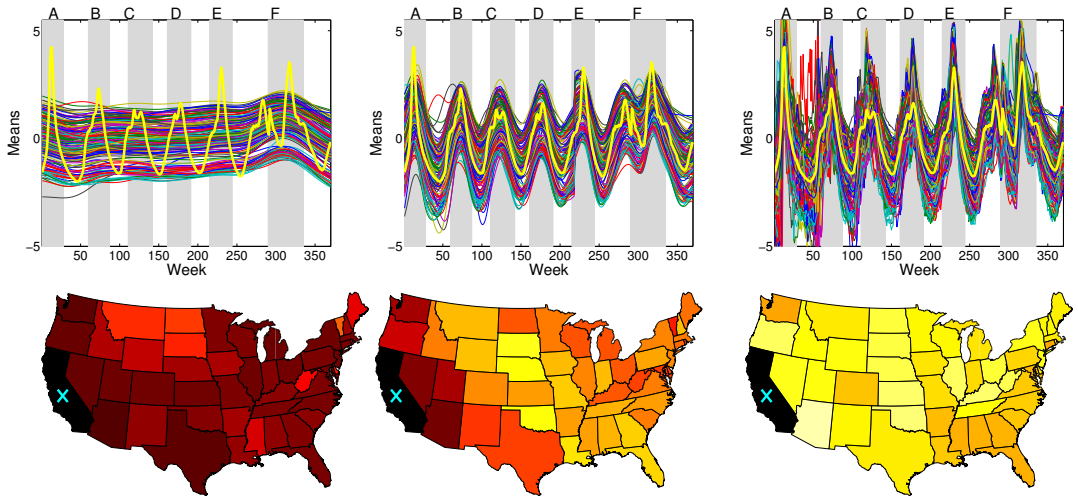


Figure 8: *Top:* Based on the nonparametric Nadaraya-Watson kernel estimator of Yin et al. (2010) using three different bandwidth settings ($(\kappa/2)^{-1/2} = 0.07, 0.02, 0.0008$), plots of the nonparametric mean estimate $\hat{\mu}_j(x)$ for each of the 183 regions, as in Figure 6(b). The estimate is based on averaging samples $[500 : 10 : 1000]$ from a stochastic EM chain that iterated between imputing missing values and computing the kernel estimate. Note that in the rightmost panel, the y-axis is truncated and the estimates in Event A actually extend to above 12. *Bottom:* Associated plots of correlations between California and all other states during February 2006 based on the nonparametric Nadaraya-Watson kernel estimator of the covariance function $\hat{\Sigma}(x)$. The color scale is exactly the same as in Figures 12 and 7.

a lower-dimensional selection (or projection) of observations. For example, the common GARCH models cannot handle missing data and are limited to typically no more than 5 dimensions. On the other hand, our proposed algorithm can readily utilize all information available to model the heteroscedasticity present here.

In an attempt to make a comparison, we propose a stochastic EM algorithm (Diebolt and Ip, 1995) for handling missing data within the framework of the nonparametric Nadaraya-Watson kernel estimator of Yin et al. (2010). Details are provided in Appendix H. The results based on a Gaussian kernel, as employed in Yin et al. (2010), are summarized in Figure 8. We examine three settings for the kernel bandwidth parameter: one (0.0008) based on the cross validation technique proposed in Yin et al. (2010) using the last portion of the time series without any missing values, one (0.02) tuned to match the smoothness of the mean function estimated from the Bayesian nonparametric method proposed herein (see Figure 6(b)), and one large setting (0.07) that leads to substantial sharing of information, but over-smooths the mean. The leave-one-out cross validation method leads to a very small bandwidth because of the specific temporal structure of the data (intuitively, the best estimate of a missing flu rate is achieved by averaging the nearest neighbors in time). However, this setting leads to poor predictive performance in the presence of consecutive

missing values. In Figure 8(a), we see the unreasonably large mean values estimated at the beginning of the time series when there is substantial missing data. The smoothness of the mean function using a bandwidth of 0.02 captures the global changes in flu activity, leaving the covariance to explain the residual correlations in the observations, better matching our goals. However, the covariance, as visualized through the California correlation map of February 2006 (Figure 8(middle)), lacks key geographic structure such as the strong correlation between California and New York. This correlation is present during other flu events, and is unlikely to be truly missing from this event. Instead, the failure to capture this and other correlations is likely due to the increased uncertainty from the substantial early missing data and lack of global sharing of information. Using a much larger bandwidth of 0.07 necessarily leads to more sharing of information, and results in the presence of these correlations. The resulting over-smooth mean function, however, does not capture global flu variations. On the other hand, our Bayesian nonparametric method is able to maintain a local description of the data while sharing information across the entire time series, thus ameliorating sensitivity to missing data.

5.2.2 Model Calibration

The plots of Section 5.2.1 qualitatively demonstrate that we are able to capture time-varying changes in the spatial conditional correlation structure of the (log) Google-estimated ILI rates. Despite not encoding spatial structure in our latent-factor-based model, we note that some local geographic structure has emerged, while still allowing for long-range correlations and temporal changes in this structure. We now turn to a quantitative assessment of the fit and robustness of our model. To this end, we examine posterior predictive intervals of randomly heldout data. More specifically, from the available observations (omitting the significant number of truly missing observations), we randomly held out 10% of the values uniformly across time and regions. We then simulated from our Gibbs sampler treating these values as missing data and analytically marginalizing them from the complete data likelihood, just as we do for the truly missing values. Based on each of our MCMC samples, we form $\boldsymbol{\mu}(x)$ and $\Sigma(x)$ for each $x = 1, \ldots, 370$ and compute the predictive distribution for the heldout data given any available *state-level* observations at week x (i.e., we condition on a subset of observed regions, ignoring non-state-level measurements). Averaging over MCMC samples, we then form 95% posterior predictive intervals and associated coverage rates for each x. We run this experiment of randomly holding out 10% of the observed data twice.

As a comparison, we consider an artificially generated homoscedastic model where we simply form $\hat{\Sigma} = \sum_{i=1}^{370} \Sigma(x_i)$ for each of our MCMC samples. In this case, both models have exactly the same mean regression, $\boldsymbol{\mu}(x)$. Likewise, the underlying $\boldsymbol{\mu}(x)$ and $\Sigma(x)$ (and thus $\hat{\Sigma}$) were all informed using the same low-dimensional embedding of the observations, harnessing the previously described benefits of such a latent factor approach. The only difference is whether we consider the week-specific covariance, $\Sigma(x)$, or instead its mean, $\hat{\Sigma}$, in forming our predictive intervals for week x. Considering the two experiments separately, we find that 94.4% and 94.6% of the heldout observations were covered by our 95% posterior predictive intervals, respectively. This result indicates that our joint mean-covariance regression model is well-calibrated and robust to the rather simple mean model. In compar-

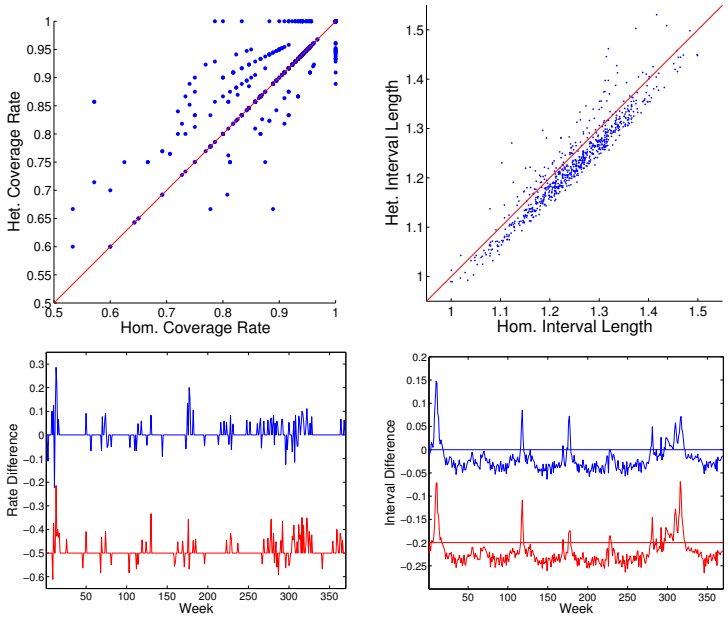


Figure 9: Comparison of posterior predictive intervals using our heteroscedastic model versus a homoscedastic model. *Top:* Scatter plots of week-specific coverage rates and interval lengths, aggregated over two experiments. *Bottom:* Differences in coverage rates and interval lengths by week, separating the experiments via an offset for clarity.

ison, the artificially generated homoscedastic model had coverage rates of 93.7% and 93.8%, respectively. Importantly, the better calibrated coverage rates of our heteroscedastic model came from shorter predictive intervals with average lengths of 1.2272 and 1.2268 compared to 1.2469 and 1.2475 for the homoscedastic model.

Figure 9 explores the differences between these posterior predictive intervals on a week-by-week basis. In Figure 9 (top) we see that a majority of the week-specific intervals have higher coverage rates and shorter interval lengths (i.e., most coverage rate comparisons are on or above the $x - y$ line whereas most interval length comparisons are below this line of equal performance). Time courses of the rates and interval lengths are shown separately for the two experiments in Figure 9 (bottom), where the temporal patterns in these differences become clear. There are stretches of weeks with identical coverage rates, leading to the similarity in overall rates for the two methods, though with the heteroscedastic approach using shorter intervals. The difference of going from overall rates of roughly 93.7% to 94.5% is attributed to certain bursts of time where capturing heteroscedasticity is really key. These time points can sometimes be attributed to the heteroscedastic approach providing wider intervals.

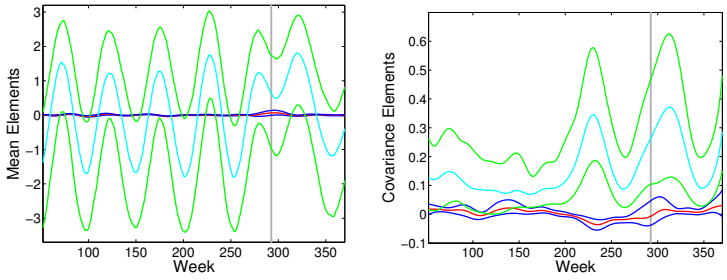


Figure 10: Comparison of posterior mean estimates of $\boldsymbol{\mu}(x)$ and $\Sigma(x)$ using the full data versus using the data with the outlying weeks of April 26, 2009 and May 3, 2009 removed. Denote the resulting estimates by $\{\hat{\boldsymbol{\mu}}(x), \hat{\Sigma}(x)\}$ and $\{\hat{\boldsymbol{\mu}}^{out}(x), \hat{\Sigma}^{out}(x)\}$, respectively. The two outlying weeks are highlighted by the gray shaded region. The blue lines indicate 0.05 and 0.95 pointwise-quantiles of the components (*left*) $\hat{\mu}_j(x) - \hat{\mu}_j^{out}(x)$ and (*right*) $\hat{\Sigma}_{ij}(x) - \hat{\Sigma}_{ij}^{out}(x)$. The red line is the median. The green and cyan lines are the corresponding 0.05, 0.5, and 0.95 quantiles of the full data $\hat{\mu}_j(x)$ and $\hat{\Sigma}_{ij}(x)$ shown for scale. We omit the first year with significant missing data to hone in on the smaller scale variability in subsequent years. No quantiles (blue/green lines) overlapped in this first year for the covariance elements, and trends were of the same scale for the mean elements.

5.2.3 Sensitivity to Outliers

As discussed in Section 5.1, there has been some debate about the accuracy of the Google Flu Trends estimates. In Cook et al. (2011), the two weeks of April 26, 2009 and May 3, 2009 were highlighted as having inflated Google estimates based on the significant media attention spurred by the H1N1 virus. In theory, our mean-covariance regression model has two defenses against such outliers. The first is due to the implicit shrinkage and regularization achieved through our use of a small set of latent basis functions. The second is the fact that an outlier at time x only has limited impact on inferences at time points x' that are "far" from x. This is formalized by in Lemma 5, where we see that the covariance between $\Sigma_{ij}(x)$ and $\Sigma_{uv}(x')$ decays with $(x - x')^2$ (for univariate x and our squared exponential kernel $c(x, x')$).

To empirically examine the impact of these outliers on our inferences, we removed all of the data from the weeks of April 26, 2009 and May 3, 2009 and simulated from our Gibbs sampler treating these data as missing. In Figure 10, we examine the differences between the posterior mean estimates of $\boldsymbol{\mu}(x)$ and $\Sigma(x)$ using the full data and this data set with the outlying weeks removed. We see that the most significant differences in our estimates are localized in time around these removed outlying weeks, as we would expect. Likewise, the sheer magnitude of these differences (which of course are larger for the harder-to-estimate covariance process) are quite small relative to the scale of the parameters in our model. These results demonstrate a robustness to outliers and allude to a robustness to certain types of malicious attacks.

5.2.4 MCMC Details, Sensitivity Analysis, and Convergence Diagnostics

We simulated 5 chains each for 10,000 MCMC iterations, discarded the first 5,000 for burn-in, and thinned the chains by examining every 10 samples. Each chain was initialized with parameters sampled from the prior. The hyperparameters were set as in the simulation study, except with larger truncation levels $\bar{L} = 10$ and $\bar{k} = 20$ and with the Gaussian process length-scale hyperparameter set to $\kappa_\psi = \kappa = 100$ to account for the time scale (weeks) and the rate at which ILI incidences change. By examining posterior samples of Θ, we found that the chosen truncation level was sufficiently large. To assess convergence, we performed the modified Gelman-Rubin diagnostic of Brooks and Gelman (1998) on the MCMC samples of the variance terms $\Sigma_{jj}(x_i)$. We also performed hyperparameter sensitivity, letting $\kappa_\psi = \kappa = 200$ to induce less temporal correlation and using a larger truncation level of $\bar{L} = \bar{k} = 20$ with less stringent shrinkage hyperparameters $a_1 = a_2 = 2$ (instead of $a_1 = a_2 = 10$). The results were essentially identical to those presented. Note that after taking the log transformation, the data were preprocessed by removing the empirical mean of each region and scaling the entire data set by one over the largest variance of any of the 183 time series.

5.2.5 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. For a sense of scaling of computations, the $p = 10$, $n = 100$ simulation study of Section 4 took 10 minutes for 10,000 Gibbs iterations while the $p = 30$, $n = 500$ scenario of took 3 hours for 10,000 Gibbs iterations. In terms of memory and storage, our method only requires maintaining samples of a $p \times L$ matrix Θ, the p elements of Σ_0, and an $L \times k \times q \times n$ matrix for the basis functions $\xi(x)$. (Compare to maintaining the $p \times p \times q \times n$ dimensional matrix for the Nadaraya-Watson estimates of $\Sigma(x)$ in the stochastic EM algorithm to which we compared.)

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 $\Sigma(\boldsymbol{x})$ to vary flexibly over $\boldsymbol{x} \in \mathcal{X}$, where $\mathcal{X}$ is some arbitrary (potentially multivariate) predictor space. As a concrete example, we considered multivariate heteroscedastic modeling of the Google Flu Trends data set, where p represents the 183 regions and x a weekly index. Key to this analysis is our model's ability to (i) scale to the full 183 regions (large p) and (ii) cope with the significant missing data without relying on imputing these values. Inherent to both of these capabilities is our predictor-dependent latent factor model that enables efficient sharing of information in a low-dimensional subspace. The factor loadings are based on a quadratic mixing over a collection of basis elements, assumed herein to be Gaussian process random functions, defined over $\mathcal{X}$. The Gaussian processes define a continuous evolution over $\mathcal{X}$ (e.g., time in the flu analysis), allowing us cope with an irregular grid of observations. 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 data sets. In our Google Flu Trends analysis, we demonstrated the scalability, calibration, and robustness of our formulation.

There are many possible extensions of the proposed covariance regression framework. The most immediate are those that fall into the categories of (i) avoiding the multivariate Gaussian assumption, and (ii) scaling to data sets with larger numbers of observations.

In terms of (i), a natural possibility is to embed the proposed model within a richer hierarchical framework. For example, a promising approach is to use a Gaussian copula while allowing the marginal distributions to be unknown as in Hoff (2007). One can also use more flexible distributions for the latent variables and residuals, such as mixtures of Gaussians. 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).

In terms of (ii), our sampler relies on $\bar{L} \times \bar{k}$ draws from an n-dimensional Gaussian (i.e., posterior draws of our Gaussian process random basis 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 data sets, 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. One could also consider replacing the chosen basis elements with a basis expansion, wavelets, or simply autoregressive (i.e., band-limited) Gaussian processes. Including flat basis elements allows the model to collapse on homoscedasticity, enabling testing for heteroscedasticity.

It is also interesting to consider extensions that harness a known, predictor-independent structured covariance. One approach is to assume a *low rank plus sparse* model (instead of our low rank plus diagonal) in which the residuals have a sparse conditional dependence structure. For example, in the Google flu application the residuals could be modeled via a Markov random field to capture static local spatial dependencies while the low-rank portion captures time variation about this nominal structure. One could similarly extend to unknown sparse structures. Such formulations might allow for fewer latent factor dimensions.

There are also a number of interesting theoretical directions related to showing posterior consistency and rates of convergence including in cases in which the dimension p increases with sample size n.

Acknowledgments

The authors would like to thank Surya Tokdar for helpful discussions on the proof of prior support for the proposed covariance regression formulation. This work was supported in part by NSF Award 0903022 and DARPA Grant FA9550-12-1-0406 negotiated by AFOSR.

Appendix A: Proofs of Theorems and Lemmas

In Lemma 7, we show that our proposed factorization of $\Sigma(x)$ in (3) and (8) is sufficiently flexible to characterize any $\{\Sigma(x), x \in \mathcal{X}\}$ for sufficiently large k. Let $\mathcal{X}_\xi$ denote the space of all possible $L \times k$ arrays of $\mathcal{X} \to \Re$ functions, $\mathcal{X}_{\Sigma_0}$ all $p \times p$ diagonal matrices with non-negative entries, and $\mathcal{X}_\Theta$ all $p \times L$ real-valued matrices such that $\Theta\Theta'$ has finite elements. Recall that our modeling specification considers $L \to \infty$.

Lemma 7 *Given $\Sigma : \mathcal{X} \to \mathcal{P}_p^+$, for sufficiently large k there exists $\{\xi(\cdot), \Theta, \Sigma_0\} \in \mathcal{X}_\xi \otimes \mathcal{X}_\Theta \otimes \mathcal{X}_{\Sigma_0}$ such that $\Sigma(x) = \Theta\xi(x)\xi(x)'\Theta' + \Sigma_0$, for all $x \in \mathcal{X}$.*
Proof *Assume without loss of generality that $\Sigma_0 = 0_{p\times p}$ and take $k \geq p$. Consider*

$$\Theta = [I_p \; 0_{p\times 1} \; 0_{p\times 1} \; \ldots], \quad \xi(x) = \begin{pmatrix} chol(\Sigma(x)) & 0_{p\times k-p} \\ 0_{1\times p} & 0_{1\times k-p} \\ 0_{1\times p} & 0_{1\times k-p} \\ \vdots & \vdots \end{pmatrix}. \tag{13}$$

Then, $\Sigma(x) = \Theta\xi(x)\xi(x)'\Theta^T$ for all $x \in \mathcal{X}$. ■

Proof [Proof of Theorem 2] 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,

$$\Pi_\Sigma\left(\sup_{x\in\mathcal{X}} ||\Sigma(x) - \Sigma^*(x)||_2 < \epsilon\right) = \Pi_\Sigma\left(\max_{x_0\in\mathcal{X}_0} \sup_{x\in B_{\epsilon_0}(x_0)} ||\Sigma(x) - \Sigma^*(x)||_2 < \epsilon\right). \tag{14}$$

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

$$\Pi_\Sigma\left(\max_{x_0\in\mathcal{X}_,} Z(x_0) < \epsilon\right) > 0 \iff \Pi_\Sigma\left(Z(x_0) < \epsilon\right) > 0, \, \forall x_0 \in \mathcal{X}_0, \tag{15}$$

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

$$\begin{aligned}
&\Pi_\Sigma\left(\sup_{x\in B_{\epsilon_0}(x_0)} ||\Sigma(x) - \Sigma^*(x)||_2 < \epsilon\right) \\
&\quad \geq \Pi_\Sigma\Bigg(\sup_{x\in B_{\epsilon_0}(x_0)} ||\Sigma^*(x_0) - \Sigma^*(x)||_2 + \sup_{x\in B_{\epsilon_0}(x_0)} ||\Sigma(x_0) - \Sigma(x)||_2 \\
&\qquad\qquad + ||\Sigma(x_0) - \Sigma^*(x_0)||_2 < \epsilon\Bigg) \\
&\quad \geq \Pi_\Sigma\left(\sup_{x\in B_{\epsilon_0}(x_0)} ||\Sigma^*(x_0) - \Sigma^*(x)||_2 < \epsilon/3\right) \\
&\qquad \cdot \Pi_\Sigma\left(\sup_{x\in B_{\epsilon_0}(x_0)} ||\Sigma(x_0) - \Sigma(x)||_2 < \epsilon/3\right) \Pi_\Sigma\left(||\Sigma(x_0) - \Sigma^*(x_0)||_2 < \epsilon/3\right)
\end{aligned} \tag{16}$$

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

$$||\Sigma^*(x_0) - \Sigma^*(x)||_2 < \epsilon/3, \quad \forall ||x - x_0||_2 < \epsilon_{0,1}. \tag{17}$$

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

$$[\Lambda(x)\Lambda(x)']_{ij} = \sum_{m=1}^{k} g_{im}(x) g_{jm}(x), \quad \forall x \in \mathcal{X}. \tag{18}$$

Eq. (18) 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

$$\Pi_\Sigma \left(\sup_{x \in B_{\epsilon_{0,2}}(x_0)} ||\Sigma(x_0) - \Sigma(x)||_2 < \epsilon/3 \right) = 1. \tag{19}$$

To examine the third term, we first note that

$$\begin{aligned} &\Pi_\Sigma \left(||\Sigma(x_0) - \Sigma^*(x_0)||_2 < \epsilon/3 \right) \\ &\qquad = \Pi_\Sigma \left(||\Theta\xi(x_0)\xi(x_0)'\Theta' + \Sigma_0 - \Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'} - \Sigma_0^*||_2 < \epsilon/3 \right), \end{aligned} \tag{20}$$

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 $\Sigma^*(x_0) = \Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'} + \Sigma_0^*$ with $\Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}$ having rank k^*. Such a factorization exists by the assumption of Σ^* being k^*-decomposable. If $k^* = p$, Lemma 7 states that such a decomposition exists for *any* Σ^*. We can then bound this prior probability by

$$\begin{aligned} &\Pi_\Sigma \left(||\Sigma(x_0) - \Sigma^*(x_0)||_2 < \epsilon/3 \right) \\ &\qquad \geq \Pi_\Sigma \left(||\Theta\xi(x_0)\xi(x_0)'\Theta' - \Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}||_2 < \epsilon/6 \right) \\ &\qquad\qquad \Pi_{\Sigma_0} \left(||\Sigma_0 - \Sigma_0^*||_2 < \epsilon/6 \right) \\ &\qquad \geq \Pi_\Sigma \left(||\Theta\xi(x_0)\xi(x_0)'\Theta' - \Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}||_2 < \epsilon/6 \right) \\ &\qquad\qquad \Pi_{\Sigma_0} \left(||\Sigma_0 - \Sigma_0^*||_\infty < \epsilon/(6\sqrt{p}) \right), \end{aligned} \tag{21}$$

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 $\Sigma_0 = diag(\sigma_1^2,\ldots,\sigma_p^2)$ with $\sigma_i^2 \overset{i.i.d.}{\sim} \text{Ga}(a_\sigma, b_\sigma)$, the support of the gamma prior implies that

$$\Pi_{\Sigma_0}\left(||\Sigma_0 - \Sigma_0^*||_\infty < \epsilon/(6\sqrt{p})\right) = \Pi_{\Sigma_0}\left(\max_{1\leq i\leq p}|\sigma_i^2 - \sigma_i^{*2}| < \epsilon/(6\sqrt{\pi})\right) > 0. \tag{22}$$

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 Θ a real matrix with $\text{rank}(\Theta) = p$,

$$\Theta\xi(x_0)\xi(x_0)'\Theta' \mid \Theta \sim \text{W}(k, \Theta\Theta'). \tag{23}$$

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

$$\Pi_\Sigma\left(||\Theta\xi(x_0)\xi(x_0)'\Theta' - \Theta^*\xi^*(x_0)\xi^*(x_0)'\Theta^{*'}||_2 < \epsilon/6\right) > 0. \tag{24}$$

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. (16) completes the proof. ■

Theorem 8 *Assuming $\mathcal{X}$ compact, for every finite k and $L \to \infty$ (or L finite), $\Lambda(\cdot) = \Theta\xi(\cdot)$ is almost surely continuous on $\mathcal{X}$.*

Proof [Proof of Theorem 8] We can represent each element of $\Lambda(\cdot)$ as follows:

$$\begin{aligned}[\Lambda(\cdot)]_{jk} &= \lim_{L\to\infty}\left[\begin{bmatrix}\theta_{11} & \theta_{12} & \ldots & \theta_{1L}\\ \theta_{21} & \theta_{22} & \ldots & \theta_{2L}\\ \vdots & \vdots & \ddots & \vdots\\ \theta_{p1} & \theta_{p2} & \ldots & \theta_{pL}\end{bmatrix}\begin{bmatrix}\xi_{11}(\cdot) & \xi_{12}(\cdot) & \ldots & \xi_{1k}(\cdot)\\ \xi_{21}(\cdot) & \xi_{22}(\cdot) & \ldots & \xi_{2k}(\cdot)\\ \vdots & \vdots & \ddots & \vdots\\ \xi_{L1}(\cdot) & \xi_{L2}(\cdot) & \ldots & \xi_{Lk}(\cdot)\end{bmatrix}\right]_{jk}\\ &= \sum_{\ell=1}^{\infty}\theta_{j\ell}\xi_{\ell k}(\cdot).\end{aligned} \tag{25}$$

If $\xi_{\ell k}(x)$ is continuous for all ℓ, k and $s_n(x) = \sum_{\ell=1}^{n}\theta_{j\ell}\xi_{\ell k}(x)$ 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 $n \geq N$

$$\Pr\left(\sup_{x\in\mathcal{X}}|g_{jk}(x) - s_n(x)| < \epsilon\right) = 1, \tag{26}$$

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

$$\sup_{x\in\mathcal{X}}|\theta_{jn}\xi_{nk}(x)| \leq M_n. \tag{27}$$

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

$$\sup_{x \in \mathcal{X}} |\theta_{jn} \xi_{nk}(x)| \leq |\theta_{jn}| c_{nk}. \tag{28}$$

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

$$\begin{aligned} E_{\Theta,c}\left[\sum_{n=1}^{\infty} M_n\right] &= E_{\Theta}\left[E_{c|\Theta}\left[\sum_{n=1}^{\infty} |\theta_{jn}| c_{nk} \mid \Theta\right]\right] = E_{\Theta}\left[\sum_{n=1}^{\infty} |\theta_{jn}| \bar{c}\right] \\ &= \bar{c} \sum_{n=1}^{\infty} E_{\Theta}\left[|\theta_{jn}|\right], \end{aligned} \tag{29}$$

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. ■

Lemma 9 *Assuming the prior specification of expression (9) with $a_2 > 2$ and $\gamma > 2$, the rows of Θ are absolutely summable in expectation: $\sum_{\ell} E(|\theta_{j\ell}|) < \infty$, satisfying Assumption 2.1.*

Proof [Proof of Lemma 9] Recall that $\theta_{j\ell} \sim \mathcal{N}(0, \phi_{j\ell}^{-1} \tau_{\ell}^{-1})$ with $\phi_{j\ell} \sim \text{Ga}(\gamma/2, \gamma/2)$ and $\tau_{\ell} = \prod_{h=1}^{\ell} \delta_h$ for $\delta_1 \sim \text{Ga}(a_1, 1)$, $\delta_h \sim \text{Ga}(a_2, 1)$. Using the fact that if $x \sim \mathcal{N}(0, \sigma^2)$ then $E[|x|] = \sigma\sqrt{2/\pi}$ and if $y \sim \text{Ga}(a, b)$ then $1/y \sim \text{Inv-Ga}(a, 1/b)$ with $E[1/y] = 1/(b \cdot (a-1))$, we derive that

$$\begin{aligned} \sum_{\ell=1}^{\infty} E_{\theta}[|\theta_{j\ell}|] &= \sum_{\ell=1}^{\infty} E_{\phi,\tau}[E_{\theta|\phi,\tau}[|\theta_{j\ell}| \mid \phi_{j\ell}, \tau_{\ell}]] = \sqrt{\frac{2}{\pi}} \sum_{\ell=1}^{\infty} E_{\phi,\tau}[\phi_{j\ell}^{-1} \tau_{\ell}^{-1}] \\ &= \sqrt{\frac{2}{\pi}} \sum_{\ell=1}^{\infty} E_{\phi}[\phi_{j\ell}^{-1}] E_{\tau}[\tau_{\ell}^{-1}] = \frac{4}{\gamma(\gamma-2)} \sqrt{\frac{2}{\pi}} \sum_{\ell=1}^{\infty} E_{\delta}\left[\prod_{h=1}^{\ell} \frac{1}{\delta_h}\right] \\ &= \frac{1}{a_1 - 1} \frac{4}{\gamma(\gamma-2)} \sqrt{\frac{2}{\pi}} \sum_{\ell=1}^{\infty} \left(\frac{1}{a_2 - 1}\right)^{\ell-1}. \end{aligned} \tag{30}$$

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

Proof [Proof of Lemma 4] Recall that $\Sigma(x) = \Theta \xi(x) \xi(x)' \Theta' + \Sigma_0$ with $\Sigma_0 = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. The elements of the respective matrices are independently distributed as $\theta_{i\ell} \sim \mathcal{N}(0, \phi_{i\ell}^{-1} \tau_{\ell}^{-1})$, $\xi_{\ell k}(\cdot) \sim \text{GP}(0, c)$, and $\sigma_i^{-2} \sim \text{Gamma}(a_\sigma, b_\sigma)$. 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

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

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

$$E[\Theta\Theta']_{ij} = \sum_{\ell}\sum_{\ell'} E[\theta_{i\ell}\theta_{j\ell'}] = \sum_{\ell} E[\theta_{i\ell}^2]\delta_{ij} = \sum_{\ell} \text{var}(\theta_{i\ell})\delta_{ij} = \sum_{\ell} \phi_{i\ell}^{-1}\tau_{\ell}^{-1}\delta_{ij}.$$

■

Proof [Proof of Lemma 5] One can use the conditionally Wishart distribution of $\Theta\xi(x)\xi(x)'\Theta'$ to derive $\text{cov}(\Sigma_{ij}(x), \Sigma_{uv}(x))$. 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,

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

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

$$g_{in}(x) = \sum_{\ell} \theta_{i\ell}\xi_{\ell n}(x), \tag{31}$$

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

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

We replace each $g_{kn}(x)$ by the form in Eq. (31), 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,

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

The Gaussian process moments are given by

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

from which we derive that

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

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

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

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

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

■

Proof [Proof of Lemma 6] The first-order stationarity follows immediately from the stationarity of the Gaussian process dictionary elements $\xi_{\ell k}$ 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', Lemma 5 implies that the defined process is wide sense stationary. ■

Appendix B: Derivation of Gibbs Sampler

In this Appendix, we derive the conditional distribution for sampling the Gaussian process dictionary elements. Combining Eq. (1) and Eq. (8), we have that

$$y_i = \Theta \begin{bmatrix} \xi_{11}(x_i) & \xi_{12}(x_i) & \dots & \xi_{1k}(x_i) \\ \xi_{21}(x_i) & \xi_{22}(x_i) & \dots & \xi_{2k}(x_i) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{L1}(x_i) & \xi_{L2}(x_i) & \dots & \xi_{Lk}(x_i) \end{bmatrix} \eta_i + \epsilon_i = \Theta \begin{bmatrix} \sum_{m=1}^k \xi_{1m}(x_i)\eta_{im} \\ \vdots \\ \sum_{m=1}^k \xi_{Lm}(x_i)\eta_{Lm} \end{bmatrix} + \epsilon_i \tag{32}$$

implying that

$$y_{ij} = \sum_{\ell=1}^L \sum_{m=1}^k \theta_{j\ell}\eta_{im}\xi_{\ell m}(x_i) + \epsilon_{ij}. \tag{33}$$

Conditioning on $\xi(\cdot)^{-\ell m}$, we rewrite Eq. (32) as

$$y_i = \eta_{im} \begin{bmatrix} \theta_{1\ell} \\ \vdots \\ \theta_{p\ell} \end{bmatrix} \xi_{\ell m}(x_i) + \tilde{\epsilon}_i, \quad \tilde{\epsilon}_i \sim \mathcal{N}\left(\mu_{\ell m}(x_i) \triangleq \begin{bmatrix} \sum_{(r,s)\neq(\ell,m)} \theta_{1r}\eta_{is}\xi_{rs}(x_i) \\ \vdots \\ \sum_{(r,s)\neq(\ell,m)} \theta_{pr}\xi_{rs}(x_i) \end{bmatrix}, \Sigma_0 \right). \tag{34}$$

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

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \eta_{1m}\theta_{\cdot\ell} & 0 & \dots & 0 \\ 0 & \eta_{2m}\theta_{\cdot\ell} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_{nm}\theta_{\cdot\ell} \end{bmatrix} \begin{bmatrix} \xi_{\ell m}(x_1) \\ \xi_{\ell m}(x_2) \\ \vdots \\ \xi_{\ell m}(x_n) \end{bmatrix} + \begin{bmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \vdots \\ \tilde{\epsilon}_n \end{bmatrix} \tag{35}$$

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

$$\begin{bmatrix} \xi_{\ell m}(x_1) \\ \xi_{\ell m}(x_2) \\ \vdots \\ \xi_{\ell m}(x_n) \end{bmatrix} \mid \{y_i\}, \Theta, \eta, \xi(\cdot), \Sigma_0 \sim \mathcal{N}\left(\tilde{\Sigma} A_{\ell m}' \begin{bmatrix} \Sigma_0^{-1} & & \\ & \ddots & \\ & & \Sigma_0^{-1} \end{bmatrix} \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix}, \tilde{\Sigma} \right)$$

$$= \mathcal{N}\left(\tilde{\Sigma} \begin{bmatrix} \eta_{1m} \sum_{j=1}^p \theta_{j\ell} \sigma_j^{-2} \tilde{y}_{1j} \\ \vdots \\ \eta_{nm} \sum_{j=1}^p \theta_{j\ell} \sigma_j^{-2} \tilde{y}_{nj} \end{bmatrix}, \tilde{\Sigma} \right), \tag{36}$$

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,

$$\tilde{\Sigma}^{-1} = K^{-1} + A_{\ell m}' \begin{bmatrix} \Sigma_0^{-1} & & \\ & \ddots & \\ & & \Sigma_0^{-1} \end{bmatrix} A_{\ell m}$$

$$= K^{-1} + \text{diag}\left(\eta_{1m}^2 \sum_{j=1}^p \theta_{j\ell}^2 \sigma_j^{-2}, \ldots, \eta_{nm}^2 \sum_{j=1}^p \theta_{j\ell}^2 \sigma_j^{-2} \right). \tag{37}$$

Appendix C: Hyperparameter Sampling and Empirical Bayes

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. Taking $\mu(x) = 0$ for simplicity, our posterior is based on marginalizing the Gaussian processes random vectors $\underline{\xi}_{\ell m} = [\xi_{\ell m}(x_1) \ \ldots \ \xi_{\ell m}(x_n)]'$. Noting that

$$\begin{bmatrix} y_1' & y_2' & \ldots & y_n' \end{bmatrix}' = \sum_{\ell m} \left[\text{diag}(\eta_{\cdot m}) \otimes \theta_{\cdot \ell} \right] \underline{\xi}_{\ell m} + \begin{bmatrix} \epsilon_1' & \epsilon_2' & \ldots & \epsilon_n' \end{bmatrix}', \tag{38}$$

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

$$\begin{bmatrix} y_1' & \cdots & y_n' \end{bmatrix}' \mid \kappa, \Theta, \eta, \Sigma_0 \sim$$

$$\text{N}_{np}\left(\sum_{\ell, m} \left[\text{diag}(\eta_{\cdot m}) \otimes \theta_{\cdot \ell} \right] K_\kappa \left[\text{diag}(\eta_{\cdot m}) \otimes \theta_{\cdot \ell} \right]' + I_n \otimes \Sigma_0 \right). \tag{39}$$

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. (39) 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 ξ given K_κ and K_κ given ξ. 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 preselect a value for κ using a data-driven heuristic, which leads to a quasi-empirical Bayes approach. Lemma 5 implies that the autocorrelation $ACF(x) = \text{corr}(\Sigma_{ij}(0), \Sigma_{ij}(x))$ is simply specified by $c(0, x)$. As given by Eq. (11), when we choose a Gaussian process kernel $c(x, x') = \exp(-\kappa||x - x'||_2^2)$, we have $ACF(x) = \exp(-\kappa||x||_2^2)$. Thus, we see that the length-scale parameter κ directly determines the shape of the autocorrelation function. If one can devise a procedure for estimating the autocorrelation function from the data, one can set κ accordingly. We propose the following, most easily implemented for scalar predictor spaces $\mathcal{X}$, but also feasible (in theory) for multivariate $\mathcal{X}$.

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. If the bin contains fewer than p observations, add a small diagonal component to ensure positive definiteness.
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 $i = 1, \ldots, n$, 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 $-\log(ACF(x)) = \kappa||x||_2^2$, 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.)

Appendix D: Initialization of Gibbs Sampler

Experimentally we found that our sampler was fairly insensitive to initialization (after a short burn-in period) and one can just initialize each of Θ, ξ, Σ_0, η_i, and the shrinkage parameters $\phi_{j\ell}$ and δ_h from their respective priors. However, in certain scenarios, the following more intricate initialization can improve mixing rates. 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 outlined above. 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 their posterior 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) ξ given $\{y_i\}$, Θ, Σ_0, and the data-driven estimates of η, ξ; (ii) Θ given $\{y_i\}$, Σ_0, η, and the sampled ξ; (iii) Σ_0 given $\{y_i\}$, Θ, η, and ξ; and (iv) determining a new data-driven approximation to ξ based on the newly sampled Θ.

Appendix E: Other Prior Specifications

Intuitively, the chosen prior on Θ flexibly shrinks the columns of this matrix towards zero as the column index increases, implying that the effect of dictionary elements $\xi_{\ell k}(x)$ on the induced covariance matrix $\Sigma(x)$ decreases with row index ℓ. Harnessing this idea, one can extend the framework to allow for variable-smoothness dictionary elements by introducing row-dependent bandwidth parameters κ_ℓ. For example, one could encourage increasingly *bumpy* dictionary elements $\xi_{\ell k}(\cdot)$ for large ℓ in order to capture multiple resolutions of smoothness in the covariance regression. The prior on Θ would then encourage the smoother dictionary elements to be more prominent in forming $\Sigma(x)$, with the bumpier elements being more heavily regularized. One could, of course, also consider other dictionary element specifications such as based on basis expansions or with a finite autoregressive (band-limited covariance) structure. Such specifications could ameliorate some of the computational burden associated with Gaussian processes, but might induce different prior support for the covariance regression.

Likewise, just as we employed a shrinkage prior on Θ to be more robust to the choice of $\bar{L}$, one could similarly cope with $\bar{k}$ by considering an augmented formulation in which

$$\Lambda(x) = \Theta\xi(x)\Gamma, \tag{40}$$

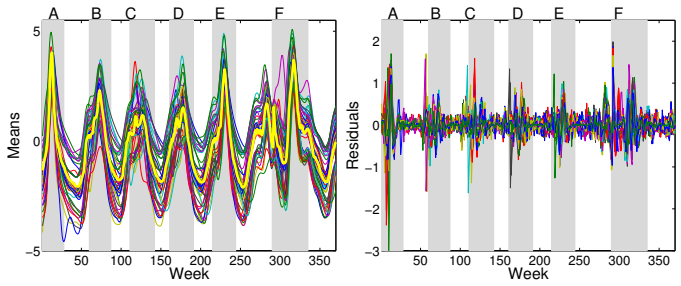
where $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_k)$ is a diagonal matrix of parameters that shrink the columns of $\xi(x)$ towards zero. One can take these shrinkage parameters to be distributed as

$$\gamma_i \sim \text{N}(0, \omega_i^{-1}), \quad \omega_i = \prod_{h=1}^{i} \zeta_h, \quad \zeta_1 \sim \text{Ga}(a_3, 1), \quad \zeta_h \sim \text{Ga}(a_4, 1) \quad h = 2, \ldots, k. \tag{41}$$

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. Computations in this augmented model are a straightforward extension of the developed Gibbs sampler.

Appendix F: Simulation Studies

For the simulation studies of Case 2, the Wishart matrix discounting method to which we compared is given as follows, with details in Section 10.4.2 of Prado and West (2010). Let $\Phi_t = \Sigma_t^{-1}$. The Wishart matrix discounting model assumes $\Sigma_t^{-1} \mid y_{1:t-1}, \beta \sim W(\beta h_{t-1}, (\beta D_{t-1})^{-1})$, with $D_t = \beta D_{t-1} + y_t y_t'$ and $h_t = \beta h_{t-1} + 1$, such that $E[\Sigma_t^{-1} \mid y_{1:t-1}] = E[\Sigma_{t-1}^{-1} \mid y_{1:t-1}] = h_{t-1} D_{t-1}^{-1}$, but with certainty discounted by a factor determined by β. The update with observation y_t is conjugate, maintaining a Wishart posterior on Σ_t^{-1}. 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. Increasing h_t can mitigate the large errors for high xs seen in Figure 4(b) and (d), 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.



(a) (b)

Figure 11: (a) Plot of smoothing spline fits $\hat{f}_j(x)$ for each of the 183 Google Flu Trends regions. The thick yellow line indicates the empirical mean of the log Google-estimated ILI rates, $\log y_{ij}$, across regions j. (b) Residuals $\log r_{ij} - \hat{f}_j(x_i)$. The shaded gray regions indicate the flu events of Figure 5.

Appendix G: Exploratory Data Analysis

To examine the spatial correlation structure of the Google-estimated ILI rates and how these correlations vary across time, we performed the following exploratory data analysis. First, we consider a log transform of our rate data and a model

$$\log r_{ij} = f_j(x_i) + \epsilon_{ij}, \quad i = 1, \ldots, 370, \; j = 1, \ldots, 183, \tag{42}$$

with $f_j(\cdot)$ taken to be a region-specific smoothing spline. These spline fits are shown along with the residuals ϵ_{ij} in Figure 11. We then examine the spatial correlations of these residuals in Figure 12. We omit data prior to Event B because of the extent of missing values. Due to the dimensionality of the data (183 dimensions) and limited number of observations (157 event and 127 non-event observations), we simply consider state-level observations plus District of Columbia (resulting in 51 dimensions) and then aggregate the data over Events B-F to create a "flu event" maximum likelihood (ML) estimate, $\hat{\Sigma}^{flu}$. We likewise examine an aggregation of data between events to form a "non-event" estimate $\hat{\Sigma}^{nonflu}$.

From Figure 12, we see that the correlations between regions is much lower during non-event periods than event periods. For event periods, there is clear spatial correlation, defined both locally and with long-range dependencies. Note that because of the dimensionality and limited data, these exploratory methods cannot handle the full set of regions nor examine smoothly varying correlations. Instead, the plots simply provide insight into the geographic structure of the correlations and the fact that this structure is clearly different within and outside of flu event periods. As such, an i.i.d. model for $\boldsymbol{\epsilon}_i$ is inappropriate, motivating our heteroscedastic model of Section 2. In Section 5, we analyze how our proposed covariance regression model enables analysis of the changing extent and intensity of the correlations as a function of time. The method allows us to harness all of the available data, both across regions and time.

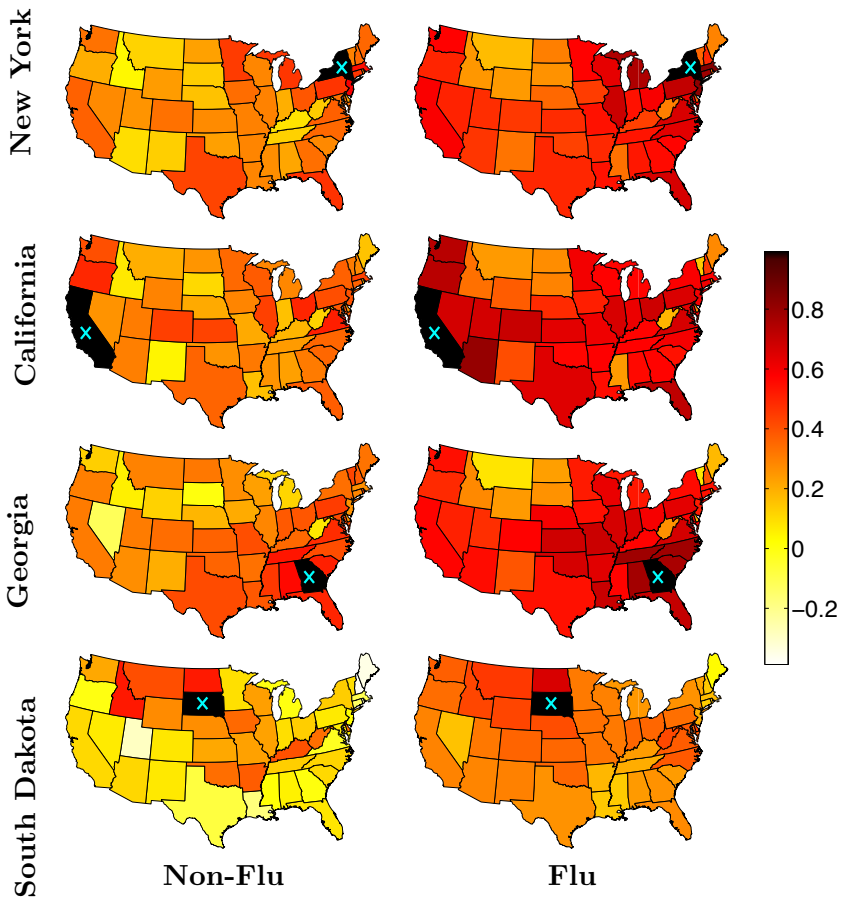


Figure 12: For each of four geographically distinct states (New York, California, Georgia, and South Dakota), plots of correlations between the state and all other states based on the sample covariance estimate from state-level data. The estimates are for data aggregated over non-event periods following event B (left) and event periods B-F (right). The data are taken to be the residuals of smoothing spline estimates fit independently for each region using log ILI rates. 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.

Appendix H: Details on Nadaraya-Watson Approach

Our proposed stochastic EM algorithm for the nonparametric Nadaraya-Watson kernel estimator iterates between (i) sampling missing values from the predictive distribution associated with the current kernel estimates of the mean and covariance functions and the available data, and (ii) computing the kernel-estimate of the mean and covariance functions using the available data and imputed missing values. We initialize by pooling all available data to form a static mean and covariance estimate from which the missing values are initially sampled. Due to the high-dimensionality compared to the limited bandwidth, we add a diagonal element $1e^{-6}I_p$ to the estimate $\hat{\Sigma}(x)$ to ensure positive definiteness.

References

H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Twitter improves seasonal influenza prediction. In *HEALTHINF*, pages 61–70, 2012.

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

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. *Biometrika*, 98 (2):291–306, 2011.

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

J. S. Brownstein, C. J. Wolfe, and K. D. Mandl. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the United States. *PLoS Medicine*, 3(10):e401, 2006.

D. Butler. When Google got flu wrong: US outbreak foxes a leading web-based method for tracking seasonal. *Nature*, 494:155–156, 2013.

CDC. Influenza vaccine effectiveness studies, January 2004. URL `http://www.cdc.gov/media/pressrel/fs040115.htm`.

V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *Annals of Statistics*, 40(4):1935–1967, 2012.

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.

S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE*, 6(8):e23610, 2011.

J. Diebolt and E.H.S. Ip. *Markov Chain Monte Carlo in Practice*, chapter Stochastic EM: methods and application, pages 259–273. Chapman & Hall, 1995.

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 epidemics with Google Flu Trends data and a state-space SEIR model. *Journal of the American Statistical Assocaition*, 107(500): 1410–1426, 2012.

D. Durante, B. Scarpa, and D. B. Dunson. Locally adaptive factor processes for multivariate time series. *The Journal of Machine Learning Research*, 15(1):1493–1522, 2014.

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

B. K. Fosdick and P. D. Hoff. Separable factor analysis with applications to mortality data. *Annals of Applied Statistics*, 8(1):120–147, 2014.

E. B. Fox and D. B. Dunson. Bayesian nonparametric covariance regression. *arXiv preprint arXiv:1101.2017*, 2011.

W. A. Fuller. *Introduction to Statistical Time Series*, volume 428. John Wiley & Sons, 2009.

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. 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.

R. Harris. Google's flu tracker suffers from sniffles. *http://www.npr.org/blogs/health/2014/03/13/289802934/googles-flu-tracker-suffers-from-sniffles*, March 2014.

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

D. Higdon, C. Nakhleh, J. Gattiker, and B. Williams. A Bayesian calibration approach to the thermal problem. *Computer Methods in Applied Mechanics and Engineering*, 197 (29):2431–2441, 2008.

P. D. Hoff and X. Niu. A covariance regression model. *Statistica Sinica*, 22:729–753, 2012.

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

M. B. Hooten, J. Anderson, and L. A. Waller. Assessing North American influenza dynamics with a statistical SIRS model. *Spatial and Spatio-temporal Epidemiology*, 1(2):177–185, 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. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of NAACL-HLT*, pages 789–795, 2013.

D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

C. Leng, W. Zhang, and J. Pan. Semiparametric mean-covariance regression analysis for longitudinal data. *Journal of the American Statistical Association*, 105(489):181–193, 2010.

J.S. Liu, W.H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1): 27–40, 1994.

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

M.A. Martínez-Beneito, D. Conesa, A. López-Quílez, and A. López-Maside. Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine*, 27(22):4455–4468, 2008.

A. S. Mugglin, N. Cressie, and I. Gemmell. Hierarchical statistical modelling of influenza epidemic dynamics in space and time. *Statistics in Medicine*, 21(18):2703–2721, 2002.

R. Paulo. Default priors for Gaussian processes. *Annals of Statistics*, pages 556–582, 2005.

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.

C. E. Rasmussen and C. K. .I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

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.

T. Sakai, H. Suzuki, A. Sasaki, R. Saito, N. Tanabe, and K. Taniguchi. Geographic and temporal trends in influenzalike illness, Japan, 1992-1999. *Emerging Infectious Diseases*, 10(10):1822–1826, 2004.

J. H. Stark, R. Sharma, S. Ostroff, D. A. T. Cummings, B. Ermentrout, S. Stebbins, D. S. Burke, and S. R. Wisniewski. Local spatial and temporal processes of influenza in Pennsylvania, USA: 2003–2009. *PLoS ONE*, 7(3):e34245, 2012.

C. Viboud, P. Y. Boëlle, K. Pakdaman, F. Carrat, A. J. Valleron, A. Flahault, et al. Influenza epidemics in the United States, France, and Australia, 1972-1997. *Emerging Infectious Diseases*, 10(1):32–39, 2004.

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

A. G. Wilson and Z. Ghahramani. Generalized Wishart processes. In *Uncertainty in Artificial Intelligence*, 2011.

J. Yin, Z. Geng, R. Li, and H. Wang. Nonparametric covariance model. *Statistica Sinica*, 20:469–479, 2010.

W. Zhang and C. Leng. A moving average Cholesky factor model in covariance modelling for longitudinal data. *Biometrika*, 99(1):141–150, 2012.