Title

Multivariate spatiotemporal functional principal component analysis for modeling hospitalization and mortality rates in the dialysis population.

Journal

Biostatistics, 25(3)

Authors

Qian, Qi
Nguyen, Danh
Telesca, Donatello
et al.

Publication Date

2024-07-01

DOI

10.1093/biostatistics/kxad013

Peer reviewed

Biostatistics (2024) **25**, 3, *pp.* 718–735
https://doi.org/10.1093/biostatistics/kxad013
Advance Access publication on June 20, 2023

Multivariate spatiotemporal functional principal component analysis for modeling hospitalization and mortality rates in the dialysis population

QI QIAN

Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

DANH V. NGUYEN

Department of Medicine, University of California, Irvine, CA 92868, USA

DONATELLO TELESCA

Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

ESRA KURUM

Department of Statistics, University of California, Riverside, CA 92521, USA

CONNIE M. RHEE

Department of Medicine, University of California, Irvine, CA 92868, USA and Harold Simmons Center for Chronic Disease Research and Epidemiology, University of California School of Medicine, Irvine, CA 92868, USA

SUDIPTO BANERJEE, YIHAO LI, DAMLA SENTURK*

Department of Biostatistics, University of California, Los Angeles, CA 90095, USA
dsenturk@ucla.edu

Summary

Dialysis patients experience frequent hospitalizations and a higher mortality rate compared to other Medicare populations, in whom hospitalizations are a major contributor to morbidity, mortality, and healthcare costs. Patients also typically remain on dialysis for the duration of their lives or until kidney transplantation. Hence, there is growing interest in studying the spatiotemporal trends in the correlated outcomes of hospitalization and mortality among dialysis patients as a function of time starting from transition to dialysis across the United States Utilizing national data from the United States Renal Data System (USRDS), we propose a novel multivariate spatiotemporal functional principal component analysis model to study the joint spatiotemporal patterns of hospitalization and mortality rates among dialysis patients. The proposal is based on a multivariate Karhunen–Loéve expansion that describes leading directions of variation across time and induces

*To whom correspondence should be addressed.

spatial correlations among region-specific scores. An efficient estimation procedure is proposed using only univariate principal components decompositions and a Markov Chain Monte Carlo framework for targeting the spatial correlations. The finite sample performance of the proposed method is studied through simulations. Novel applications to the USRDS data highlight hot spots across the United States with higher hospitalization and/or mortality rates and time periods of elevated risk.

Keywords: Conditional autoregressive model; End-stage kidney disease; Multivariate conditional autoregressive model; Multivariate functional principal component analysis; United States Renal Data System.

1. Introduction

End-stage kidney disease (ESKD) affected more than 809, 103 individuals in the United States as of 2019, a 41% increase from 2009. About 70% of patients with incident ESKD were on dialysis, a life-sustaining treatment (USRDS, 2021). The mortality rate among patients receiving dialysis is higher than Medicare populations with heart failure, acute myocardial infarction, or cancer. Moreover, patients on dialysis have a high burden of complex comorbid conditions and experience frequent hospitalizations (at about twice per year), which is a major source of morbidity and mortality (USRDS, 2021). Hence, hospitalization and survival are intricately related in the dialysis population.

ESKD patients typically remain on dialysis for the rest of their lives (or until kidney transplantation) and their needs change as they stay longer on dialysis. Our own works and others have shown that mortality risk and hospitalization rates change over time after patients transition to dialysis (higher hospitalization and mortality rates have been reported within the first year and first 6 months on dialysis, respectively (Estes *and others*, 2018; Li *and others*, 2018)). In addition to temporal variation, hospitalization and mortality rates also vary significantly across the United States, contributing to spatial variation (Li *and others*, 2021). Understanding the geospatial patterns of outcomes for dialysis patients has been an important objective of the United States Renal Data System (USRDS) annual reporting. Hence, studying the spatiotemporal trends in the multivariate outcome of hospitalization risk and mortality rates for regions across the United States is an important goal in identifying "hot spots"/regions and critical time periods (after transitioning to dialysis) of high risk and elevated rates for more targeted patient monitoring.

There is extensive literature on multivariate spatiotemporal modeling, particularly in environmental health, criminology, road traffic analysis, and disease mapping. The existing multivariate space-time approaches can broadly be classified into four settings, in which time or space or both are modeled as a discrete or a continuous process. In our application to modeling the multivariate outcome of hospitalization and mortality rates, we target rates across health service areas (HSAs) in the United States (regions with relatively self-contained infrastructure for the provision of hospital care) and across time on dialysis. Hence in our application, space is discrete, since HSAs across the United States are fixed, and time is continuous, since we are interested in drawing inference along the continuous time index of duration on dialysis. While there is extensive literature on multivariate spatiotemporal modeling when space and time are viewed as discrete, the literature on discrete space and continuous time models is limited. When space and time are viewed as discrete, a Markov random field structure in the form of conditionally autoregressive (CAR) or multivariate conditionally autoregressive (MCAR) specifications are commonly used for spatial modeling, while time is typically modeled via an autoregressive structure (see Banerjee *and others* (2015) for a comprehensive review). For discrete space and continuous time models, Zhang *and others* (2006) proposed a separable structure for the multivariate space-time covariance, while Hepler *and others* (2021) and Baer *and others* (2021) considered nonseparable space-time covariances. However, both of the latter approaches use quite restricted linear or user-defined parametric forms in temporal modeling.

We propose multivariate spatiotemporal functional principal components analysis (MST-FPCA) to study the joint spatiotemporal patterns of hospitalization and mortality rates among US dialysis patients. The proposed model is based on a multivariate Karhunen–Loéve expansion which models temporal trends nonparametrically via a data-driven lower dimensional multivariate eigenfunction bases. Spatial correlations are induced through a CAR model assumed among region-specific scores, leading to a nonseparable structure on the multivariate space-time covariance. Functional principal components decompositions for longitudinal and functional processes have been a major modeling theme in the functional data analysis literature, with focus centering on efficient and interpretable representations of functional variability in highly structured settings (Ramsay and Silverman, 2005). More recently, there has been interest in analyzing multiple trajectories, with dependencies among the curves created by spatial or temporal proximity (Crainiceanu *and others*, 2009; Staicu *and others*, 2010; Scheffler *and others*, 2020; Campos *and others*, 2022). In the multivariate functional setting, several developments rely upon generalizations of the Karhunen–Loève representation (Jacques and Preda, 2014; Chiou *and others*, 2014), with extensions developed to handle functions observed over different evaluation domains (Happ and Greven, 2018), or to handle multivariate spatiotemporal data in a separable space-time covariance framework (Di Salvo *and others*, 2015; Ruggieri *and others*, 2018). However, to date, there has been no work proposed to our knowledge on multivariate FPCA that can incorporate spatial correlations in the observed data and model nonseparable multivariate space-time covariance structures.

Similar to developments proposed in Happ and Greven (2018), we propose a computationally efficient estimation procedure for MST-FPCA, which relies only on univariate FPCA expansions. An MCAR structure is induced on the vector of region-specific scores from the univariate FPCA expansions to target the correlation between outcomes in the multivariate response. The estimated between-outcome correlation is then incorporated into estimation of the multivariate eigenfunctions. Finally, the region-specific scores and spatial variation parameters are targeted via a Markov Chain Monte Carlo (MCMC) framework, conditional on the estimated multivariate eigenfunctions. MST-FPCA achieves computational efficiency by coupling dimension reduction in modeling nonparametric time trends through data-driven lower dimensional multivariate eigenfunctions with modeling of spatial correlations via a parametric CAR structure. Section 2 outlines the proposed MST-FPCA model, along with the proposed estimation and inference procedures. Applications of the proposed methodology to data from a large national database, USRDS, allow us to model spatiotemporal trends jointly in hospitalization risk and mortality rates across the United States, as outlined in Section 3. Finite sample properties of MST-FPCA are studied in simulations of Section 4, followed by a brief discussion given in Section 5.

2. Proposed MST-FPCA

2.1. *Model specification*

Let $i = 1, \ldots, n$ index regions, $k = 1, \ldots, T$ index time (months) after transition to dialysis and $j = 1, \ldots, J$ index the different dimensions of the J-dimensional outcome vector, $\boldsymbol{X}_i(t_k) = \{X_i^{(1)}(t_k), \ldots, X_i^{(J)}(t_k)\}^\top$. In our application to USRDS data, our multivariate outcome contains region-specific monthly hospitalization and mortality rates (i.e., $J = 2$), where region-specific rates are obtained as averages of dialysis facility-specific rates. Hospitalization and mortality rates at the dialysis facility level are defined as the ratio of the total number of patient hospitalizations or deaths to the total patient follow-up time for that specific facility at month k, respectively. We multiplied the rates by 12 so that the rate unit can be interpreted as a rate per person-year (PPY) consistent with annual national reporting from the USRDS. In our application to USRDS data, region units

are taken to be HSAs across the United States and region-specific rates are analyzed for a total of 24 months (i.e., 2-year follow-up) after transitioning to dialysis. Note that we opt to model the multivariate response as continuous data similar to the works of Li *and others* (2021) and amenable to the proposed FPCA framework.

To study the spatiotemporal variation in hospitalization and mortality rates jointly, the proposed MST-FPCA decomposes the J-dimensional multivariate response vector $\boldsymbol{X}_i(t) = \{X_i^{(1)}(t), \ldots, X_i^{(J)}(t)\}^\top$ $(t \in \mathcal{T})$ by

$$\boldsymbol{X}_i(t) = \boldsymbol{\mu}(t) + \sum_{\ell=1}^{\infty} \rho_{i\ell} \boldsymbol{\psi}_\ell(t) + \boldsymbol{\epsilon}_i(t). \tag{2.1}$$

In (2.1), $\boldsymbol{\mu}(t) = \{\mu^{(1)}(t), \ldots, \mu^{(J)}(t)\}^\top$ denotes the overall mean function, $\boldsymbol{\psi}_\ell(t) = \{\psi_\ell^{(1)}(t), \ldots, \psi_\ell^{(J)}(t)\}^\top$ denotes the multivariate eigenfunctions, $\rho_{i\ell}$ are region-specific principal component (PC) scores, and $\boldsymbol{\epsilon}_i(t) = \{\epsilon_i^{(1)}(t), \ldots, \epsilon_i^{(J)}(t)\}^\top$, $\epsilon_i^{(j)}(t) \sim_{ind} N(0, \sigma_j^2)$, denotes the measurement error. The multivariate eigenfunctions form an orthonormal system, i.e.,

$$\langle\langle \psi_\ell(t), \psi_{\ell'}(t) \rangle\rangle := \sum_{j=1}^{J} \langle \psi_\ell^{(j)}(t), \psi_{\ell'}^{(j)}(t) \rangle_2 = \sum_{j=1}^{J} \int_{\mathcal{T}} \psi_\ell^{(j)}(t) \psi_{\ell'}^{(j)}(t) \mathrm{d}t = \delta_{\ell\ell'},$$

where $\delta_{\ell\ell'} = 1$ for $\ell = \ell'$ and $\delta_{\ell\ell'} = 0$ otherwise. Under the classical multivariate FPCA framework (Ramsay and Silverman, 2005; Happ and Greven, 2018), the multivariate PC scores $\{\boldsymbol{\rho}_\ell = (\rho_{1\ell}, \ldots, \rho_{n\ell})^\top : \ell = 1, 2, \ldots\}$, are assumed to be uncorrelated, with zero means and $\mathrm{Var}(\boldsymbol{\rho}_{i\ell}) = \lambda_\ell$, where λ_ℓ denotes the eigenvalues. Similar to univariate FPCA, the multivariate eigenfunctions describe directions of leading modes of variation in the different dimensions of the functional response, while the eigenvalues quantify the amount of variation explained along the identified modes of variation. In practice, the expansion is truncated to include L eigencomponents based on the fraction of variance explained (FVE), where 2–3 eigencomponents are retained in most applications, leading to effective dimension reduction of the high-dimensional data.

Different from multivariate FPCA, we induce a CAR structure on the region-specific PC scores $\rho_{i\ell}$ to capture dependencies due to facility- and region-specific practice patterns and infrastructure. Specifically, let the $n \times n$ adjacency matrix $W = \{w_{ii'}\}$ describe the neighborhood structure of regions, where $w_{ii'} = 1$ if regions i and i' $(i \neq i')$ are neighbors, denoted by $i \sim i'$, and $w_{ii'} = 0$ otherwise. By convention, the diagonal elements of W are set to zero. Further let D be the diagonal matrix consisting of elements $d_i = \sum_{i \sim i'} w_{ii'}$, denoting the total number of neighbors of region i. Then, the full conditional distribution for the ℓth PC score for region i, $\rho_{i\ell}$, is specified as a weighted average of the ℓth PC scores from neighbors of region i, via the Markov Random Field $\rho_{i\ell} | \{\rho_{i'\ell}\}_{i \neq i'} \sim N(\nu \sum_{i' \sim i} w_{ii'} \rho_{i'\ell}/d_i, \alpha_\ell/d_i)$ with a variance component α_ℓ and a spatial correlation parameter ν. Through Brook's lemma, the joint distribution of the ℓth PC scores $\boldsymbol{\rho}_\ell = (\rho_{1\ell}, \ldots, \rho_{n\ell})^\top$ takes the form $\boldsymbol{\rho}_\ell \sim N\left(\mathbf{0}, \alpha_\ell (D - \nu W)^{-1}\right)$, where the spatial correlation parameter ν is constrained to lie between bounds given by the inverse of the minimum and maximum eigenvalues of the matrix $D^{-1/2} W D^{-1/2}$ in order for the precision matrix $(D - \nu W)/\alpha_\ell$ to be positive definite (Banerjee *and others*, 2015). The CAR model induced on the PC scores stabilizes estimation especially for smaller regions, smoothing out scores across neighbors. Note that even though the multiple outcomes considered share the same PC scores in the proposed modeling, this does not imply that the outcomes also share the same spatiotemporal correlation structure. Since the eigenfunctions are allowed to vary across outcomes (through the multivariate eigenfunctions), the proposed modeling is able to

induce different spatiotemporal correlations across outcomes (via the linear combination of PC scores multiplying multivariate eigenfunctions) that are nonseparable.

2.2. *Estimation and inference*

Similar to Happ and Greven (2018), the proposed MST-FPCA is also built on links to univariate FPCA of the different dimensions of the multivariate response vector for computational efficiency. The proposed estimation algorithm begins by expansions of the different dimensions of $\boldsymbol{X}_i(t)$ employing univariate FPCA (see Step 1 in Estimation Algorithm). The dependencies among the vector of univariate PC scores are modeled next, via an MCAR correlation structure, and used in targeting the multivariate eigenfunctions. Specifically, the variation in univariate PC scores is decomposed into a between eigencomponent (Σ_b) and a within eigencomponent (Σ_w) variation, capturing dependencies between the eigencomponents of the different dimensions of the response and spatial correlations across regions within eigencomponents, respectively (Step 2). The multivariate eigenfunctions $\boldsymbol{\psi}_\ell(t)$ are targeted using the estimated univariate eigenfunctions and eigenvectors of the estimated Σ_b (Step 3). Hence, by utilizing the dependency between the eigencomponents of the univariate response (through Σ_b) in deriving the multivariate eigenfunctions, MST-FPCA avoids multivariate FPCA on the higher dimensional covariance processes, which can be quite cumbersome in higher dimensions. Conditional on the estimated mean and multivariate eigenfunctions the multivariate PC scores $\rho_{i\ell}$ and spatial correlation parameters α_ℓ and ν are targeted within a hierarchical modeling framework via MCMC (Step 4). Finally, the region-specific multivariate response trajectories are reconstructed with pointwise credible intervals (CIs), relying on the estimated multivariate eigenfunctions and posterior distribution of the multivariate PC scores obtained in Step 4.

Direct estimation of the multivariate eigenfunctions in the first three steps, instead of expansion on a known bases set has two advantages. First it achieves dimension reduction, in the sense that in most applications a small number of multivariate eigenfunctions are enough to capture most of the variation in the data, leading to computational savings in the Bayesian modeling of Step 4. In addition, and perhaps more importantly, the multivariate eigenfunctions provide extra interpretations in model building. Because they are estimated from the data, they represent dominant modes of variation in time. Hence in the case of the multivariate eigenfunctions, they lead to the study and comparison of modes of temporal variation across outcomes.

Note that there are multiple computational savings utilized in the proposed estimation algorithm that makes the implementation of the proposed methodology feasible for decompositions of high-dimensional multivariate response. First, the proposed algorithm relies only on univariate FPCA decompositions. Second, the MCAR model induced on the univariate PC score vector is modeled as a linear combination of independent latent Gaussian processes with a CAR correlation structure. This lower dimensional representation allows for the efficient fitting of the MCAR model in WinBUGS. The hierarchical model with a CAR correlation structure on the multivariate PC scores in Step 4 is also implemented in WinBUGS, leading to easy implementation of the proposed algorithm. The R code and a tutorial for implementing MST-FPCA are made publicly available on Github (https://github.com/dsenturk/MST-FPCA). The proposed estimation algorithm is outlined in the table below, with specific steps discussed in further detail in this section.

Step 1: The proposed MST-FPCA begins by univariate FPCA employed in each dimension $X_i^{(j)}(t)$ of the multivariate outcome. A penalized spline smoother is used to obtain the univariate mean function $\hat{\mu}^{(j)}(t)$ (where the smoothing parameter can be selected via generalized cross-validation). The

Estimation Algorithm

Step 1: Employ univariate FPCA in each dimension ($X_i^{(j)}(t)$) of the multivariate response vector $\boldsymbol{X}_i(t)$ to target the estimated univariate eigenfunctions and univariate PC score vector $\hat{\boldsymbol{\xi}}$.

Step 2: Decompose dependencies among $\hat{\boldsymbol{\xi}}$ into a between eigencomponent Σ_b and a within eigencomponent Σ_w (between region) variation via an MCAR correlation structure.

Step 3: Target the multivariate eigenfunctions $\boldsymbol{\psi}_\ell(t)$ using the estimated univariate eigenfunctions and eigenvectors of the estimated Σ_b.

Step 4: Fixing the estimated mean and multivariate eigenfunctions, target the multivariate PC scores $\rho_{i\ell}$ and spatial correlation parameters α_ℓ and ν within a hierarchical modeling framework via MCMC.

Step 5: Provide inference on the multivariate response trajectories using the estimated multivariate eigenfunctions and posterior distribution of the multivariate PC scores obtained in Step 4.

estimated mean is used to center the observed data, $\hat{X}_i^{c(j)}(t) = X_i^{(j)}(t) - \hat{\mu}^{(j)}(t)$, leading to the empirical covariance, $\hat{G}^{(j)}(t,t') = \sum_{i=1}^n \hat{X}_i^{c(j)}(t)\hat{X}_i^{c(j)}(t')/n$. Following common practice, the diagonal entries of $\hat{G}^{(j)}(t,t')$ are removed before the covariance surface is smoothed using 2D penalized smoothing splines (Ramsay and Silverman, 2005). This is because the i.i.d measurement error inflates the error along the diagonal of the covariance. In addition, the smoothing parameters are selected by restricted maximum likelihood (Li *and others*, 2021). Once the covariance operators are obtained, estimators of the univariate eigenfunctions $\{\hat{\phi}_m^{(j)}(t) : m = 1, \ldots, M_j\}$, and PC scores $\{\hat{\xi}_{im}^{(j)} : i = 1, \ldots, n; m = 1, \ldots, M_j\}$ are recovered by FPCA. The numbers of eigencomponents, M_j, retained in the univariate FPCA expansions are determined by the FVE, where we use FVE $> 99\%$ in numerical applications to retain enough information at the initial step of the algorithm.

Step 2: Next, the vector of univariate PC scores $\hat{\boldsymbol{\xi}}_i^\top = (\hat{\xi}_{i1}^{(1)}, \ldots, \hat{\xi}_{iM_1}^{(1)}, \ldots, \hat{\xi}_{i1}^{(J)}, \ldots, \hat{\xi}_{iM_J}^{(J)})$ are modeled via an MCAR model. Let $\Xi = [\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_n]^\top$ denote the $n \times M^+$ score matrix with $M^+ = M_1 + M_2 + \cdots + M_J$ denoting the total number of eigencomponents retained across all J dimensions. Stacking the columns of Ξ leads to the $nM^+ \times 1$ score vector $\hat{\boldsymbol{\xi}} = \text{vec}(\Xi)$ that is modeled as a linear combination of M^+ $n \times 1$ independent latent spatial Gaussian processes $\mathbf{f}_\ell \sim N(\mathbf{0}, (D - \nu W)^{-1})$, $\ell = 1, \ldots, M^+$, as in Jin *and others* (2007), where D and W are as defined in Section 2.1, and ν denotes the common spatial smoothing parameter,

$$\underset{nM^+\times 1}{\hat{\boldsymbol{\xi}}} = \underset{nM^+\times 1}{\boldsymbol{\xi}} + \underset{nM^+\times 1}{\mathbf{e}} = (\underset{M^+\times M^+}{A} \otimes \underset{n\times n}{I}) \underset{nM^+\times 1}{\mathbf{f}} + \underset{nM^+\times 1}{\mathbf{e}}. \tag{2.2}$$

In (2.2), $\mathbf{f} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_{M^+}^\top)^\top$ denotes the vector of stacked latent Gaussian processes with the joint distribution $\mathbf{f} \sim N\left(\mathbf{0}, I_{M^+} \otimes (D - \nu W)^{-1}\right)$, $A = \{a_{\ell\ell'}\}$, $1 \le \ell' \le \ell \le M^+$, denotes an $M^+ \times M^+$ full rank lower triangular matrix and $\mathbf{e} \sim N(\mathbf{0}, \tau^2 I_{nM^+})$ denotes the vector of measurement errors, assumed to be uncorrelated with $\boldsymbol{\xi}$. With the proposed specification in (2.2), the covariance of $\boldsymbol{\xi}$, denoted by Σ, can be decomposed into a between eigencomponent (Σ_b) and a within eigencomponent (Σ_w) variation,

$$\begin{aligned}\underset{nM^+\times nM^+}{\Sigma} &= (\underset{M^+\times M^+}{A} \otimes \underset{n\times n}{I})\text{Bdiag}\{(D - \nu W)^{-1}, \ldots, (D - \nu W)^{-1}\}(A^\top \otimes I) \\ &= \underset{M^+\times M^+}{AA^\top} \otimes \underset{n\times n}{(D - \nu W)^{-1}} \equiv \underset{M^+\times M^+}{\Sigma_b} \otimes \underset{n\times n}{\Sigma_w},\end{aligned}$$

where $\Sigma_b = AA^\top$, $\Sigma_w = (D - \nu W)^{-1}$ and Bdiag$(\cdot)$ denotes a block-diagonal matrix. Hence, an MCAR (ν, Σ_b) structure is induced on $\boldsymbol{\xi}$. Representation of $\boldsymbol{\xi}$ as a linear combination of independent spatial Gaussian processes modeled with a CAR structure, via the lower triangular matrix A, allows for easy implementation of the MCAR model in WinBUGS, without the computational burden of operations with large covariance matrices and without the need to check for positive-definiteness of covariance matrices in each iteration of the algorithm. While the between-eigencomponent variation captures the dependency between the eigencomponents from univariate expansions, the within eigencomponent (between region) variation captures the spatial dependency among the regional units. Note that while a separable correlation structure is induced on the univariate PC scores, this does not imply a separable correlation structure between time and space, nor the same level spatial smoothing for the multivariate outcome vector. The goal of the modeling of $\boldsymbol{\xi}$ through an MCAR structure with a common between eigencomponent dependence across regions is to incorporate the estimated between eigencomponent dependency into estimation of the multivariate eigenfunctions in Step 3.

The parameters of the MCAR model are targeted via MCMC. Elementwise priors are imposed on the lower-triangular matrix A: $a_{\ell\ell} \sim \text{Lognormal}(\mu_{a_{\ell\ell}}, \sigma^2_{a_{\ell\ell}})$ and $a_{\ell\ell'} \sim N(\mu_{a_{\ell\ell'}}, \sigma^2_{a_{\ell\ell'}})$ for $1 \leq \ell' < \ell \leq M^+$. In addition, an Inverse Gamma (IG) (a_{τ^2}, b_{τ^2}) prior is imposed on the measurement error variance τ^2, and a Uniform prior is used for the spatial parameter ν with the parameters constrained to lie between bounds given by the inverse of the minimum and maximum eigenvalues of the matrix $D^{-1/2}WD^{-1/2}$ (denoted by a_ν, b_ν, respectively). The posterior distribution we seek can be expressed as

$$\pi(a_{\ell\ell}, a_{\ell\ell'}, \nu, \tau^2|\hat{\boldsymbol{\xi}}) \propto N(\hat{\boldsymbol{\xi}}|\boldsymbol{\xi}, \tau^2) \times N\left(\boldsymbol{\xi}|\mathbf{0}, AA^\top \otimes (D - \nu W)^{-1}\right) \times \prod_{\ell=1}^{M^+} \text{Lognormal}(a_{\ell\ell}|\mu_{a_{\ell\ell}}, \sigma^2_{a_{\ell\ell}}),$$
$$\times \prod_{\ell=1}^{M^+} \prod_{\ell'=1}^{\ell-1} N(a_{\ell\ell'}|\mu_{a_{\ell\ell'}}, \sigma^2_{a_{\ell\ell'}}) \times \text{Unif}(\nu|a_\nu, b_\nu) \times \text{IG}(\tau^2|a_{\tau^2}, b_{\tau^2}).$$

The model parameters are sampled from the posterior distributions using MCMC with Gibbs sampling and random walk metropolis as implemented in WinBUGS.

Step 3: The between eigencomponent dependence targeted via Σ_b is incorporated in estimation of the multivariate eigenfunctions as

$$\hat{\psi}^{(j)}_\ell(t) = \sum_{m=1}^{M_j} [\hat{\mathbf{c}}_\ell]^{(j)}_m \hat{\phi}^{(j)}_m(t), \quad \ell = 1, \ldots, M^+, \quad j = 1, \ldots, J, \tag{2.3}$$

where each dimension of the multivariate eigenfunction is targeted as a linear combination of the univariate eigenfunctions estimated for that dimension with weights $[\hat{\mathbf{c}}_\ell]^{(j)}_m$ equal to the mth entry of $[\hat{\mathbf{c}}_\ell]^{(j)} \in \mathbb{R}^{M_j}$, the jth block of the ℓth eigenvector $\hat{\mathbf{c}}_\ell$ of $\hat{\Sigma}_b$ (the posterior mean of Σ_b). The form in (2.3) follows from the fact that the variance components α_ℓ are also the eigenvalues of the between eigencomponent covariance matrix Σ_b of the univariate PC scores as shown in Appendix A of the supplementary material available at *Biostatistics* online. We defer the reader to Appendix A of the supplementary material available at *Biostatistics* online for more details.

Step 4: Once the multivariate mean and eigenfunctions are estimated, we target the region-specific PC scores $(\rho_{i\ell})$, spatial parameters (α_ℓ and ν), and measurement error variance (σ^2_j) of MST-FPCA,

conditional on $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{\psi}}_\ell(t), M^+\}$, via the multivariate hierarchical model

$$\boldsymbol{X}_i(t) = \hat{\boldsymbol{\mu}}(t) + \sum_{\ell=1}^{M^+} \rho_{i\ell}\hat{\boldsymbol{\psi}}_\ell(t) + \boldsymbol{\epsilon}_i(t),$$

$$\underset{n\times 1}{\boldsymbol{\rho}_\ell} = (\rho_{1\ell}, \ldots, \rho_{n\ell})^\top \sim N(\underset{n\times 1}{\mathbf{0}}, \alpha_\ell(\underset{n\times n}{D} - \nu \underset{n\times n}{W})^{-1}), \quad \underset{J\times 1}{\boldsymbol{\epsilon}_i(t)} \sim N(\underset{J\times 1}{\mathbf{0}}, \underset{J\times J}{\mathrm{Diag}(\sigma_1^2, \ldots, \sigma_J^2)}),$$

$$\sigma_j^2 \sim \mathrm{IG}(a_{\sigma_j^2}, b_{\sigma_j^2}), \quad \alpha_\ell \sim \mathrm{IG}(a_{\alpha_\ell}, b_{\alpha_\ell}), \quad \nu \sim \mathrm{Unif}(a_\nu, b_\nu),$$

where $\mathrm{Diag}(\sigma_1^2, \ldots, \sigma_J^2)$ denotes the $J \times J$ diagonal matrix with $\sigma_j^2, j = 1, \ldots, J$, on the diagonal (Step 4). Hence, the posterior distribution we seek can be expressed as

$$\begin{aligned}\pi(\boldsymbol{\rho}_\ell, \nu, \alpha_\ell, \sigma_j^2 | \boldsymbol{X}_i(t), \hat{\boldsymbol{\theta}}) \propto & \prod_{i=1}^{n}\prod_{k=1}^{T} N\left(\boldsymbol{X}_i(t_k) | \{\hat{\boldsymbol{\mu}}(t_k) + \sum_{\ell=1}^{M^+} \rho_{i\ell}\hat{\boldsymbol{\psi}}_\ell(t_k)\}, \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_J^2)\right) \\ & \times \prod_{\ell=1}^{M^+} N(\boldsymbol{\rho}_\ell | \mathbf{0}, \alpha_\ell(\mathrm{D} - \nu\mathrm{W})^{-1}) \times \prod_{\ell=1}^{M^+} \mathrm{IG}(\alpha_\ell | a_{\alpha_\ell}, b_{\alpha_\ell}) \\ & \times \mathrm{Unif}(\nu | a_\nu, b_\nu) \times \prod_{j=1}^{J} \mathrm{IG}(\sigma_j^2 | a_{\sigma_j^2}, b_{\sigma_j^2}).\end{aligned}$$

The model parameters are sampled from the posterior distributions using MCMC as implemented in WinBUGS. Since M^+ is the total number of univariate eigencomponents retained across the J dimensions, the total number (L) of multivariate eigencomponents used in MST-FPCA could be chosen to be smaller than M^+ in applications where M^+ may be large. Here, we recommend the use of the estimated variance components α_ℓ in formulating the FVE, since they are established as eigenvalues of the between eigencomponent covariance matrix Σ_b (see Appendix A of the supplementary material available at *Biostatistics* online for details).

Step 5: In the final step, region-specific trajectories are constructed for the multivariate outcome using the estimated MST-FPCA model components: $\hat{\boldsymbol{X}}_i(t) = \hat{\boldsymbol{\mu}}(t) + \sum_{\ell=1}^{L} \hat{\rho}_{i\ell}\hat{\boldsymbol{\psi}}_\ell(t)$. In addition, $(1-\gamma)$ pointwise CIs for region-specific trajectories are obtained by

$$\hat{\boldsymbol{X}}_i(t) \pm \Phi^{-1}\left(1 - \frac{\gamma}{2}\right)\sqrt{\mathrm{diag}\left[\mathrm{Var}\{\hat{\boldsymbol{X}}_i(t) - \boldsymbol{X}_i(t) | \hat{\boldsymbol{\theta}}\}\right]},$$

where $\mathrm{Var}\{\hat{\boldsymbol{X}}_i(t) - \boldsymbol{X}_i(t) | \hat{\boldsymbol{\theta}}\} \approx \hat{\Psi}\hat{\Omega}\hat{\Psi}^\top$, with $\hat{\Psi} = \{\hat{\boldsymbol{\psi}}_1(t), \ldots, \hat{\boldsymbol{\psi}}_L(t)\}$, $\Phi(\cdot)$ denotes the Gaussian cumulative distribution function, $\mathrm{diag}(\cdot)$ denotes the diagonal of a matrix and $\Omega_{L\times L} = \mathrm{Var}\{\boldsymbol{\rho}_i | \boldsymbol{X}(t), \hat{\boldsymbol{\theta}}\}$ for $\boldsymbol{\rho}_i = (\rho_{i1}, \ldots, \rho_{iL})^\top$ is targeted using the posterior variance of the estimated multivariate PC scores.

When the different dimensions of the multivariate outcome have different domains/ranges or if they exhibit different amounts of variation, standardization of the data as a preliminary step may be necessary in order to obtain interpretable multivariate functional principal components (Jacques and Preda, 2014; Chiou *and others*, 2014; Happ and Greven, 2018). The standardization can be carried out by rescaling data in each dimension using weights $s_j = \left\{\int_{\mathcal{T}} \hat{\mathrm{Var}}\left(X^{(j)}(t)\right) \mathrm{d}t\right\}^{-1}$, such that the integrated variance along the rescaled data $\tilde{X}^{(j)}(t) = s_j^{1/2} X^{(j)}(t)$ equals one (Happ and Greven,

2018). In this way, all dimensions of the multivariate outcome contribute equal amounts of variation to the analysis, similar to the standardization of classical multivariate PCA. In our data application, this rescaling is utilized. Note that when standardization is employed as a preliminary step in MST-FPCA, the algorithm uses the rescaled data in the first three steps in deriving the multivariate eigenfunctions. Once the multivariate eigenfunctions are estimated, the multivariate hierarchical model of Step 4 and inference in Step 5 utilizes the observed data as outlined in this section.

3. Data analysis

The USRDS collects data on nearly all patients with ESKD in the United States. Our study cohort includes patients aged 18 years or older who transitioned to dialysis between January 1, 2005, and September 30, 2013. Patients were followed for a maximum of 2 years, beginning on day 91 of dialysis (after 90 days to establish a stable treatment modality), with the last date of follow-up on December 31, 2015. Facility-specific hospitalization and mortality rates PPY are calculated monthly over the 2 year follow-up and are averaged within regions to yield the region-specific multivariate outcome. Consistent with national USRDS reporting, regions are taken to be Health Service Areas (HSAs), which are regions with relatively self-contained infrastructure for the provision of hospital care in the contiguous United States, including the District of Columbia. The final study cohort contains 367 regions/HSAs, where we defer detailed descriptions of the study cohort, exclusion rules, and preprocessing steps to Appendix B of the supplementary material available at *Biostatistics* online. The mean region-specific hospitalization and mortality rates are 1.810 and 0.073 PPY, respectively, where the raw hospitalization and mortality rate trajectories over the 2-year follow-up are given in Figure 1(a) and (c). Note that the range of the two outcomes is quite different and hence a preliminary rescaling is applied with $s_1 = 0.40$ and $s_2 = 215.07$, for hospitalization and mortality, respectively, as discussed in Section 2.2 before MST-FPCA, to guarantee that variation along both dimensions of the multivariate outcome is captured in the estimated multivariate eigenfunctions.

3.1. *Estimated MST-FPCA components*

The mean hospitalization and mortality trajectories are given in Figure 1(b) and (d), where the highest rates in both outcomes are observed within three months after transitioning to dialysis and where both (average) rates steadily decline during the first year on dialysis, consistent with previous literature (Foley *and others*, 2014). For example, the mean hospitalization rate at 1 month is 2.114 PPY and declines 16.84% to 1.758 PPY by 12 months; and similarly, the mean mortality rate at 1 month is 0.087 PYY and decreases to 0.069 PPY by 12 months (20.69% decline). While hospitalization rates continue to decline after the first year, the mortality rates remain relatively stable in the second year on dialysis. A total of four and two eigencomponents are retained in univariate FPCA expansions of hospitalization and mortality, respectively, explaining 99.2% and 99.3% of the total variation. The estimated spatial correlation parameter $\hat{\eta}$ equals 0.923, leading to the within eigencomponent correlations in the range of 0.27 to 0.63 in neighboring HSAs. The estimated multivariate eigenfunctions with respective estimated spatial variance parameters $\hat{\alpha}_1 = 3.090$, $\hat{\alpha}_2 = 1.966$, $\hat{\alpha}_3 = 0.211$, $\hat{\alpha}_4 = 0.117$, $\hat{\alpha}_5 = 0.103$, and $\hat{\alpha}_6 = 0.065$, are given in Figure S2 in the supplementary material available at *Biostatistics* online. The two leading multivariate eigenfunctions (given in Figures S2(a) and (b) in the supplementary material available at *Biostatistics* online) describe mostly constant variation in both hospitalization and mortality rates, with slightly higher variation in mortality in the first year of dialysis. The third, fourth, and fifth multivariate eigenfunctions (given in Figures S2 (c) and (d) in the supplementary material available at *Biostatistics* online) mainly explain variation in hospitalization, which highlight variation in the first and last 6 months on dialysis (third eigenfunction),

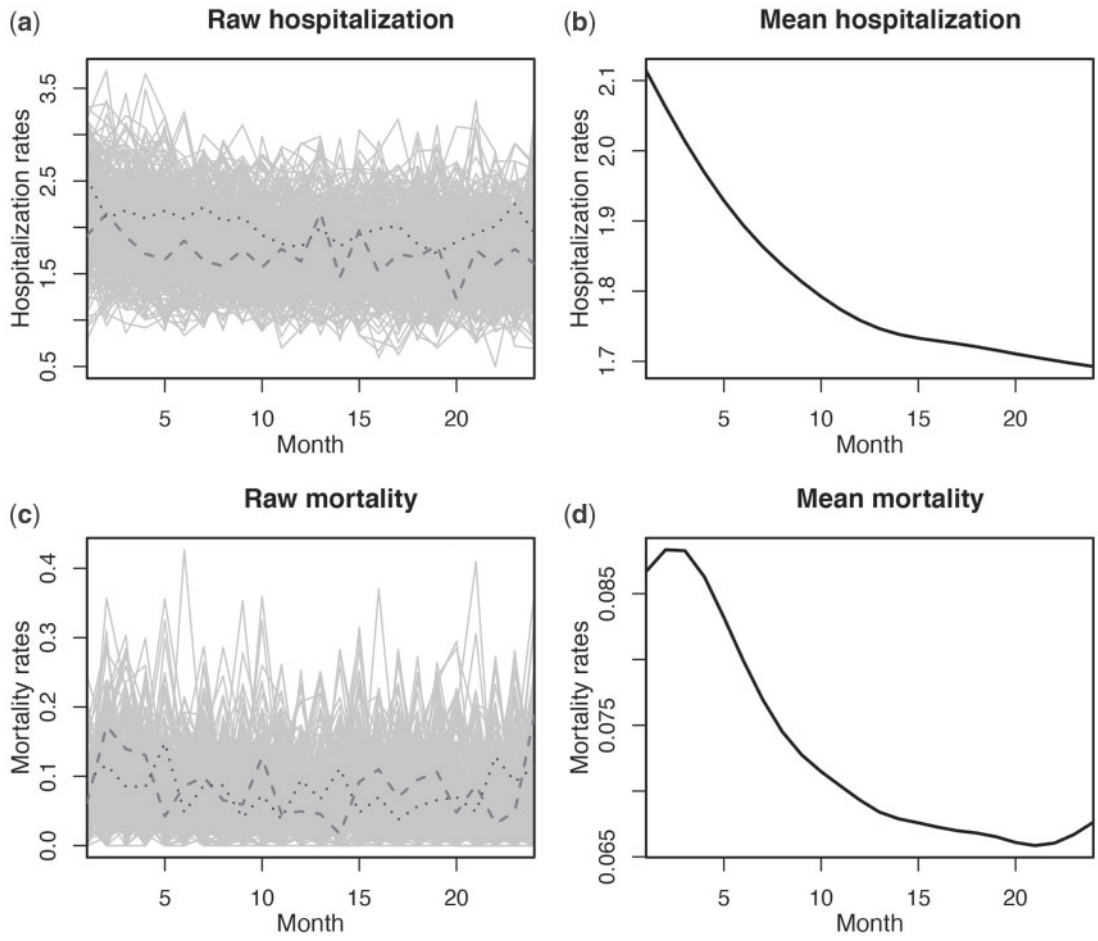

Fig. 1. Raw region-specific hospitalization (a) and mortality (c) trajectories. Raw trajectories for a randomly chosen two regions are depicted with dashed and dotted lines. The estimated mean trajectories for hospitalization and mortality are provided in (b) and (d), respectively.

followed by variation in the middle 1 year and end of the 2-year follow-up (fourth eigenfunction) and finally at 1 year and 18 months on dialysis (fifth eigenfunction). The last multivariate eigenfunction (given in Figures S2 (e) and (f) in the supplementary material available at *Biostatistics* online) mainly explains variation in mortality, which highlights variation in mortality at initiation and 1 year and 18 months after transitioning to dialysis. The leading time-varying variation in hospitalization observed within the first 6 months of dialysis is consistent with higher hospitalization rates observed at initiation of dialysis, while higher variation in the last 6 months of follow-up may be related to the decrease in the total number of patients towards the end of the 2-year follow-up.

3.2. *Inference for region-specific hospitalization and mortality rates*

The six estimated multivariate eigenfunctions are retained in the multivariate reconstruction of the region-specific hospitalization and mortality trajectories. The raw and predicted hospitalization and

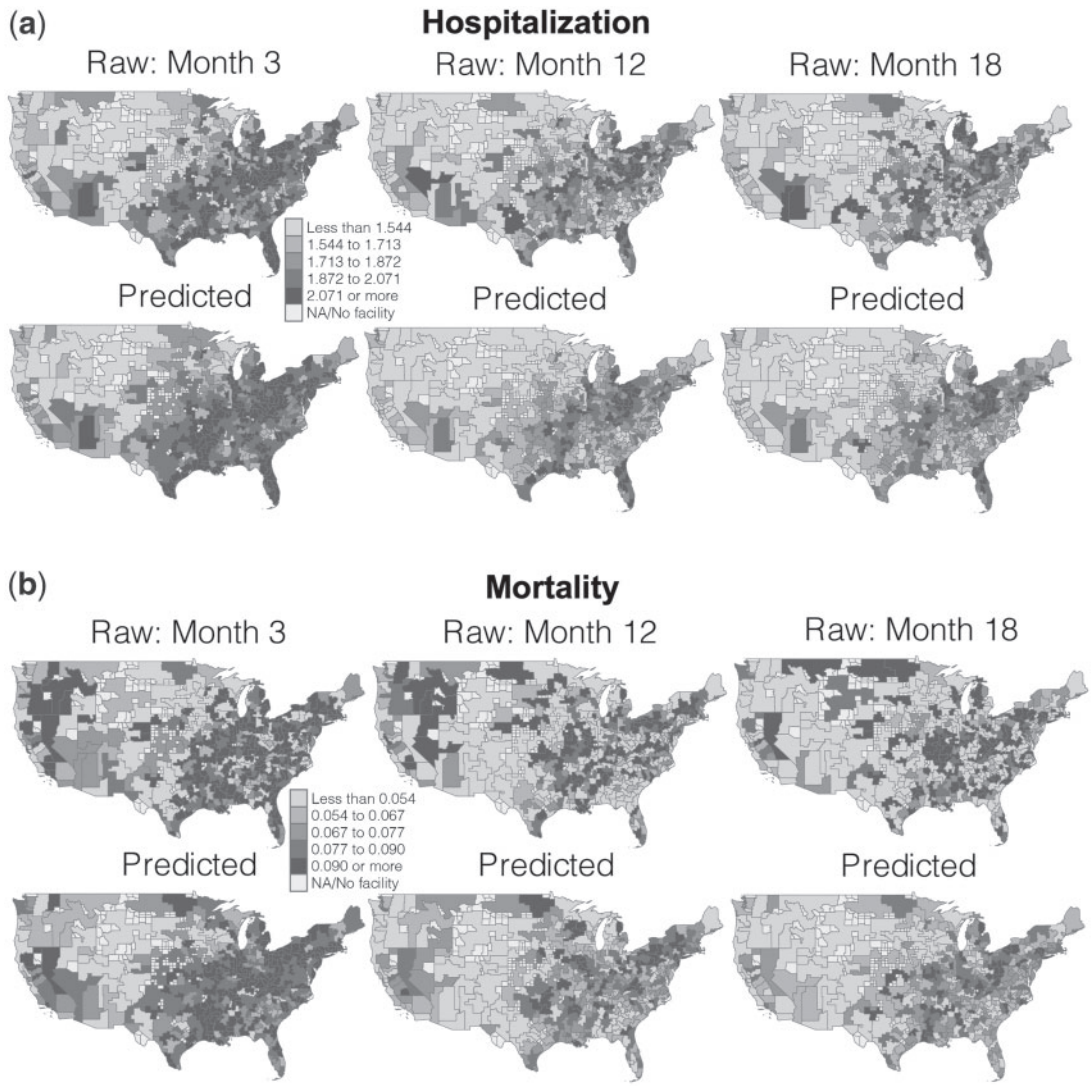

Fig. 2. The raw and predicted hospitalization (a) and mortality (b) rates from the 3rd, 12th, and 18th months on dialysis for all HSAs.

mortality rates from the 3rd, 12th, and 18th months on dialysis are displayed for all HSAs in Figure 2. There is a distinct pattern ("band") of higher rates (darker blue) in both hospitalization and mortality from Massachusetts to southern Texas. While regions display higher rates of hospitalization more consistently in this band, mortality rates are more variable with regions of elevated and lower mortality rates. In addition, there are outcome-specific patterns that emerge where some HSAs in northern California, Oregon, Montana, and Idaho have relatively low hospitalization rates but elevated mortality rates, whereas some HSAs in Florida and Arizona have relatively high hospitalization rates but not very high mortality rates. Furthermore, consistent with the estimated mean functions for hospitalization and mortality displayed in Figure 1 (b) and (d), respectively, the hospitalization and

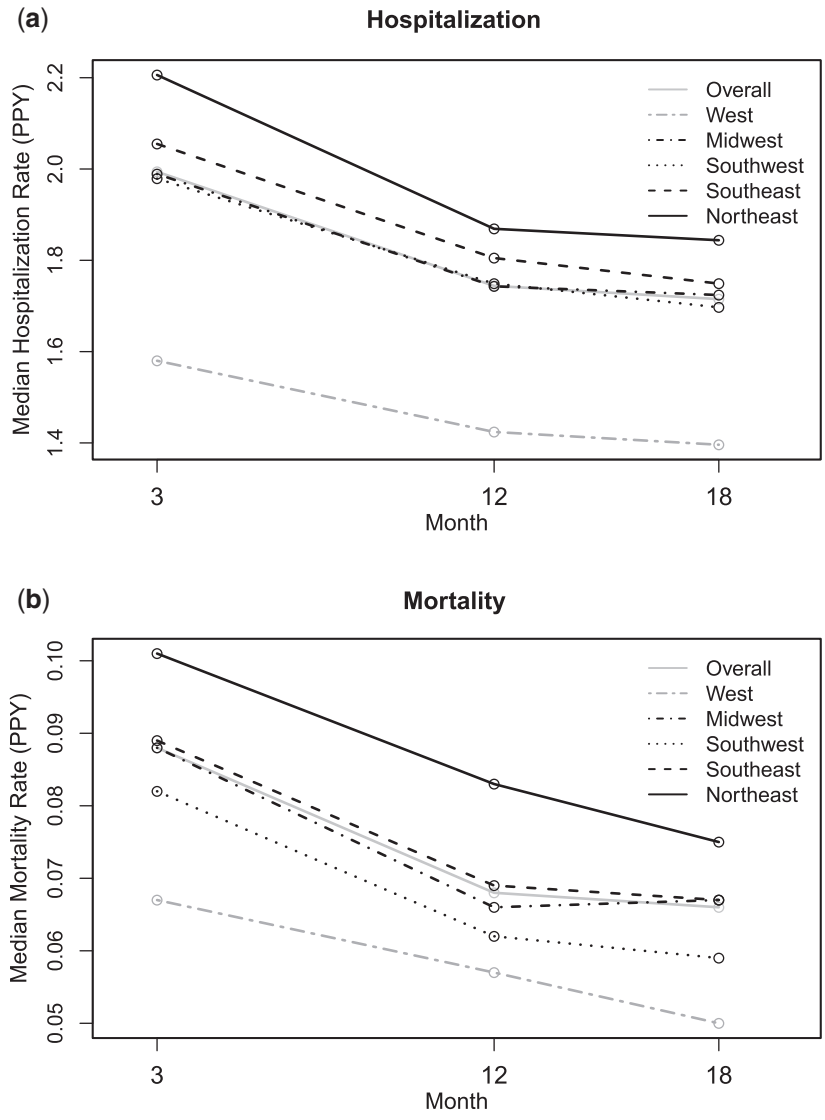

Fig. 3. Median predicted hospitalization (a) and mortality (b) rates PPY for five major zones, as well as the overall rates across all HSAs.

mortality rates are the highest in the first few months on dialysis, where hospitalization rates consistently decrease over the 2-year follow-up and mortality rates decrease within the first year on dialysis and remain roughly stable in the second year.

For a more in-depth study of the variation in hospitalization and mortality rates across the United States, we divide the regions into five major zones: West (41 HSAs), Midwest (98 HSAs), Southwest (41 HSAs), Southeast (141 HSAs), and Northeast (46 HSAs). Table S1 in the supplementary material available at *Biostatistics* online displays the median and the 5th and 95th percentiles of the predicted region-specific hospitalization and mortality rates at selected months 3, 12, and 18 across the five zones and Figure 3 displays the median rates. In addition, consistent with the estimated

mean functions for hospitalization and mortality in Figure 1 (b) and (d), respectively, both overall hospitalization and mortality rates decrease at a faster rate in the first year on dialysis, where they continue to decrease in the second year as well across the five zones but at a slower rate than the first year. More specifically, the median hospitalization rates decline from 3 to 18 months on dialysis by 0.184–0.362 PPY across five zones, which represents a 11.65%, 13.32%, 14.25%, 14.89%, and 16.41% decline in the West, Midwest, Southwest, Southeast, and Northeast, respectively. Similarly, the mortality rates decline in the first year on dialysis from 3 to 12 months by 0.010 to 0.022 PPY across five zones, which represents a 14.93%, 25.00%, 24.39%, 22.47%, and 18.82% decline in the West, Midwest, Southwest, Southeast, and Northeast, respectively (see Table S1 in the supplementary material available at *Biostatistics* online).

Figure S3 in the supplementary material available at Biostatistics online depicts the predicted hospitalization and mortality trajectories for five HSAs with median rates from the five zones. The predicted trajectories for HSAs from the five zones are plotted (in red) along with their 95% pointwise CI, overlaying predictions for a representative HSA with median rates across the entire United States (in blue) for comparison. Consistent with previous observations, hospitalization rates are the lowest in the West, where the CI for rates from the representative HSA does not overlap with the CI for the HSA from the West for most of the 2-year follow-up. While mortality rates are still the lowest in the West, the CIs overlap indicating more variability in mortality and lack of significant separation in the rates. This also aligns with the previous observation that some HSAs in the West (northern California) have higher mortality but lower hospitalization rates (Figure 2). In addition, the Northeast has the highest hospitalization and mortality rates, consistently over the 2-year follow-up where the CIs that compare rates to a representative HSA do not separate. Rates from HSAs from the remaining three zones in the Midwest, Southwest, and Southeast are more similar to those from the representative HSA.

Finally to assess model fit, we study the residuals obtained from MST-FPCA fits and compare the MST-FPCA fits to fits from three alternate methods of modeling the data. The map of absolute value of residuals from the fits at the 3rd, 12th, and 18th months on dialysis for all HSAs is given in Figure S1 in the supplementary material available at *Biostatistics* online. The residuals are quite small for hospitalizations, with larger deviations observed in less than 5% of HSA in the harder to model outcome of mortality. Most of these larger deviations are observed on the band of higher rates from Massachusetts to southern Texas and correspond to higher than expected mortality (positive residual). In addition, there does not seem to be an obvious spatial correlation in the residuals, implying that MST-FPCA is able to model spatial correlations in the data effectively. The first alternate method MST-FPCA is compared to (referred to as multivariate FPCA [M-FPCA]) assumes that the multivariate PC scores are i.i.d. across regions, ignoring spatial correlations in the data. M-FPCA is fitted through the algorithm of Happ and Greven (2018). The second comparative approach considered is called the Multivariate "Besag-York-Mollié" model (MBYM) proposed by Boulieri *and others* (2017). MBYM utilizes a random walk of order one to induce temporal correlations, and a multivariate intrinsic conditional autoregressive (ICAR) structure to model spatial patterns, leading to a separable discrete-time, discrete-space modeling framework. MBYM is easy to fit via the built-in `mv.car` package in WinBUGS, hence is utilized as a comparative method. The last alternate method considered is to fit univariate spatiotemporal FPCA separately in each dimension, which we refer to as U-FPCA. Similar to MST-FPCA, the U-FPCA approach targets the univariate eigenfunctions in the first step via FPCA. Then in each dimension, a CAR structure is induced among PC scores to capture dependencies among regions. The detailed description of the three alternate models is deferred to Appendix C of the supplementary material available at *Biostatistics* online. We compare the four fits using relative mean squared deviation error (MSDE), i.e., $\text{MSDE}_{\hat{x}^{(j)}(t)} = \int\{X_i^{(j)}(t) - \hat{X}_i^{(j)}(t)\}^2 / \int\{X_i^{(j)}(t)\}^2$. The mean MSDEs across regions are (0.0179, 0.0189,

0.0178, 0.0167) and (0.2587, 0.2564, 0.2552, 0.2454) in hospitalization and mortality for (M-FPCA, MBYM, U-FPCA, MST-FPCA), respectively. The run times for the four fits are (6.33, 81.31, 8.77, 12.02) minutes, for (M-FPCA, MBYM, U-FPCA, MST-FPCA), respectively. MST-FPCA and U-FPCA are fit with two parallel chains with 15000 iterations (5000 burn-in) and 3500 iterations (1000 burn-in) in the MCMC involved in Steps 2 and 4, respectively. M-FPCA and MYBM are also fit using two parallel chains with 5000 iterations (1000 burn-in) and 30 000 iterations (10 000 burn-in), respectively. The MST-FPCA leads to smaller MSDEs in modeling hospitalizations and modest improvements in modeling mortality, where the computational efficiency is pretty close to M-FPCA which ignores the spatial correlations and U-FPCA in lower dimensions. More information on comparisons of MST-FPCA to the three alternate methods in simulation studies are summarized in the next section.

4. Simulation studies

Simulation studies are conducted to examine the finite sample properties of the proposed MST-FPCA, as well as to compare its performance to the performance of M-FPCA, MBYM, and U-FPCA. Two simulation set-ups are considered: (i) independent multivariate response and (ii) correlated multivariate response, with varying number of time points ($T = 24, 50$), number of regions ($n = 49\,367$), different levels of error variance ($\sigma^2 = 0.02, 0.5$) and multivariate response from two and three dimensions. Simulation results are briefly reviewed here, where we defer details to Appendices D and E of the supplementary material available at *Biostatistics* online. Error measures for all model components get smaller with decreasing noise level σ^2. Similarly, all error measures decrease with increasing number of time points and regions, as expected, where the trends are stronger in estimation of multivariate eigenfunctions. Even with smaller number of time points, number of regions, and higher level of error variance, error measures from MST-FPCA signal a good fit (see Table 1).

The prediction and inference for the multivariate region-specific trajectories based on M-FPCA, MBYM, U-FPCA, and MST-FPCA are evaluated using MSDE and the coverage probability (CP) and length of the associated 95% pointwise CIs. Ignoring the spatial correlation in M-FPCA leads to higher MSDEs in recovery of the multivariate trajectories and lower CPs (compared to MST-FPCA and U-FPCA). In addition, the joint modeling of correlated outcomes via MST-FPCA leads to better efficiency than univariate modeling via U-FPCA, as expected, and the separable covariance structure of MBYM across time and space, coupled with constant spatial correlation assumed across outcomes via ICAR, is too restrictive and leads to worse performance compared to the other comparative methods (see Table 2).

5. Discussion

We proposed a novel MST-FPCA model to study the joint spatiotemporal patterns of hospitalization and mortality rates among dialysis patients in the United States. The proposed MST-FPCA not only allows for leveraging information from nearby rates (over geographic regions and time) but also across correlated outcomes to achieve a more comprehensive assessment of variation in the multivariate outcome. The proposed estimation involves effective dimension reduction in modeling variation across time through multivariate eigenfunctions, which provide additional interpretations on directions of dominant temporal variation. Spatial correlations on univariate PC scores are modeled via MCAR. Both the multivariate eigenfunctions and the proposed MCAR modeling are built on computationally feasible representations involving lower dimensional building blocks (univariate

Table 1. *The mean MSE and MSDE values for estimation of the model components from the correlated response simulation set-up with varying number of time points T, varying number of regions n and different levels of measurement error variance σ^2. Results are based on 200 Monte Carlo runs*

T	24 time points				50 time points			
n	49 regions		367 regions		49 regions		367 regions	
σ^2	0.02	0.5	0.02	0.5	0.02	0.5	0.02	0.5
	MSDE							
$\hat{\psi}_1^{(1)}(t)$	0.079	0.116	0.030	0.033	0.065	0.104	<0.001	0.001
$\hat{\psi}_1^{(2)}(t)$	0.079	0.115	0.030	0.034	0.065	0.104	<0.001	0.001
$\hat{\psi}_2^{(1)}(t)$	0.090	0.125	0.035	0.039	0.074	0.105	0.001	0.002
$\hat{\psi}_2^{(2)}(t)$	0.090	0.127	0.035	0.039	0.074	0.104	0.001	0.002
$\hat{\psi}_3^{(1)}(t)$	0.010	0.048	<0.001	0.004	0.010	0.030	<0.001	0.002
$\hat{\psi}_3^{(2)}(t)$	0.010	0.051	<0.001	0.005	0.010	0.030	<0.001	0.003
	MSE							
$\hat{\alpha}_1$	0.057	0.127	0.047	0.051	0.006	0.030	<0.001	0.002
$\hat{\alpha}_2$	0.014	0.052	0.007	0.007	0.003	0.045	<0.001	0.002
$\hat{\alpha}_3$	0.023	0.105	0.021	0.024	0.007	0.085	0.004	0.005
$\hat{\nu}$	0.146	0.157	0.054	0.059	0.144	0.152	0.019	0.049
$\hat{\sigma}^2$	<0.001	0.002	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
$\hat{\rho}_{i1}$	0.221	0.484	0.050	0.080	0.208	0.475	0.011	0.027
$\hat{\rho}_{i2}$	0.364	0.392	0.120	0.170	0.316	0.342	0.002	0.027
$\hat{\rho}_{i3}$	0.080	0.292	0.009	0.140	0.077	0.215	0.009	0.072

eigenfunctions and independent latent Gaussian processes with a CAR correlation structure, respectively). Hence, the proposed modeling, while leading to a nonseparable space and time covariance structure in the outcomes, can still easily scale up to multivariate response in higher dimensions. In addition, the proposed modeling via MCMC allows for estimation and inference on multivariate hospitalization and mortality trajectories in order to obtain regional hot spots (with high rates in both hospitalization and mortality or with differing patterns in the outcomes) and time periods with elevated rates after transitioning to dialysis. Results point to significant spatiotemporal variation in the multivariate outcome across the United States.

Since the estimated multivariate mean and eigenfunctions are considered fixed in the Bayesian modeling (Step 4), the proposed inference can underestimate uncertainty in small data applications where estimation error in targeting the mean and eigenfunctions are nonnegligible. For small data applications, corrected inference has been proposed in functional data settings via parametric bootstrap (Li *and others*, 2021). A similar extension could enable MST-FPCA to incorporate the stochasticity in estimation of the mean and multivariate eigenfunctions and is the topic for future research. We follow previous works (Quick *and others*, 2013; Li *and others*, 2021) in modeling rates directly, amenable to functional data analysis techniques used, however, another extension of MST-FPCA that is of interest is in modeling generalized outcome. Multivariate FPCA has been extended for generalized outcome utilizing a semiparametric latent process (Jiang *and others*, 2022). Multivariate eigenfunctions can be incorporated into a generalized Bayesian hierarchical framework in such an extension, with a need for a new set of tools for computational savings. Finally, the proposed methodology can be extended for modeling time-varying risk factors that may explain parts

Table 2. *The mean MSDE of the predicted multivariate region-specific trajectories and the CP and length of the associated 95% pointwise CIs based on M-FPCA, MBYM, U-FPCA, and MST-FPCA from both independent and correlated response simulation set-ups for number of time points $T = 24$, number of regions $n = 367$ and varying measurement error variance σ^2. Results are based on 200 Monte Carlo runs*

Number of regions, n	367 regions							
Noise level, σ^2	0.02				0.5			
Model	M-FPCA	MBYM	U-FPCA	MST-FPCA	M-FPCA	MBYM	U-FPCA	MST-FPCA
	Independent response							
MSDE: $\hat{x}_i^{(1)}(t)$	0.071	1.031	0.016	0.015	0.505	1.048	0.363	0.372
MSDE: $\hat{x}_i^{(2)}(t)$	0.068	0.515	0.016	0.016	0.478	0.604	0.330	0.336
Length$^{(1)}$	0.194	0.409	0.194	0.194	0.867	0.516	0.870	0.874
Length$^{(2)}$	0.192	0.427	0.192	0.192	0.857	0.664	0.868	0.865
CP$^{(1)}$ (%)	90.11	21.67	94.09	94.08	91.87	27.06	94.53	94.42
CP$^{(2)}$ (%)	90.44	42.62	94.30	94.29	91.11	54.99	94.57	94.05
	Correlated response							
MSDE: $\hat{x}_i^{(1)}(t)$	0.073	1.032	0.016	0.011	0.474	1.048	0.284	0.188
MSDE: $\hat{x}_i^{(2)}(t)$	0.072	1.031	0.016	0.012	0.471	1.048	0.291	0.190
Length$^{(1)}$	0.157	0.408	0.194	0.179	0.750	0.515	0.877	0.755
Length$^{(2)}$	0.157	0.408	0.194	0.179	0.749	0.514	0.874	0.754
CP$^{(1)}$ (%)	91.13	21.61	94.09	95.13	91.52	27.01	94.56	96.19
CP$^{(2)}$ (%)	90.49	22.64	93.72	94.63	90.68	28.08	94.48	96.06
	Run times							
	M-FPCA		MBYM		U-FPCA		MST-FPCA	
Time (min)	6.330		81.307		8.774		12.022	

of the variation observed. The extension would expand the mean function $\mu(t)$ (and the possibly time-varying covariate effect functions) on a common basis system whose coefficients can be targeted within the proposed Bayesian hierarchical modeling framework. This can lead to additional insights on potentially modifiable risk factors that may contribute to elevated hospitalization and mortality risk and inform targeted patient care.

6. Software

The R code and documentation for implementing the MST-FPCA on simulated datasets are provided on Github at https://github.com/dsenturk/MST-FPCA.

Supplementary material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

Funding

The National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK092232 to D.S., D.V.N., E.K., S.B., C.M.R., Q.Q., and Y.L.). The data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of the data presented here are the responsibility of the author(s) and in no way should be seen as an official policy or interpretation of the United States government.

Conflict of Interest: None declared.

References

Baer, D. R., Lawson, A. B. and Joseph, J. E. (2021). Joint space–time bayesian disease mapping via quantification of disease risk association. *Statistical Methods in Medical Research* **30**, 35–61.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data*, 2nd Edition. New York: CRC Press.

Boulieri, A., Liverani, S., de Hoogh, K. and Blangiardo, M. (2017). A space–time multivariate Bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**, 119–139.

Campos, E., Wolfe Scheffler, A., Telesca, D., Sugar, C., DiStefano, C., Jeste, S., Levin, A. R., Naples, A., Webb, S. J., Shic, F. *and others*. (2022). Multilevel hybrid principal components analysis for region-referenced functional electroencephalography data. *Statistics in Medicine*. **41**, 3737–3757.

Chiou, J., Chen, Y. and Yang, Y. (2014). Multivariate functional principal component analysis: a normalization approach. *Statistica Sinica*, **24**, 1571–1596.

Crainiceanu, C. M., Staicu, A.-M. and Di, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association* **104**, 1550–1561.

Di Salvo, F., Ruggieri, M. and Plaia, A. (2015). Functional principal component analysis for multivariate multidimensional environmental data. *Environmental and Ecological Statistics* **22**, 739–757.

Estes, J. P., Nguyen, D. V., Chen, Y., Dalrymple, L. S., Rhee, C. M., Kalantar-Zadeh, K. and Şentürk, D. (2018). Time-dynamic profiling with application to hospital readmission among patients on dialysis. *Biometrics* **74**, 1383–1394.

Foley, R. N., Chen, S.-C., Solid, C. A., Gilbertson, D. T. and Collins, A. J. (2014). Early mortality in patients starting dialysis appears to go unregistered. *Kidney International* **86**, 392–398.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association* **113**, 649–659.

Hepler, S. A., Waller, L. A. and Kline, D. M. (2021). A multivariate spatiotemporal change-point model of opioid overdose deaths in Ohio. *The Annals of Applied Statistics* **15**, 1329–1342.

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* **71**, 92–106.

Jiang, J., Lin, H., Zhong, Q. and Li, Y. (2022). Analysis of multivariate non-gaussian functional data: a semiparametric latent process approach. *Journal of Multivariate Analysis* **189**, 104888.

Jin, X., Banerjee, S. and Carlin, B. P. (2007). Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 817–838.

Li, Y., Nguyen, D. V., Banerjee, S., Rhee, C. M., Kalantar-Zadeh, K., Kürüm, E. and Şentürk, D. (2021). Multilevel modeling of spatially nested functional data: spatiotemporal patterns of hospitalization rates in the United States dialysis population. *Statistics in Medicine* **40**, 3937–3952.

Li, Y., Nguyen, D. V., Chen, Y., Rhee, C. M., Kalantar-Zadeh, K. and Şentürk, D. (2018). Modeling time-varying effects of multilevel risk factors of hospitalizations in patients on dialysis. *Statistics in Medicine* **37**, 4707–4720.

Quick, H., Banerjee, S. and Carlin, B. P. (2013). Modeling temporal gradients in regionally aggregated California asthma hospitalization data. *The Annals of Applied Statistics* **7**, 154.

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

Ruggieri, M., Plaia, A. and Di Salvo, F. (2018). Comparing spatial and spatio-temporal fpca to impute large continuous gaps in space. In: Mola F. *and others* (editors), *Classification,(Big) Data Analysis and Statistical Learning*. Switzerland: Springer, 201–208.

Scheffler, A., Telesca, D., Li, Q., Sugar, C. A., Distefano, C., Jeste, S. and Şentürk, D. (2020). Hybrid principal components analysis for region-referenced longitudinal functional EEG data. *Biostatistics* **21**, 139–157.

Staicu, A.-M., Crainiceanu, C. M. and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics* **11**, 177–194.

USRDS. (2021). United States Renal Data System 2021 Annual Data Report: "Epidemiology of Kidney Disease in the United States". *Technical Report*, National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

Zhang, S., Sun, D., He, C. Z. and Schootman, M. (2006). A Bayesian semi-parametric model for colorectal cancer incidences. *Statistics in Medicine* **25**, 285–309.

[*Received November 9, 2022; revised March 14, 2023; accepted for publication May 30, 2023*]