Biostatistics (2023) **24**, 2, *pp.* 358–371
doi: 10.1093/biostatistics/kxab032
Advance Access publication on August 26, 2021

Predicting the onset of breast cancer using mammogram imaging data with irregular boundary

SHU JIANG*

Division of Public Health Sciences, Washington University School of Medicine, MO, USA, 63110

JIGUO CAO

Department of Statistics and Actuarial Science, Simon Fraser University, BC, Canada, V5A 1S6

GRAHAM A. COLDITZ

Division of Public Health Sciences, Washington University School of Medicine, MO, USA, 63110

BERNARD ROSNER

Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, MA, USA, 02115 Department of Biostatistics, Harvard T.H. Chan School of Public Health, MA, USA, 02115

jiang.shu@wustl.edu

Summary

With mammography being the primary breast cancer screening strategy, it is essential to make full use of the mammogram imaging data to better identify women who are at higher and lower than average risk. Our primary goal in this study is to extract mammogram-based features that augment the well-established breast cancer risk factors to improve prediction accuracy. In this article, we propose a supervised functional principal component analysis (sFPCA) over triangulations method for extracting features that are ordered by the magnitude of association with the failure time outcome. The proposed method accommodates the irregular boundary issue posed by the breast area within the mammogram imaging data with flexible bivariate splines over triangulations. We also provide an eigenvalue decomposition algorithm that is computationally efficient. Compared to the conventional unsupervised FPCA method, the proposed method results in a lower Brier Score and higher area under the ROC curve (AUC) in simulation studies. We apply our method to data from the Joanne Knight Breast Health Cohort at Siteman Cancer Center. Our approach not only obtains the best prediction performance comparing to unsupervised FPCA and benchmark models but also reveals important risk patterns within the mammogram images. This demonstrates the importance of utilizing additional supervised image-based features to clarify breast cancer risk.

Keywords: Functional principal component analysis; Image analysis; Supervised learning; Survival analysis; Triangulation.

1. Introduction

Accurate assessment of risk is a top priority in oncology due to the population burden of cancer and advancing technologies for precision oncology. Breast cancer is the leading cancer diagnosis among

*To whom correspondence should be addressed.

women, with more than 2.1 million new cases identified worldwide each year (Bray *and others*, 2018). With routine mammography being the primary breast cancer screening strategy, it is essential to make full use of the mammogram images and derive features that can provide accurate mammography-based breast cancer risk to better identify women who are at very high or low future risk. The problem that motivated this study comes from the Joanne Knight Breast Health (JKBH) Cohort at Siteman Cancer Center. The cohort was established to link breast cancer risk factors, mammographic breast density, and blood markers in a diverse population of women undergoing routine mammographic screening. The primary goal of this article is to build a prognostic model for prevention, using additional unexplored image-based markers, augmenting the well-known breast cancer risk factors. The challenge lies in modeling such high-dimensional mammogram imaging data, as the total number of pixels will greatly exceed the number of women in the cohort, making the model nonidentifiable. An efficient dimension reduction technique is thus needed.

Functional principal component analysis (FPCA) is a popular approach for dimension reduction. It aims to identify major sources of variability among functional data, for example, data recorded as curves, images, or other objects over a continuum, usually time or spatial locations (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). The objective of FPCA is to project the infinite-dimensional functional data to a low-dimensional vector space defined by the functional principal components (FPCs). A considerable amount of literature has been published on FPCA for 1D functional data. For instance, Ramsay and Dalzell (1991) and Rice and Silverman (1991) developed smoothed FPCA methods to control the smoothness of FPCs. Yao *and others* (2005) proposed to calculate the FPC scores based on a conditional expectation when functional data are collected on sparse points. Chen and Lei (2015), Lin *and others* (2016), and Nie and Cao (2020) proposed to estimate FPCs which are only nonzero in a small interval in order to enhance the interpretability of FPCs. Supervised methods with functional covariates have also been considered in the literature. Kong *and others* (2018) and Gellar *and others* (2015) considered Cox regression model with functional covariates with survival outcome. Li *and others* (2016) proposed the supervised sparse functional principal component analysis, where they assumed that additional scalar supervision variables drive the low-rank structure of the functional predictor. Nie *and others* (2018) considered supervised FPCA in identifying the set of FPCs to boost the prediction performance. Zhang *and others* (2021) discussed the supervised principal component regression on a newly defined expected integrated residual sum of squares.

Less work has been proposed on FPCA methods for imaging data, that is, 2D functional data. Zipunnikov *and others* (2011) explored a connection between the singular value decomposition and FPCA for analyzing brain images by transferring images to a long vector. Other approaches for functional or smooth principal component analysis for tensor data have been proposed (see Huang *and others*, 2009; Allen, 2013; Lin *and others*, 2015 for example). Of note, these articles are based on the assumption that the images are bounded within a rectangular domain.

One unique challenge that arises in mammogram images is that the breast area appearing in the mammograms are irregularly shaped. Mammograms from two randomly chosen women within the JKBH cohort are shown in Figure S1 of the Supplementary material available at *Biostatistics* online, where it is apparent that the breast area is bounded within a semicircular region.

When the images are bounded within an arbitrary domain, triangulation is the most convenient tool to partition the domain into pieces. Bivariate spline, smooth piecewise polynomial functions over triangulations, is a natural approach that can be used to preserve important features of imaging data (Lai and Wang, 2013). Wang *and others* (2020) recently considered modeling the imaging data under the FPCA framework with triangulation. However, the conventional FPCA method does not consider the relationship between the functional predictor and the response variable (failure time under the survival context). Thus, the ordering of the eigenfunctions does not indicate the degree of association with the failure time, which may be sub-optimal when our goal lies in prediction.

In this article, we propose a novel supervised FPCA (sFPCA) over triangulations method for extracting image-based features that are ordered by the magnitude of association with failure time. This newly proposed method adopts the bivariate splines over triangulation to accommodate the irregularly shaped boundary of the mammogram imaging data to avoid "leakage" over complex domains. The sFPCA method is demonstrated to be versatile in simulation studies, where effective dimension reduction is achieved by choosing the top few sFPCs that cumulatively explain a large proportion of association with the failure time. We then apply the proposed sFPCA method to the Joanne Knight Breast Health Cohort at Siteman Cancer Center to leverage new insights from the rich set of mammograms in relation to the onset of breast cancer. The prediction performance for the proposed method is compared to the conventional FPCA method, and further compared to benchmark models.

To this end, we summarize the main contributions of this article. First, we accommodate the irregular boundary of the mammogram imaging data with bivariate splines over triangulations. Second, building on the foundation of triangulation, we propose a novel supervised FPCA over triangulations method that enables us to estimate the supervised FPCs ordered by the magnitude of association with the failure time. Third, we provide an eigenvalue decomposition algorithm that is computationally efficient. Last, we leverage new insights from the motivating mammogram imaging data in the Joanne Knight Breast Health Cohort at Siteman Cancer Center.

This article is organized as follows. In Section 2, we introduce notation and the model setup. Specifically, we discuss the bivariate splines over triangulation and propose a novel sFPCA method that enables us to estimate the supervised FPCs. The connection between prediction under the survival context with the estimated sFPCs is also discussed here. We investigate the finite sample performance of the proposed method via intensive simulation studies in Section 3. The proposed method is then applied to the motivating Joanne Knight Breast Health Cohort in Section 4. Concluding remarks are given in Section 5.

2. MODEL AND ESTIMATION

We first let T_i and C_i be the time of event occurrence and time of censoring for an individual i, respectively. The observed time is denoted by $\widetilde{T}_i = \min(T_i, C_i)$, with $\triangle_i = I(T_i < C_i)$ indicating that the observed time is an event time. Further, we let Ω be a bounded 2D domain of arbitrary shape, and $\mathbf{s} = (s_1, s_2)$ represent a particular point $\in \Omega \subset \mathbb{R}^2$. The imaging data can then be defined as $\{Z_i(\mathbf{s}), \forall \mathbf{s} \in \Omega\}$, $i = 1, \ldots, n$.

Under the functional framework, the observed imaging data Z_i can be viewed as realizations of a stochastic process $\{Z(\mathbf{s}), \forall \mathbf{s} \in \Omega\}$. If Z is a square-integrable process in space with continuous covariance function, the process is known to have a Karhunen–Loève expansion,

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \sum_{k=1}^{\infty} \lambda_k \psi_k(\mathbf{s}),$$

where $\mu(\mathbf{s})$ is the mean function, λ_k is the kth functional score that is defined as the inner product between Z and the FPC ψ_k, that is, $\langle Z - \mu, \psi \rangle = \int_{\mathbf{s}\in\Omega} (Z(\mathbf{s}) - \mu(\mathbf{s}))\psi_k(\mathbf{s})\mathrm{d}\mathbf{s}$. Without loss of generality, we assume that Z is demeaned from here on, that is, $\mu(\mathbf{s}) = 0$. In practice, this infinite summation is usually truncated to the first $K^\dagger$ terms that are sufficient to explain the first $K^\dagger$ major sources of variation in the functional data. However, the projection of the functional data $Z(\mathbf{s})$ onto the first $K^\dagger$ corresponding FPCs ordered by the explained variation may not be associated with the response variable, which is suboptimal when our goal lies in prediction.

In what follows, we will first consider the estimation of $Z(\mathbf{s})$ using bivariate splines that are piecewise polynomial functions over a 2D triangulated domain in Section 2.1 to accommodate the semicircular shaped domain in mammogram images. Then, building on approximation over triangulations, we will introduce the supervised FPCA (sFPCA) method in Section 2.2, where the features extracted will be

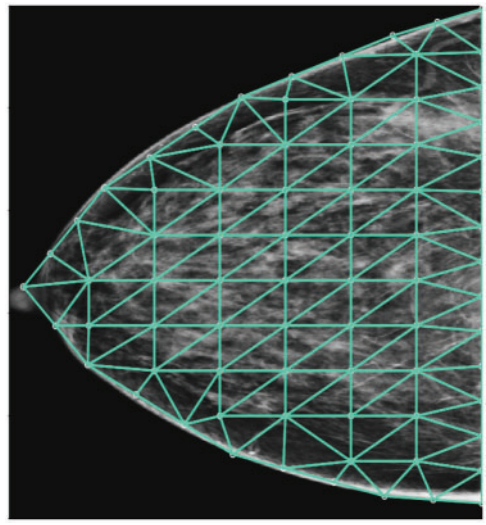

Fig. 1. An example of triangulation within mammogram with 115 triangles.

ordered by the magnitude of association with the failure time. Finally, we establish a connection between the sFPCA method and survival analysis in Section 2.3.

2.1. *Bivariate spline basis approximation over triangulation*

Triangulation is a well-known and effective tool for handling data with complex boundaries and/or interior holes (Lai and Wang, 2013). We let τ be a triangle that is the convex hull of three points that are not collinear. A collection $\triangle = \{\tau_1, \ldots, \tau_M\}$ of triangles is called a triangulation of $\Omega = \bigcup_{m=1}^{M} \tau_m$ if any nonempty intersection between a pair of triangles in $\triangle$ is either a common vertex or a common edge. For an integer $r \geq 0$, let $C^r(\Omega)$ be the collection of all r-times continuously differentiable functions over Ω. Then, let $T_d^r(\triangle) = \{t \in C^r(\Omega) : t|_\tau \in \mathbb{P}_d, \tau \in \triangle\}$ be a spline space of degree d and smoothness r over triangulation $\triangle$, where $t|_\tau$ is the polynomial piece of spline t restricted to triangle τ, and $\mathbb{P}_d$ is the space of all polynomials of degree less than or equal to d. We then define the Bernstein basis polynomial of degree d as $B_{ijk}^{\tau,d}(\mathbf{s}) = (i!j!k!)^{-1} d! b_1^i b_2^j b_3^k$, where b_1, b_2, and b_3 are the barycentric coordinates of $\mathbf{s} \in \Omega$ relative to $\tau \in \triangle$, and $i + j + k = d$. We show the triangulation over the mammogram imaging data in Figure 1. We can see that the semicircular boundary is well-approximated over triangulations. For illustration purposes, we have also constructed the Bernstein polynomial basis function with $d = 2$ in Figure S2 of the Supplementary material available at *Biostatistics* online.

In practice, the image data are only measured at a finite grid of pixels, $\mathbf{s}_j \in \Omega$, for $j = 1, \ldots, J$. We can first smooth the image data $\{\mathbf{s}_j, Z_i(\mathbf{s}_j)\}_{j=1}^J$ individually by employing the bivariate spline smoothing method. Let $\mathbf{B}(\mathbf{s}) = (B_1(\mathbf{s}), \ldots, B_K(\mathbf{s}))^T$ be a vector of degree-d bivariate Bernstein basis polynomials for $T_d^r(\triangle)$, where K is the number of Bernstein basis polynomials. Our objective is to estimate the 2D function $g_i(\mathbf{s}) = \mathbf{a}_i^T \mathbf{B}(\mathbf{s})$ by minimizing $\sum_{i=1}^n \sum_{j=1}^J \{Z_i(\mathbf{s}_j) - g_i(\mathbf{s}_j)\}^2 + \rho_i \cdot \mathtt{PEN}(g_i)$, where a_i is the vector of coefficients to the bivariate Bernstein basis polynomials, ρ_i is the roughness penalty parameter, and the roughness penalty $\mathtt{PEN}(g_i) = \sum_{\tau \in \triangle} \int_\tau \sum_{p+q=2} \binom{2}{p} (D_{s_1}^p D_{s_2}^q g_i)^2 \mathrm{d}s_1 \mathrm{d}s_2$ with $D_{s_1}^p$ being the pth-order derivative in the direction s_1, $\mathbf{s} = (s_1, s_2)$. We also need to impose some linear constraints on the spline coefficients $\mathbf{a}$, $\mathbf{Ha} = 0$, such that the smoothness requirement of the splines are met. Examples of the

H matrix can be found in Yu *and others* (2020). The minimization problem can also be converted to a conventional unrestricted penalized regression problem (Zhou and Pan, 2014).

2.2. *Supervised functional principal component analysis with failure time data*

Up to this point, we have discussed a specific type of bivariate spline basis over triangulation to approximate the imaging data $Z_i(\mathbf{s})$. However, the projection of the functional data $Z_i(\mathbf{s})$ onto the corresponding basis functions may not be associated with the failure time. Thus, under the supervised framework, we aim to allocate a specific set of orthonormal basis functions $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots)$, i.e., $||\boldsymbol{\phi}|| = 1$, $\langle \boldsymbol{\phi}_k, \boldsymbol{\phi}_{k'} \rangle = 0$ for $k < k'$, such that $Q(\boldsymbol{\phi})$ is maximized,

$$Q(\boldsymbol{\phi}) = \frac{\theta \text{var}(\langle \mathbf{Z}, \boldsymbol{\phi} \rangle) + (1-\theta)\text{cov}^2(\log(\widetilde{\mathbf{T}}), \langle \mathbf{Z}, \boldsymbol{\phi} \rangle)}{||\boldsymbol{\phi}||}, \tag{2.1}$$

for $0 < \theta \leq 1$. The orthonormal basis function $\boldsymbol{\phi}$ is called the supervised functional principal components (sFPCs). Note that when $\theta = 1$, (2.1) is equivalent to the conventional FPCA method. The rationale for using a squared covariance term is 2-fold: (i) it ensures that this term is always positive and (ii) it facilitates the eigenvalue decomposition process discussed next. The optimal value of θ can be estimated by conducting cross-validation on a user-specified grid in (0, 1], where it can be chosen to maximize the prediction performance.

Due to the presence of right-censoring, the covariance function needs to be adjusted by using inverse probability censoring weights (IPCW). For an ith observation, the IPCW can be expressed as $w_i = \frac{\triangle_i}{\widehat{G}(\log(\widetilde{T}_i))}$, where $\widehat{G}(t) = P(\log(C) > t)$, $i = 1, \ldots, n$ (Van der Laan and Robins, 2003; Welchowski *and others*, 2019). Under the assumption of independent censoring, the cumulative censoring distribution can be estimated with the Kaplan–Meier estimator. The covariance term can then be estimated by,

$$\text{cov}(\log(\widetilde{\mathbf{T}}), \langle \mathbf{Z}, \boldsymbol{\phi} \rangle) = \frac{1}{n} \sum_{i=1}^{n} w_i \langle \mathbf{Z}_i, \boldsymbol{\phi} \rangle (\log(\widetilde{T}_i) - \overline{T}),$$

where $\overline{T} = \frac{1}{n} \sum_{i=1}^{n} w_i \log(\widetilde{T}_i)$.

We now discuss an eigenvalue decomposition method for estimating the set of supervised FPCs (sFPCs) to accommodate imaging data with irregular boundaries and/or interior holes. We first start with a set of bivariate spline basis approximation over triangulations defined in Section 2.1 and rewrite $\{Z_i(\mathbf{s}_j) = \mathbf{a}_i^T \mathbf{B}(\mathbf{s}_j)\}_{j=1}^J$, where $\mathbf{B}(\mathbf{s}_j)$ denote the column vector $(B_1(\mathbf{s}_j), \ldots, B_K(\mathbf{s}_j))^T$. Suppose each function $\phi_k(\mathbf{s})$ in (2.1) has an expansion $\phi_k(\mathbf{s}) = \boldsymbol{\gamma}_k^T \mathbf{B}(\mathbf{s})$, then such set of coefficients, $\boldsymbol{\gamma}_k = (\gamma_{k,1}, \ldots, \gamma_{k,K})^T$, will serve as surrogate for identifying the sFPCs. The empirical score can thus be written as $\langle \mathbf{Z}_i, \boldsymbol{\phi} \rangle = \boldsymbol{\gamma}^T \mathbf{M} \mathbf{a}_i = \boldsymbol{\gamma}^T \mathbf{M}_i$, where $\mathbf{M}(k, k') = \langle \mathbf{B}_k, \mathbf{B}_{k'} \rangle$, and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K)^T$. The variance of the scores can then be estimated empirically as $\text{var}(\langle \mathbf{Z}, \boldsymbol{\phi} \rangle) = \frac{1}{n} \boldsymbol{\gamma}^T \mathbf{M} \mathbf{a}^T \mathbf{a} \mathbf{M} \boldsymbol{\gamma}$. The empirical covariance function, on the other hand, can be written as $\text{cov}(\mathbf{Y}, \langle \mathbf{Z}, \boldsymbol{\phi} \rangle) = \frac{1}{n} \boldsymbol{\gamma}^T \mathbf{M} \mathbf{a}^T (\mathbf{Y} \circ \mathbf{w})$, where $\mathbf{Y} = ((\log(\widetilde{T}_1) - \overline{T}), \ldots, (\log(\widetilde{T}_n) - \overline{T}))^T$ is the vector of demeaned outcome measures, with $\mathbf{w} = (w_1, \ldots, w_n)^T$ denoting the vector of IPCW weights, and $(\boldsymbol{a} \circ \boldsymbol{b})$ the element-wise multiplication between $\boldsymbol{a}$ and $\boldsymbol{b}$.

Then, we can reconstruct the objective function in (2.1) as,

$$Q(\boldsymbol{\phi}) = \frac{\boldsymbol{\gamma}^T \mathbf{U} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \mathbf{M} \boldsymbol{\gamma}}, \tag{2.2}$$

where

$$\mathbf{U} = \frac{\theta}{n}\mathbf{M}\mathbf{a}^T\mathbf{a}\mathbf{M} + \frac{1-\theta}{n^2}\mathbf{M}\mathbf{a}^T(\mathbf{Y}\circ\mathbf{w})(\mathbf{Y}\circ\mathbf{w})^T\mathbf{a}\mathbf{M}^T\ .$$

Note that maximizing (2.2) is equivalent to maximizing, $\boldsymbol{\delta}^T(\mathbf{M}^{-1/2})^T\mathbf{U}\mathbf{M}^{-1/2}\boldsymbol{\delta}$, subject to $\boldsymbol{\delta}^T\boldsymbol{\delta} = \mathbf{I}$, where $\boldsymbol{\delta} = \mathbf{M}^{1/2}\boldsymbol{\gamma}$. We can then estimate $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{K^\star}$ by finding the leading $K^\star$ eigenvectors of the matrix $(\mathbf{M}^{-1/2})^T\mathbf{U}\mathbf{M}^{-1/2}$. As a result, we are able to estimate $\widehat{\boldsymbol{\gamma}}_k = \mathbf{M}^{-1/2}\boldsymbol{\delta}_k$, and consequently the sFPCs as $\widehat{\phi}_k(\mathbf{s}) = \widehat{\boldsymbol{\gamma}}_k^T\mathbf{B}(\mathbf{s})$ for $\mathbf{s}\in\Omega$, $k = 1, \ldots, K^\star$. Note that $\widehat{\phi}_k(\mathbf{s})$ is guaranteed to be orthogonal because of the eigenvalue decomposition algorithm.

2.3. *Survival analysis incorporating the imaging data*

In this subsection, we will introduce the linkage between sFPCA and survival analysis. Given the set of supervised FPCs, we are able to obtain the scores of the ith image using inner product as $\widehat{\xi}_{i,k} = \langle\widehat{\mathbf{Z}}_i, \widehat{\boldsymbol{\phi}}_k\rangle$, $k = 1, \ldots, K^\star$. Note that $\widehat{Z}_i(\mathbf{s}) \approx \widehat{\mathbf{a}}_i^T\mathbf{B}(\mathbf{s})$, for $\mathbf{s}\in\Omega$. If J is sufficiently large, the inner product can also be approximate by $\sum_{j=1}^J \widehat{Z}_i(\mathbf{s}_j)\widehat{\phi}_k(\mathbf{s}_j)A(\mathbf{s}_j)$, where $A(\mathbf{s}_j)$ is the area of the pixel $\mathbf{s}_j$. With a set of demographic predictors $\mathbf{x}_i$ of length $P\times 1$ and the set of estimated scores $\widehat{\boldsymbol{\xi}}_i = (\widehat{\xi}_{i,1}, \ldots, \widehat{\xi}_{i,K^\star})^T$, we can write the survival distribution at time t as,

$$S_0(t)^{\exp(\boldsymbol{\alpha}^T\mathbf{x}_i + \boldsymbol{\beta}^T\widehat{\boldsymbol{\xi}}_i)}\ ,$$

under the proportional hazard assumption, where $S_0(t) = \exp(-\int_0^t h_0(u)\mathrm{d}u)$ with $h_0(t)$ being the baseline hazard function, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_P)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K^\star})^T$ are the regression coefficients.

3. Simulation studies

In this section, we will describe our simulation setup and examine the finite sample performance of the proposed method. We first simulate $K = 3$ 2D basis functions from tensor products of univariate Legendre polynomials basis on $[0,1]\times[0,1]$ under the resolution of 40×40. We restrain 921 pixels to fall inside the domain Ω formed by a circular boundary, similar to the mammograms in our motivating application. The 2D basis functions, $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3)^T$, are weighted by random factors to ensure orthonormality (Happ and Greven, 2018). We further simulate the individual-specific scores $\boldsymbol{\lambda}_i = (\lambda_{i,1}, \lambda_{i,2}, \lambda_{i,3})^T$ with mean 0 and covariance of $\mathrm{diag}(10, 8, 4)$. Given the orthonormal basis functions and scores, the individual-specific images are generated from the model:

$$Z_i(\mathbf{s}) = \mu(\mathbf{s}) + \sum_{k=1}^{3}\lambda_{i,k}\psi_k(\mathbf{s})\ ,\mathbf{s}\in\Omega\ ,$$

where $\mu(\mathbf{s}) = 0$ without loss of generality. We illustrate these basis functions with irregular boundaries in Figure S3 of the Supplementary material available at *Biostatistics* online.

We consider a proportional hazards model in this simulation study where,

$$h_i(t) = h_0(t)\exp\left\{\int_{\mathbf{s}\in\Omega} c(\mathbf{s})[Z_i(\mathbf{s}) - \mu(\mathbf{s})]\mathrm{d}\mathbf{s}\right\}.$$

We aim to investigate two scenarios for the image data coefficient. Under Scenario 1, we set $c(\boldsymbol{s}) = \psi_3(\boldsymbol{s})$, such that the hazard over time only depends on the third basis function. Under Scenario 2, we set $c(\boldsymbol{s}) =$

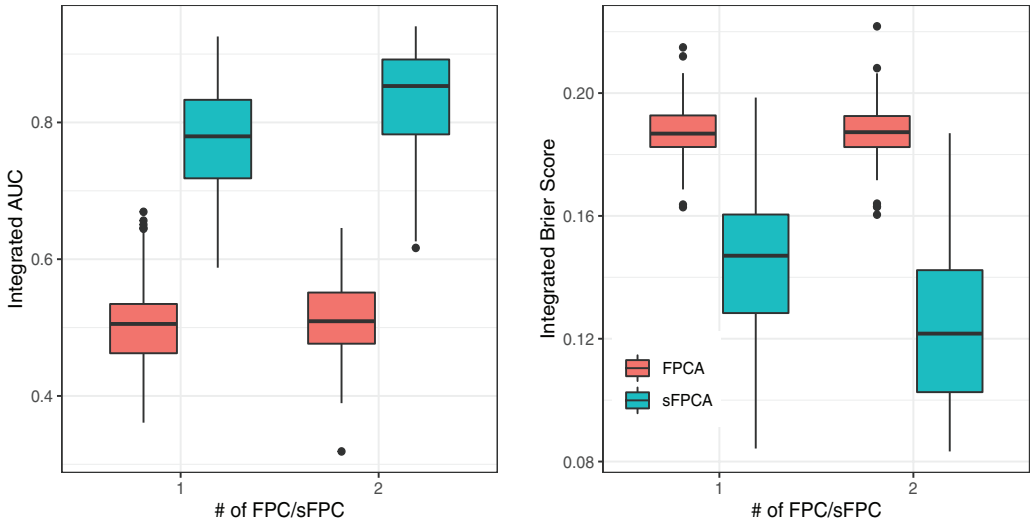

Fig. 2. Boxplots for the estimated integrated area under the receiver operating characteristic (ROC) curve (AUC) and integrated Brier scores (BS) with FPCA and supervised FPCA under simulation Scenario 1, with one and two FPCs/sFPCs, respectively.

$0.25\psi_1(\boldsymbol{s}) + 0.5\psi_2(\boldsymbol{s}) + \psi_3(\boldsymbol{s})$, such that the hazard over time is effected by a linear combination of all three basis functions. We set the maximum time to the end of the study to be 15 years. The baseline hazard function is assumed to follow a Weibull distribution $h_0(t) = \kappa\rho(\rho t)^{\kappa-1}$ with increasing risk over time, where we set $\kappa = 2$ and $\rho = 0.16$. The survival time T_i is generated from the inverse of the cumulative hazard function $H_i^{-1}(u)$, where $u \sim \text{unif}(0, 1)$. We have assumed the independent censoring scheme in this simulation study, where $C_i \sim \text{unif}(0, C_{\max})$, with $C_{\max}$ set at a value such that the % of being censored by the end of the study is approximately 30%.

We have simulated 400 individuals per data set, of which 300 are used for training the model and 100 are used as validation to avoid over-optimism. Within the training data set, the tuning parameter θ was chosen by conducting a 5-fold cross-validation on a grid from 0.001 to 1, with an increment of 0.111. We considered 80 triangles and degree $d = 2$ and $r = 1$ polynomial in this simulation study.

Figure 2 shows the prediction performance, repeated over 100 simulation runs, under Scenario 1 for the image coefficient $c(\boldsymbol{s}) = \psi_3(\boldsymbol{s})$. The prediction performance has been assessed in terms of both the model discrimination and calibration, represented with the integrated area under the receiver operating characteristic (ROC) curve (AUC) (Uno *and others*, 2007) and integrated Brier scores (Graf *and others*, 1999; Gerds *and others*, 2008). For a fair comparison, we constrain the sFPCA and conventional FPCA to take on the same number of basis functions (1 and 2). Some important observations can be gathered by looking at Figure 2. First, we observe that the sFPCA retains a higher integrated AUC and lower BS in general, as compared to the conventional FPCA. Second, we can see that the performance of the conventional FPCA did not improve with the increase in the number of FPCs. While the prediction performance improved under the sFPCA method with increased number of sFPCs used. This aligns with our expectation under Scenario 1, since the FPCA will not pick up the effect rising from the third basis function that is associated with the hazard. The sFPCA, on the other hand, should be able to allocate the set of basis functions that are ordered by association with the failure time.

We further investigated the gain in prediction performance for the sFPCA method by plotting the first sFPC against the conventional FPC from a randomly selected simulation run. Figure S4 of the

Supplementary material available at *Biostatistics* online shows the side-by-side plot of the first sFPC vs. FPC, where these were further compared with the true basis function $\psi_3(\mathbf{s})$ under scenario 1. It is apparent that the first basis of sFPCA mimics the true basis fairly well, while the conventional FPCA retains the appearance of the basis function that explains the most variation, $\psi_1(\mathbf{s})$.

Additional results under Scenario 2 are displayed in Figure S5 of the Supplementary material available at *Biostatistics* online. Similar conclusions can be drawn here as to Scenario 1. However, because the true image coefficient $c(\mathbf{s})$ is now a linear combination of all three basis functions, we do see improvements in the performance of the conventional FPCA with increased number of FPCs used. The general conclusion still remains valid, where the sFPCA outperforms the conventional FPCA, regardless of the number of basis functions used.

Software in the form of `R` code, together with a sample input data set and complete documentation is available on the github repository (https://github.com/jj113/Supervised-Triangulation).

4. The Joanne Knight Breast Health Cohort at Siteman Cancer Center

The motivating problem arose from the Joanne Knight Breast Health Cohort at Siteman Cancer Center, Washington University School of Medicine in St. Louis. Women were recruited in St. Louis, MO, from November 2008 to April 2012, and have been followed through October 2020. Upon recruitment, the breast cancer risk factors, mammographic breast density, and blood markers in a diverse population of women undergoing routine mammographic screening were documented. Recruited women are followed and recorded for pathology review confirmed incidence of breast cancer. In this analysis, we focused on the nested case–control study composed of 785 women who had a full field digital craniocaudal (CC) view mammogram available at the baseline. Individuals who had a diagnosis of breast cancer within the first 6 months since recruitment have been excluded. Of the 785 women, 246 have been diagnosed with breast cancer prior to the end of follow-up.

The set of mammogram imaging data were preprocessed before they were analyzed. As can be seen in an example presented in Figure S6(a) of the Supplementary material available at *Biostatistics* online, the original images of breast area are of distinct size and position within the mammograms for different individuals. To minimize the noise caused by this issue, we follow the approach proposed in Lee and Nishikawa (2018, 2019). We first segmented the breast area using a tight rectangular box, followed by soft tissue and label stamp removal procedures for parts outside of the breast region. Then, we resized each mammogram to 500×800 pixels using bicubic interpolation. The result of such preprocessing procedure is illustrated in Figure S6(b) of the Supplementary material available at *Biostatistics* online, where we can see that the breast areas are well matched. As such, the pixel intensity can be well averaged between the mammogram images for the left and the right breast as has often been done for the mammogram density in the literature.

We consider the Cox proportional hazards model in this analysis,

$$h_i(t) = h_0(t) \exp\left\{\alpha_1 \text{age}_i + \alpha_2 \text{weight}_i + \alpha_3 \text{height}_i + \int_{\mathbf{s}\in\Omega} c(\mathbf{s})[Z_i(\mathbf{s}) - \mu(\mathbf{s})]\text{d}\mathbf{s}\right\} \tag{4.3}$$

$$= h_0(t) \exp\left\{\alpha_1 \text{age}_i + \alpha_2 \text{weight}_i + \alpha_3 \text{height}_i + \boldsymbol{\beta}^T \hat{\boldsymbol{\xi}}_i\right\}, \tag{4.4}$$

where the baseline age, weight, and height are the demographic variables. The coefficient function $c(\mathbf{s})$ represents the effect of the demeaned mammogram image $Z_i(\mathbf{s})$ on the hazard function. Note that because $Z_i(\mathbf{s}) = \mu(\mathbf{s}) + \sum_{k=1}^{K^*} \xi_{i,k}\phi_k(\mathbf{s}), \mathbf{s} \in \Omega$, the coefficient function can also be rewritten in the form of $c(\mathbf{s}) = \sum_{k=1}^{K^*} \beta_k \phi_k(\mathbf{s})$ due to the orthogonality of $\phi_k(\mathbf{s})$, where the coefficients for the corresponding $K^\star$

Table 1. *Estimated median 5-year integrated receiver operating characteristic (ROC) curve (AUC) and integrated Brier Score (BS), with the standard error in parentheses, under the 10-fold cross-validation with benchmark model, and the inclusion of conventional and supervised FPC scores.*

	Benchmark		FPCA		sFPCA	
	AUC	BS	AUC	BS	AUC	BS
Age	0.568 (0.108)	0.065 (0.018)	—	—	—	—
Age + ψ	—	—	0.627 (0.100)	0.068 (0.019)	0.657 (0.113)	0.066 (0.018)
Age + weight + height	0.596 (0.132)	0.064 (0.017)	—	—	—	—
Age + weight + height + ψ	—	—	0.639 (0.091)	0.065 (0.019)	0.694 (0.104)	0.065 (0.018)

supervised scores are denoted by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{K^\star})^T$. The proportional hazards assumption was deemed reasonable upon formally inspecting the Schoenfeld residual plot for each of the baseline covariates.

In this analysis, the mammogram images and sFPCs are estimated as a linear combination of the bivariate spline basis functions of degree $d = 3$ and smoothness $r = 1$ defined over 115 triangles as shown previously in Figure 1. The number of triangles is chosen such that the breast area is well covered by the triangulation. Additional sensitivity analysis with the decreasing and increasing number of triangles can be found in the Supplemental Material. We have conducted a 10-fold internal cross-validation to avoid over-optimism in the prediction performance. The tuning parameter θ was chosen from a nested 5-fold cross-validation with fine increments on the grid (0, 1]. For each of the outer 10-fold cross-validation, prediction performance under several submodels was recorded. As a first pass of visualizing the results, we have randomly picked 20 individuals, 10 of which have been diagnosed with breast cancer before the end of the follow-up. Their estimated Score 1 vs. Score 2 under both the conventional and supervised FPCA are illustrated in Figure S10 of the Supplementary material available at *Biostatistics* online. The greater distinction between the cases and noncases under the supervised framework suggests that accounting for the covariance with the failure time indeed facilitates a better risk discriminatory performance. Additional illustration of the survival curve under the conventional vs. the supervised framework for two randomly chosen individual in the cohort can be found in Figure S11 of the Supplementary material available at *Biostatistics* online. The overall model performance will be formally assessed next.

To achieve a benchmark, we first estimated the 5-year integrated AUC and BS without including any conventional FPC or sFPC scores; see the first column of Table 1. Next, we assessed the prediction performance when we add in the conventional and supervised FPC scores. We propose a 2D grid search procedure for $(\theta, K^\star)$ in the real data analysis. We start with one sFPC ($K^* = 1$), then add new sFPC one by one. For a given value of K^*, we select an optimal $\hat{\theta}(K^*)$ by maximizing AUC, and the maximum is denoted as $\widehat{\mathrm{AUC}}(\hat{\theta}(K^*))$. We stop adding new FPCs if AUC does not improve significantly, that is, $\{\widehat{\mathrm{AUC}}(\hat{\theta}(K^*+1)) - \widehat{\mathrm{AUC}}(\hat{\theta}(K^*))\}/\widehat{\mathrm{AUC}}(\hat{\theta}(K^*)) < \epsilon$, where ϵ is a threshold. We set $\epsilon = 10^{-3}$ in the real data analysis. In this analysis, the number $K^\star$ varied slightly across different folds with a mean value of 10. From Table 1, we see that both the unsupervised and supervised FPCA method, on average, achieved higher 5-year integrated AUC as compared to the benchmark study. Of note, the sFPCA method retained the best prediction performance (AUC = 0.694) in all sub-models, compared to the conventional FPCA (AUC = 0.639), and the benchmark model (AUC = 0.596). These results suggest that the addition of functional scores can provide complementary information to the established risk factors for breast cancer.

Finally, we would like to give some insights on the estimated sFPCs and coefficient function. As our primary goal in this article lies in extracting mammogram features to improve prediction accuracy, we will

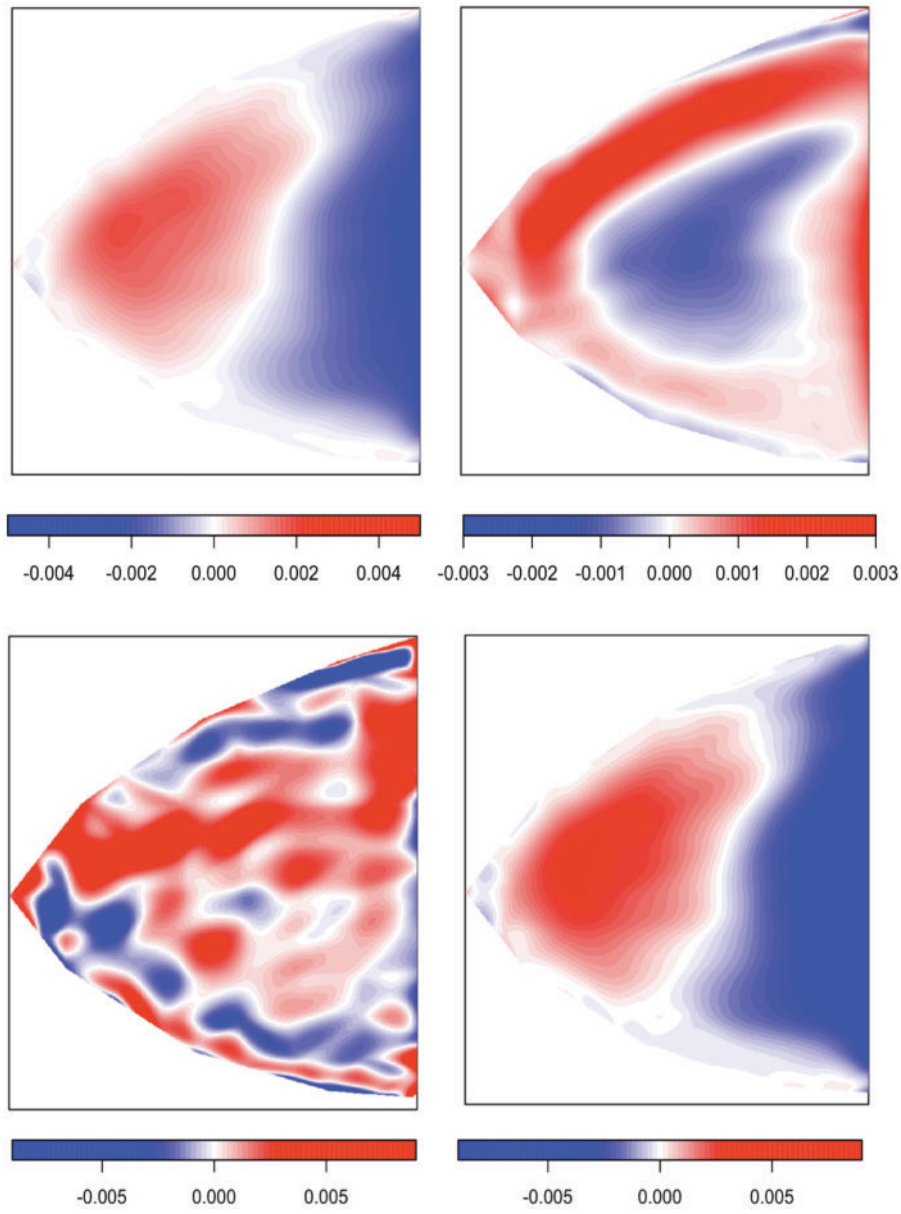

Fig. 3. The first two estimated functional principal components (first row) and supervised functional principal components (second row) for a randomly chosen fold within the cross-validation.

first elaborate on the interpretation of the estimated sFPCs. The features are represented with the scores to the sFPCs estimated by our proposed method, i.e., $\xi_{ik} = \int_{\mathbf{s}\in\Omega} \phi_k(\mathbf{s})[Z_i(\mathbf{s}) - \mu(\mathbf{s})]\mathrm{d}\mathbf{s}$. Based on the above formula, the kth sFPC score ξ_{ik} can be interpreted as the weighted average of the demeaned mammograms, while the kth sFPC $\phi_k(\mathbf{s})$ serves as the weight. Figure 3 shows the first two estimated FPC (first row) and sFPC (second row) from our cohort under a random train/test split in the cross-validation study. Here,

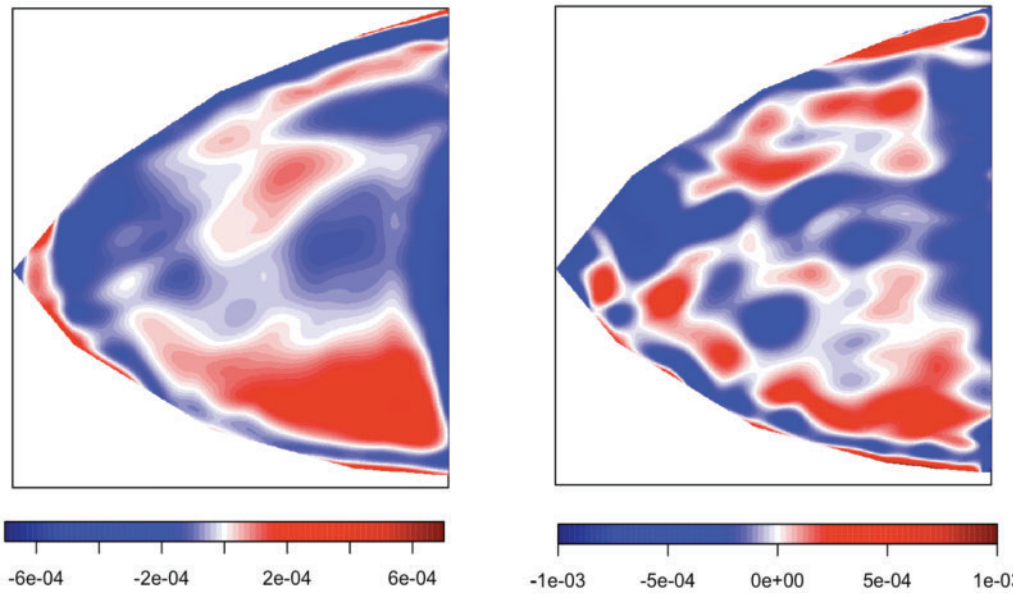

Fig. 4. The estimated coefficient surface $c(\mathbf{s})$ for the conventional FPCA (left panel) and supervised FPCA (right panel) for one fold within the cross-validation.

because the red regions represent positive estimates and the blue regions represent negative estimates, the corresponding scores can be interpreted as the weighted difference of demeaned mammograms between the red and blue regions. We notice that the first FPC estimated from the conventional FPCA method is close to the second sFPC estimated from the supervised FPCA method, telling us that the supervised FPCA method also considers this pattern of risk to be correlated with failure time.

The coefficient function can be expressed as a linear combination of sFPCs: $c(\mathbf{s}) = \sum_{k=1}^{K^*} \beta_k \phi_k(\mathbf{s})$. Figure 4 displays the estimated coefficient function from the same train/test split in the cross-validation study from Figure 3. As shown in (4.3), the coefficient function can be interpreted as the effect of the demeaned mammograms at location $\mathbf{s} \in \Omega$ to the hazard function. As the estimated coefficient function $c(\mathbf{s})$ is positive in the red regions and negative in the blue regions, the hazard function is affected by the difference of the demeaned mammogram at the red regions in comparison with the blue regions. Note that the coefficient functions estimated with the conventional and supervised FPCA methods have overlaps in some areas but are with different magnitudes of effect. The entire estimation, applied on our cohort data set, took approximately 45 mins on a standard laptop (2.9 GHz Intel Core i7, 16 GB RAM) without parallel computing.

5. DISCUSSION

With mammography being the primary breast cancer screening strategy, it is crucial to make full use of the mammogram imaging data to better discriminate women who are at higher and lower future risk. The primary objective of this study is to better capture the between-patient heterogeneity in the breast tissue by extracting mammogram-based features, in addition to the well-established breast cancer risk factors, to improve the risk prediction performance. We have proposed a supervised FPCA method in this article in facilitating feature extraction while accommodating the irregular boundary issue posed by the mammogram images using the bivariate splines that are piecewise polynomial functions over triangulations. These

extracted features based on the sFPCA over triangulations are then ordered by their association with the failure time. The finite sample performance of the proposed method has been assessed in our simulation studies, where we demonstrate improved prediction accuracy over the conventional unsupervised FPCA method. Note that as we increase the number of basis functions to infinity, the conventional FPCA would eventually span the same space with the supervised FPCA. However, it is important to maximize the signal-to-noise ratio in selecting the few that are most correlated with the survival time. Additionally, it is often more desirable to have fewer features to aid the interpretation from a clinical and practical point of view.

We further applied the proposed method to the Joanne Knight Breast Health Cohort at Siteman Cancer Center. Our results show that the sFPCA over triangulation method can achieve the best discriminatory performance as compared to the conventional FPCA and the benchmark model. These results suggest that the mammogram-based features extracted with the proposed sFPCA method can provide complementary information to augment the well-established breast cancer risk factors. Additionally, the coefficient surface estimated from the sFPCA revealed patches of hot spots within the breast area. This may facilitate future research in precision prevention, where individuals with parenchymal complexity concentrated in the identified hot areas may be followed more closely, or offered chemoprevention or other interventions, as these individuals may be at higher breast cancer risk. These interesting aspects on risk locations and subsequent lesions remain to be explored.

Supplementary material

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

Acknowledgments

The authors wish to thank the referees of this article for providing helpful comments and suggestions.

Conflict of Interest: None declared.

Funding

This research is partially supported by the NCI (R37 CA256810), Breast Cancer Research Foundation (BCRF 20-028), and the NSERC Discovery Grant (RGPIN-2018-06008).

References

Allen, G. I. (2013). Multi-way functional principal components analysis. In: *2013 5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, pp. 220–223.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A. and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68**, 394–424.

Chen, K. and Lei, J. (2015). Localized functional principal component analysis. *Journal of the American Statistical Association* **110**, 1266–1275.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. New York: Springer.

Gellar, J. E., Colantuoni, E., Needham, D. M. and Crainiceanu, C. M. (2015). Cox regression models with functional covariates for survival data. *Statistical Modelling* **15**, 256–278.

Gerds, T. A., Cai, T. and Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50**, 457–479.

GRAF, E., SCHMOOR, C., SAUERBREI, W. AND SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* **18**, 2529–2545.

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.

HUANG, J. Z., SHEN, H. AND BUJA, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *Journal of the American Statistical Association* **104**, 1609–1620.

KONG, D., IBRAHIM, J. G., LEE, E. AND ZHU, H. (2018). FLCRM: functional linear Cox regression model. *Biometrics* **74**, 109–117.

LAI, M.-J. AND WANG, L. (2013). Bivariate penalized splines for regression. *Statistica Sinica*, **23**, 1399–1417.

LEE, J. AND NISHIKAWA, R. M. (2018). Automated mammographic breast density estimation using a fully convolutional network. *Medical Physics* **45**, 1178–1190.

LEE, J. AND NISHIKAWA, R. M. (2019). Detecting mammographically occult cancer in women with dense breasts using deep convolutional neural network and radon cumulative distribution transform. *Journal of Medical Imaging* **6**, 044502.

LI, G., SHEN, H. AND HUANG, J. Z. (2016). Supervised sparse and functional principal component analysis. *Journal of Computational and Graphical Statistics* **25**, 859–878.

LIN, N., JIANG, J., GUO, S. AND XIONG, M. G. (2015). Functional principal component analysis and randomized sparse clustering algorithm for medical image analysis. *PLoS One* **10**, e0132945.

LIN, Z., WANG, L. AND CAO, J. (2016). Interpretable functional principal component analysis. *Biometrics* **72**, 846–854.

NIE, Y. AND CAO, J. (2020). Sparse functional principal component analysis in a new regression framework. *Computational Statistics & Data Analysis* **152**, 107016.

NIE, Y., WANG, L., LIU, B. AND CAO, J. (2018). Supervised functional principal component analysis. *Statistics and Computing* **28**, 713–723.

RAMSAY, J. O. AND DALZELL, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B* **53**, 539–572.

RAMSAY, J. O. AND SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd edition. New York: Springer.

RICE, J. A. AND SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B* **53**, 233–243.

UNO, H., CAI, T., TIAN, L. AND WEI, L.-J. (2007). Evaluating prediction rules for t-year survivors with censored regression models. *Journal of the American Statistical Association* **102**, 527–537.

VAN DER LAAN, M. AND ROBINS, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer Science & Business Media.

WANG, Y., WANG, G., WANG, L. AND OGDEN, R. T. (2020). Simultaneous confidence corridors for mean functions in functional data analysis of imaging data. *Biometrics* **76**, 427–437.

WELCHOWSKI, T., ZUBER, V. AND SCHMID, M. (2019). Correlation-adjusted regression survival scores for high-dimensional variable selection. *Statistics in Medicine* **38**, 2413–2427.

YAO, F., MÜLLER, H.-G. AND WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

YU, S., WANG, G., WANG, L., LIU, C. AND YANG, L. (2020). Estimation and inference for generalized geoadditive models. *Journal of the American Statistical Association* **115**, 761–774.

ZHANG, X., SUN, Q. AND KONG, D. (2021). Supervised principal component regression for functional response with high dimensional predictors. *arXiv preprint arXiv:2103.11567*.

ZHOU, L. AND PAN, H. (2014). Smoothing noisy data for irregular regions using penalized bivariate splines on triangulations. *Computational Statistics* **29**, 263–281.

ZIPUNNIKOV, V., CAFFO, B., YOUSEM, D. M., DAVATZIKOS, C., SCHWARTZ, B. S. AND CRAINICEANU, C. (2011). Functional principal component model for high-dimensional brain imaging. *NeuroImage* **58**, 772–784.

[*Received March 5, 2021; revised June 26, 2021; accepted for publication July 29, 2021*]